AP3456
The Central Flying School (CFS) Manual of Flying

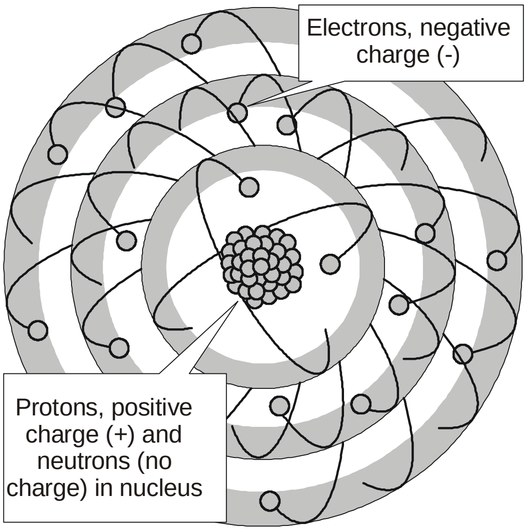
Volume 14 – Electronics

CHAPTER 1 - BASIC ELECTRICITY

Electron Theory

1. The smallest portion of any piece of matter which still retains the properties of the original substance is called a molecule. Molecules are made up of atoms which in turn are composed of a mixture of negatively and positively charged particles (see Fig 1). Normally there is an equal amount of each charge in an atom; the negatively charged particles (electrons) being balanced by the positively charge particles (protons) in the nucleus. The neutrons which form part of the nucleus have no charge.

14-1 Fig 1 An Atom


Electrons, negative charge (-)
Protons, positive charge (+) and neutrons (no charge) in nucleus

Conductors and Insulators

2. In some materials, the outer orbital electrons are loosely attached to the atom’s nucleus and are free to wander within the body of the material. If a voltage is applied across the material, these free electrons experience a force and will move under its influence. A bodily movement of electrons is called an electric current and substances in which appreciable amounts will flow are called conductors. Insulators will pass only a very small number of electrons, even at high voltages.

3. The property of a substance which tends to oppose the flow of electric current is called the resistance of the substance. Good conductors have low resistance while poor conductors (insulators) have high resistance.

Electric Units

4. The fundamental unit of electricity is the electron, but this is far too small to be of any practical use for measurement. Consequently the practical unit of electricity or electric charge, the coulomb, is much larger. The coulomb is equal to the charge on 6.28 × 10^{18} electrons. When a charge of one coulomb flows in one second through a conductor, it constitutes a current of one ampere (usually shortened to 'amp').

$$\text{ie, amperes (A)} = \frac{\text{coulombs (Q)}}{\text{seconds (t)}}$$

5. Current will only flow in a conductor if the conducting material is connected between the terminals of a power source such as a battery. The battery is said to provide an electromotive force or emf. The practical unit of emf is the volt. The electromotive force may be thought of as driving electric current in the same way as pressure will drive a current of water through a pipe. Water flows from a greater height to a lower point and similarly electric charge is said to flow from a higher to a lower potential. The amount of water flowing per second will depend upon the diameter of the pipe. Likewise, the electric current will depend upon the resistance of the conducting wire, which in turn will depend on its diameter.

6. The direction of conventional current flow is taken as being from positive to negative, which is the opposite direction to that in which the electrons travel.

Ohm's Law and Resistance

7. Ohm's law states that under constant physical conditions the current flowing in a conductor is directly proportional to the potential difference between its ends and inversely proportional to its resistance. If the potential difference is increased then the current will increase by the same factor; if resistance increases then the current decreases. Ohm's law is expressed by the following formula:

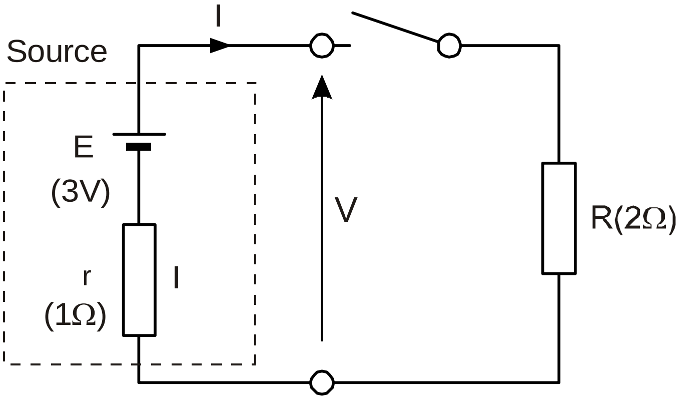
$$\text{Current (I)} = \frac{\text{Voltage (V)}}{\text{Resistance (R)}}$$

8. The unit of resistance is the Ohm, and a conductor is said to have a resistance of one ohm if it will allow a current of one ampere to flow when a potential difference of one volt is applied across it. The symbol for ohms is the Greek capital letter omega (Ω). For larger values of resistance, the units Kilohm and Megohm are used (Kilohm (K Ω) = 1,000 Ω; Megohm (M Ω) = 1,000,000 Ω)

EMF and Potential Difference

9. It is important to distinguish between electromotive force (emf) and potential difference (pd), because, although they are both measured in volts, they differ considerably. For any given power source the off-load voltage across the output terminals (pd) is equal to the driving force (emf) of the source. However, once an external load is connected across the terminals pd is no longer equal to emf. This is because all sources of power have internal resistance and voltage is dropped across this resistance once current flows in the circuit. To help distinguish between the two terms, emf is represented by the letter 'E' and pd by the letter 'V'. The following example is intended to illustrate the difference.

14-1 Fig 2 Circuit to show EMF and PD



Example 1. In Fig 2, the emf of the battery (E) is 3V and an internal resistance (r) of 1Ω is assumed. When the switch is closed, current (I) will flow through the external load (R) and the internal resistance of the battery,

$$\text{ie}\quad I = \frac{E}{R + r} = \frac{3}{2 + 1} = 1\text{A}$$

and the voltage at the battery terminal

$$(V) = IR = 1\times 2 = 2V.$$

Thus, it can be seen that although the battery emf is 3V, there is only a pd of 2V across the terminals.

Resistivity

10. The longer a piece of wire is then the higher its resistance will be. Similarly, the thicker the wire is the lower its resistance will be. However, resistance is also determined by the nature of the material being used. This is referred to as the resistivity of the material. Resistivity is expressed in ohms metres and is defined as being the resistance per metre of material with cross-sectional area of one square metre; it is represented by the Greek letter rho (ρ). Therefore:

$$\text{Resistance (R)} \;=\; \frac{\text{Resistivity }(\rho)\times\text{Length }(l)}{\text{Cross - sectional area}(a)}$$

Temperature Coefficient of Resistance

11. In the statement of Ohm’s law the words "under constant physical conditions" are used; this is because the resistance of most materials varies with temperature. For metallic conductors, resistance increases nearly uniformly with temperature, and each material is given a temperature coefficient. This coefficient is defined as the fractional increase in resistance per degree increase in temperature above 0 °C. Temperature coefficient of resistance is represented by the Greek letter alpha (α). For practical purposes the following formula can be applied:

$$R_t = R_0\left(1 + \alpha t\right)$$

Where R_t = resistance at t ºC

R_0 = resistance at 0 ºC

α = temperature coefficient of resistance

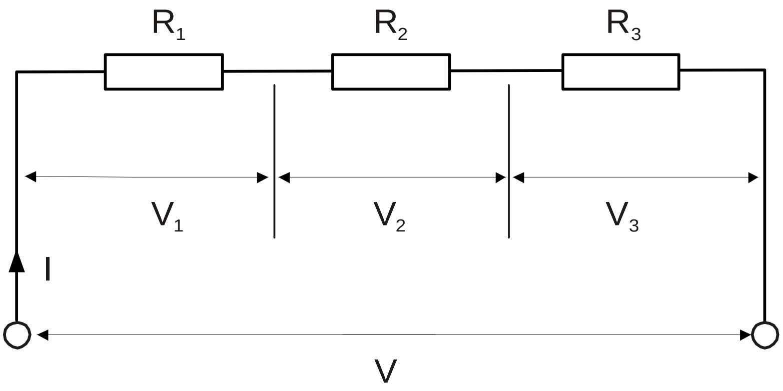
t = operating temperature

12. The temperature coefficients of alloys are generally much smaller than those of pure metals, and special alloys are used in making standard resistors which vary only slightly with changes in temperature. The resistance of carbon and glass decreases with increases in temperature so these materials are given a negative temperature coefficient.

Resistors in Series

13. When resistors are connected in series, their resultant resistance is equal to the sum of their separate values. In Fig 3, three resistors R_1, R_2, and R_3 are connected in series.

14-1 Fig 3 Resistors in Series



The voltages across the resistors are V_1, V_2, and V_3 respectively. Since the same current (I) flows through each resistor, then by applying Ohm's law:

$$V_1 = IR_1 \qquad V_2 = IR_2 \qquad V_3 = IR_3$$

The total voltage across the three resistors is given by,

$$V_T = V_1 + V_2 + V_3 \quad \text{or} \quad IR_1 + IR_2 + IR_3$$

$$= I\,(R_1 + R_2 + R_3)$$

Therefore, the three resistors in series are equivalent to a single resistor R_T of value,

$$R_T = R_1 + R_2 + R_3$$

Example 2. Three resistors of values 100 Ω, 220 Ω and 470 Ω, are connected in series. Calculate the total resistance, and find the current flow if 50V is applied across the combination.

$$R_T = R_1 + R_2 + R_3$$

$$= 100\ \Omega + 220\ \Omega + 470\ \Omega$$

$$= 790\ \Omega$$

Total resistance is 790 Ω

$$I = \frac{V}{R_T}$$

$$= \frac{50\text{V}}{790\Omega} = 0.063\text{ A} \quad \text{or} \quad 63\text{ mA}$$

A current of 63 mA flows through the circuit.

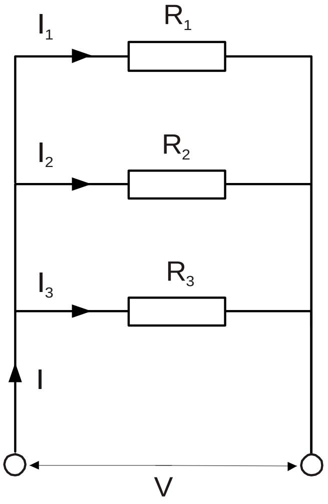
Resistors in Parallel

14. When resistors are connected in parallel, their total resistance is given by the equation:

$$\frac{1}{R_T} = \frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3}$$

15. In Fig 4 the three resistors are connected in parallel. The same voltage appears across each resistor but the currents flowing through the resistors are not the same.

14-1 Fig 4 Resistors in Parallel



Referring to Fig 4, the currents flowing through the resistors are I_1, I_2 and I_3, respectively. Once again, by applying Ohm's law:

$$I_1 = \frac{V}{R_1} \qquad I_2 = \frac{V}{R_2} \qquad I_3 = \frac{V}{R_3}$$

The total current flowing in the circuit is given by,

$$I = I_1 + I_2 + I_3 = \frac{V}{R_1} + \frac{V}{R_2} + \frac{V}{R_3}$$

$$= V\left(\frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3}\right)$$

But, $I = \frac{V}{R}$ where R is the equivalent of the three resistors. Dividing throughout by V gives:

$$\frac{1}{R} = \frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3}$$

Example 3. Three resistors, of values 100 Ω, 200 Ω, and 400 Ω, are connected in parallel. Calculate the total resistance and find the current flow if 200V is applied across the combination.

$$\frac{1}{R_T} = \frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3}$$

$$= \frac{1}{100\Omega} + \frac{1}{200\Omega} + \frac{1}{400\Omega}$$

$$= \frac{4+2+1}{400\,\Omega} = \frac{7}{400\,\Omega}$$

$$R_T = \frac{400\Omega}{7} \quad = 57.14\,\Omega$$

The total resistance is 57.14 Ω

$$I = \frac{V}{R_T} \qquad = \frac{200V}{57.14\Omega} \qquad = 3.5\ A$$

The total current flowing in the circuit is 3.5 A.

Kirchhoff's Laws

16. In the preceding paragraphs on resistors in series and parallel, two important laws due to Kirchhoff have been assumed. These are:

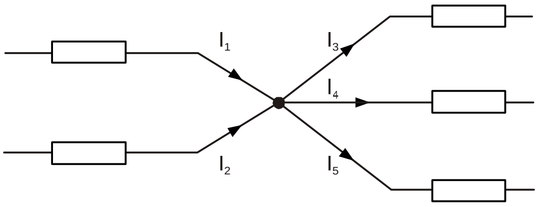
a. *The First Law (Current Law)*. This law states that at any junction in an electric circuit, the total current flowing towards that junction is equal to the total current away from the junction. For the circuit shown in Fig 5:

$$I_1 + I_2 = I_3 + I_4 + I_5$$

or

$$I_1 + I_2 - I_3 - I_4 - I_5 = 0$$

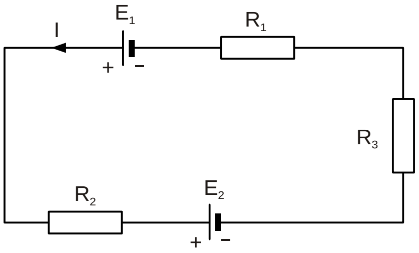
14-1 Fig 5 Kirchhoff's First Law



b. *The Second Law (Voltage Law)*. The second law states that in any closed loop in a network, the algebraic sum of the voltage drops (ie products of current and resistance) taken around the loop is equal to the resultant emf in that loop. For the circuit shown in Fig 6:

$$E_1 - E_2 = IR_1 + IR_2 + IR_3$$

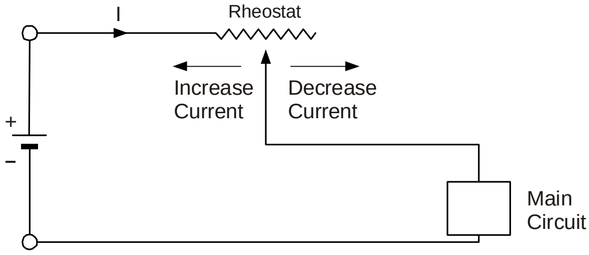
14-1 Fig 6 Kirchhoff's Second Law



Rheostats and Potentiometers

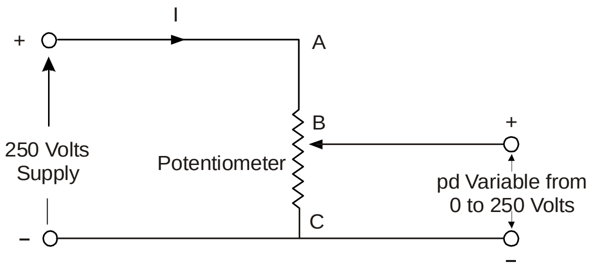
17. A rheostat is a variable resistor which is placed in series with a circuit, as shown in Fig 7, in order to control the current.

14-1 Fig 7 A Rheostat



18. A potentiometer is a tapped resistor which is connected across a pd source and used to control the voltage applied to the main circuit (see Fig 8).

14-1 Fig 8 A Potentiometer



19. Since the potentiometer is connected directly across the supply, its resistance must be large enough to prevent a heavy current being taken from the supply. If the load on the potentiometer takes an appreciable amount of current then the ratio between the two pds (A-B and B-C) will not be quite the same as the ratio of the two resistances of the potentiometer.

Energy and Power

20. When an electric current flows in a conductor, energy is expended, ie work is done. The amount of energy expended is given by the product of the quantity of the charge moved and the difference in potential through which it moves. The unit of energy is the joule, therefore:

$$\begin{aligned} \text{W (joules)} &= \text{Q (coulombs)} \times \text{V (volts)} \\ &= \text{V (volts)} \times \text{I (current)} \times \text{t (time)} \\ \text{Since Q} &= \text{current} \times \text{time} \end{aligned}$$

Power (P) is measured in watts and is the rate of expenditure of energy, or work done per second.

$$\mathrm{P}\ (\text{watts}) = \frac{\mathrm{V}(\text{volts}) \times \mathrm{I}(\text{current}) \times \mathrm{t}(\text{time})}{\mathrm{t}(\text{time})}$$

$$= \mathrm{VI} \text{ or } \mathrm{I}^2\mathrm{R} \text{ or } \frac{\mathrm{V}^2}{\mathrm{R}}$$

21. A joule can therefore be called a watt-second, but for practical purposes the joule is far too small a unit to work with, and for this reason the Kilowatt-hour (3.6×10^6 joules) is used as the standard unit of energy.

Heating Effect of a Current

22. The energy expended in a purely resistive conductor is converted into heat. The unit of heat energy is the calorie, which is defined as the amount of heat required to raise the temperature of one cubic centimetre of water through one degree centigrade. One calorie of heat is produced in a conductor for every 4.2 joules of electrical energy expended. The total heat produced in any conductor is given by:

$$\frac{\text{V (voltage)} \times \text{I (current)} \times \text{t (time)}}{4.2} \text{ calories}$$

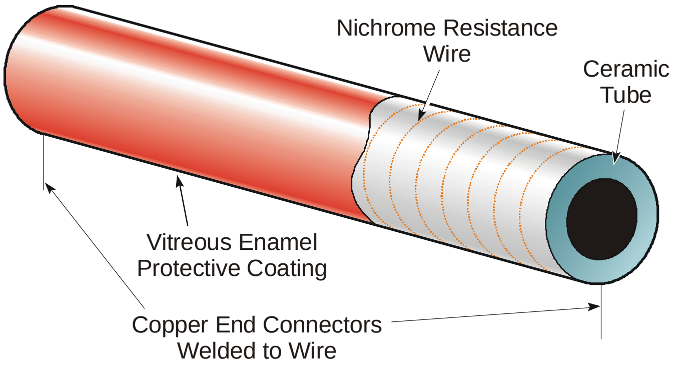
The Maximum Power Transfer Theorem

23. The power transferred from a supply source to a load is at a maximum when the resistance of the load is equal to the internal resistance of the source. Referring to the circuit shown in Fig 2, maximum power is delivered when the load is equal to 1 ohm. This theorem is particularly important in the design of electronic equipment where maximum power transference is needed between circuit stages.

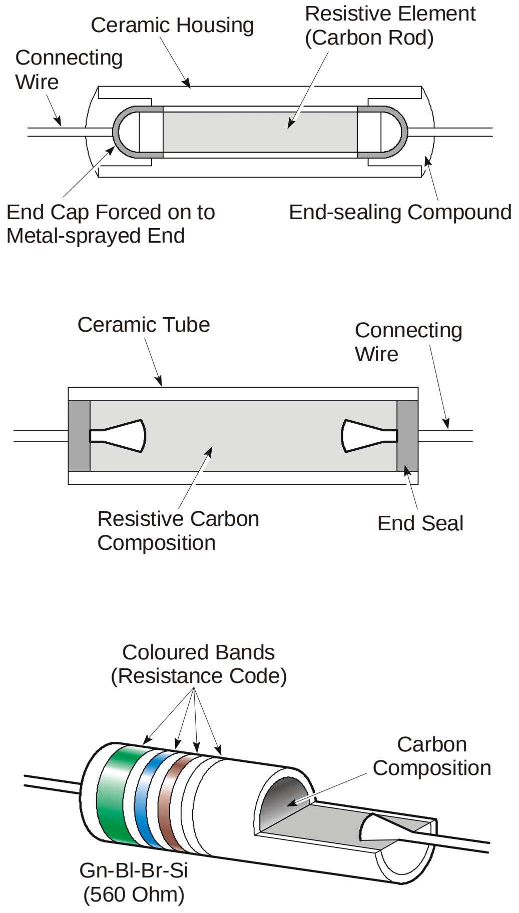
Practical Resistors

24. There are several different types of resistor in common use, however, the three main types are carbon, metal film and wire-wound. General purpose resistors are more often of the carbon type; they are inexpensive and perform reasonably well in circuits where the design requirements are not too critical. Where higher accuracy components are called for, carbon or metal film resistors (low power) and wire-wound resistors (high power) are used. Resistors can overheat when overloaded and those installed in aircraft are therefore housed within a ceramic coating or tube. Fig 9 shows the construction of a typical high power wire-wound resistor and Fig 10 shows the composition of two types of carbon resistor, and a colour coded exterior.

14-1 Fig 9 Construction of a Wire-wound Resistor


Nichrome Resistance Wire
Ceramic Tube
Vitreous Enamel Protective Coating
Copper End Connectors Welded to Wire

14-1 Fig 10 Two Types of Carbon Resistor


Ceramic Housing
Resistive Element (Carbon Rod)
Connecting Wire
End Cap Forced on to Metal-sprayed End
End-sealing Compound
Ceramic Tube
Connecting Wire
Resistive Carbon Composition
End Seal
Coloured Bands (Resistance Code)
Carbon Composition
Gn-Bl-Br-Si (560 Ohm)

Preferred Values of Resistance

25. It is desirable to keep the number of resistance values to be manufactured or held in stock to a reasonable figure. In addition, the final value of a carbon composition resistor is rarely exactly the value that was originally intended. Because of these factors a logarithmic series of preferred resistor values has been chosen. For the ± 20% tolerance range, the ohmic values are 10, 12, 22, 33, 47, and 68, while the values for the ± 10% range are 10, 12, 15, 18, 22, 27, 33, 39, 47, 56, 68, and 82. Resistors manufactured to these values (or these values multiplied by factors of 10) are termed 'preferred value resistors'.

The Resistor Colour Code

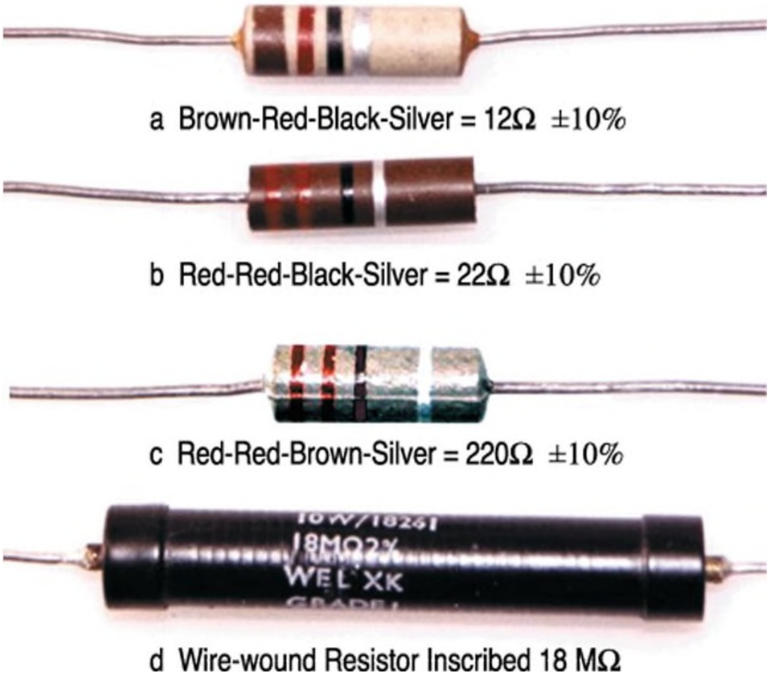
26. In order to identify the value of resistors and their tolerance range, methods of exterior markings have been devised. One such method of identification is by means of coloured bands. The bands are coloured according to a standard code; each colour representing a digit as indicated in Table 1.

Table 1 Colour Code for Fixed Resistors

Colour	Significant Figures	Multiplier	Tolerance
Silver	-	10^{-2}	± 10%
Gold	-	10^{-1}	± 5%
Black	0	1	-
Brown	1	10	± 1%
Red	2	10^{2}	± 2%
Orange	3	10^{3}	-
Yellow	4	10^{4}	-
Green	5	10^{5}	± 0.5%
Blue	6	10^{6}	± 0.25%
Violet	7	10^{7}	± 0.1%
Grey	8	10^{8}	-
White	9	10^{9}	-
None	-	-	± 20%

27. The coloured bands are painted on the body of the resistor usually biased towards one end. The colour of the band nearest the end indicates the first number. For easy recognition, on some resistors this band may be wider than the others, especially if there is likely to be any doubt over choice of ends. The colour of the second band gives the second number and the third band provides the multiplier (decimal place). The fourth band indicates the resistor’s tolerance and will most often be either silver (10% tolerance) or gold (5% tolerance). Neither of these colours is used for significant figures, so it will be clear that counting of the value should begin at the other end. If there is no fourth band, the tolerance is 20%. Fig 11 shows a selection of colour-coded resistors with their values decoded together with a wire-wound example which has its value printed on the body.

14-1 Fig 11 Examples of Resistors with their Values



a Brown-Red-Black-Silver = 12Ω ±10%

b Red-Red-Black-Silver = 22Ω ±10%

c Red-Red-Brown-Silver = 220Ω ±10%

d Wire-wound Resistor Inscribed 18 MΩ

CHAPTER 2 - MAGNETISM

Introduction

1. Around 3000 BC an iron ore known as lodestone was discovered in the Chinese desert. The ancient Chinese found that this ore possessed some unusual properties namely:

a. The attraction or repulsion of other pieces of lodestone.

b. The attraction of pieces of soft iron.

c. Orientation in the direction of the Earth's North and South poles when suspended freely.

The ability of a piece of lodestone to align itself with the Earth's North and South poles led to the piece ends being called 'poles' and were labelled North (seeking) and South (seeking) poles. Such a piece of lodestone was called a magnet and the force which it possessed referred to as magnetism.

2. Magnetism was thought of as a force in its own right until Ampere discovered that a small coil carrying an electric current behaves like a magnet. It was suggested that electric and magnetic forces are both manifestations of the same electromagnetic force. Ampere's theory, which gave a natural explanation of the fact that no isolated magnetic pole had ever been observed, is essentially similar to modern day atomic theory. Electric charges can either be positive or negative, and the forces act so as to repel like charges and attract unlike charges. A moving electric charge produces a magnetic field. The electric current flowing through a coil of wire produces a controllable magnetic field, while in the case of a normal bar magnet the moving charges are the electrons circulating inside the atoms.

3. All materials exhibit magnetic properties, the degree to which they are exhibited depends on the distribution of electrons in the outer shells of the material's atoms. Materials can be placed in one of the following main magnetic categories:

a. **Diamagnets**. Materials which oppose an applied magnetic field. Examples of such materials are copper, bismuth and hydrogen.

b. **Paramagnets**. Materials which slightly aid an applied magnetic field.

c. **Ferromagnets**. Materials which greatly aid an applied magnetic field. Examples of such materials are iron, steel, nickel and cobalt. Ferromagnetism will be discussed in more detail later in this chapter.

Magnetic Fields

4. A magnet affects the space around it in such a way that other materials placed in this space experience forces. The space in which this occurs is known as a magnetic field, and its presence can be detected using iron filings or a compass needle. The magnetic field pattern formed by iron filings around a bar magnet is shown in Fig 1. Michael Faraday referred to the apparent lines as lines of magnetic flux, and although flux does not exist as separate lines the concept of flux is useful in order to explain the effects of magnetism. Magnetic flux has the following properties:

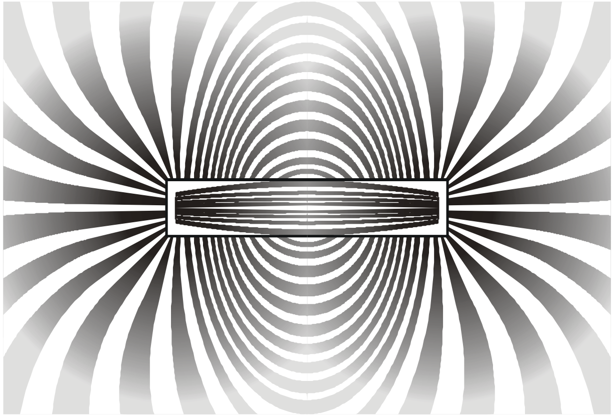
a. The direction of the apparent lines, outside the magnet, is from the North to the South pole.

b. Lines of flux form complete closed paths, they cannot end in space. Lines which do not close around the magnet join those of the Earth's magnetic field.

c. Lines of magnetic flux never intersect each other, although may become extremely distorted.

d. Where the flux is more intense, then the lines are closer together.

14-2 Fig 1 The Magnetic Field Around a Bar Magnet



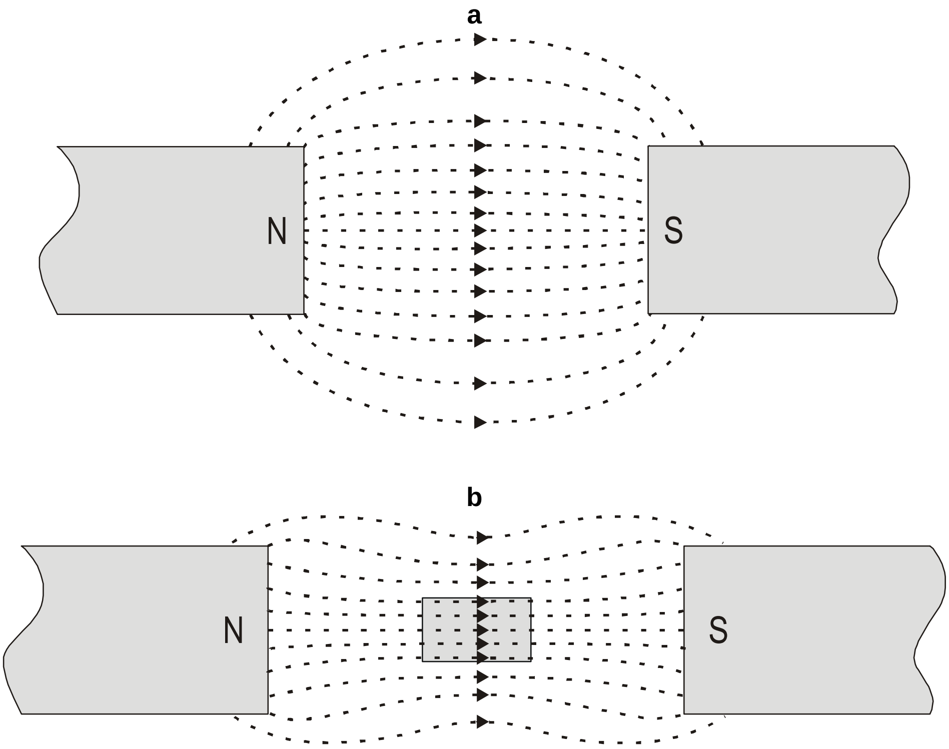
5. Magnetic lines of force associated with a given magnet are modified in the presence of other magnets and magnetic materials. Fig 2 illustrates the effective fields produced by:

a. Two adjacent, unlike poles

b. The insertion of an iron bar in the space between the poles.

In the latter case, the flux lines appear to be concentrated in the vicinity of the iron. This is because the iron becomes magnetized by the external field and produces flux lines of its own, which reinforces the original field. This is referred to as induced magnetism.

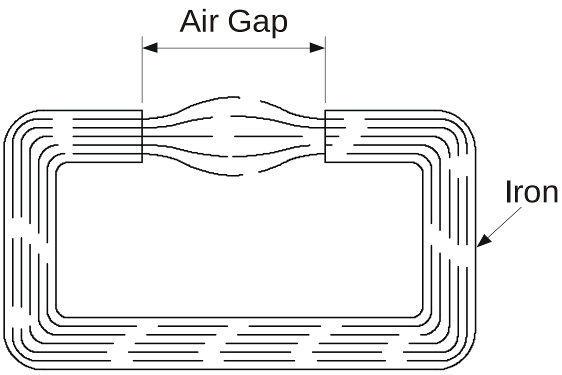
14-2 Fig 2 Modified Magnetic Field Patterns



Flux and Magnetic Circuits

6. Magnetic flux may be likened to an electric current, insomuch as it only exists in circuits. The closed path in which the magnetic flux exists is known as a magnetic circuit. An example of a magnetic circuit is shown in Fig 3. The flux 'flows' through the iron ring and across the air gap, thus completing the magnetic circuit. The symbol for magnetic flux is Φ (phi) and the unit is the Weber (Wb).

14-2 Fig 3 The Magnetic Circuit



7. The amount of Flux passing through a defined area, which is perpendicular to the direction of the flux, is referred to as the flux density, and is a far more useful quantity to work with than the total amount of flux. The symbol for flux density is B and the unit is the tesla (abbreviated as 'T'). Therefore:

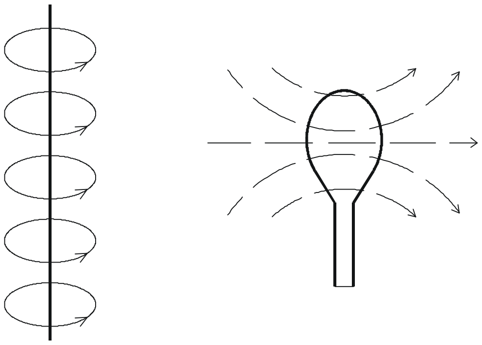
$$\text{Flux Density (B)} = \frac{\text{Total Magnetic Flux } (\Phi)}{\text{Area (A)}}$$

$$\text{and 1 tesla} = \frac{1\,\text{weber}}{1\,\text{square metre}} \text{ or 1 T = 1 Wb/m}^2.$$

Magnetic Flux by an Electric Current

8. When an electric current flows in a wire, a magnetic field is set up around the wire. This field may be represented by lines of force, with the strongest force being experienced near the wire. If the wire is bent to form a loop, the field will be strengthened at the centre of the loop (see Fig 4).

14-2 Fig 4 Magnetic Fields produced by Straight and Looped Wires

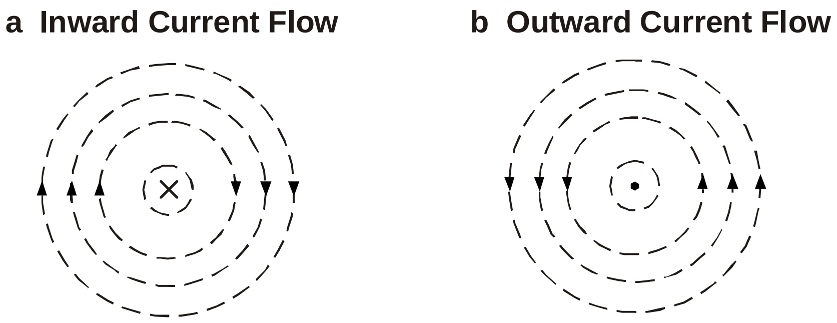


9. The direction of the magnetic field produced by a current carrying wire can be determined by applying Maxwell's corkscrew rule, which states:

> If a normal right-hand threaded corkscrew travels along a conductor in the direction of the current, then the direction of rotation of the corkscrew is in the direction of the magnetic field.

The convention that is used to show the direction of the current in a conductor, in diagrammatic form, is that the current flowing into the paper is shown by the flight of an arrow, while a current flowing outwards is shown by the arrow point. This is illustrated by Fig 5.

14-2 Fig 5 Current Flow Representation



a Inward Current Flow

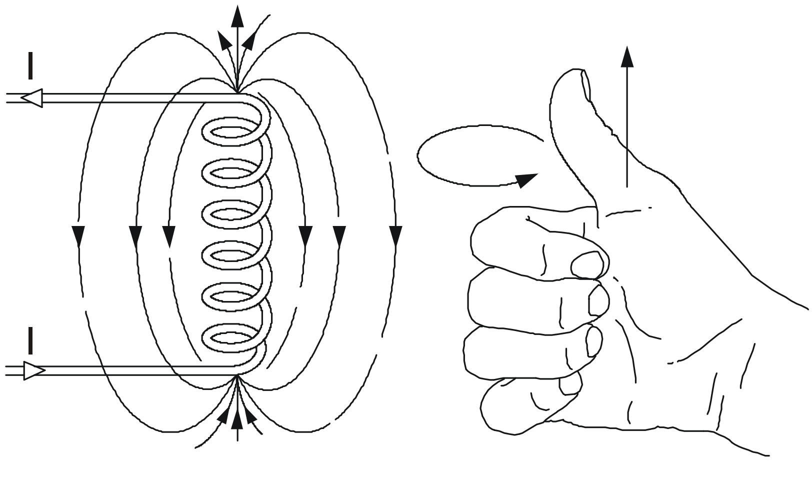
b Outward Current Flow

10. If a wire conductor is wound in the shape of a coil, then this is referred to as a solenoid. The direction of the magnetic field associated with a solenoid can be established using either the corkscrew rule or the more popular right-hand grip rule. The grip rule states:

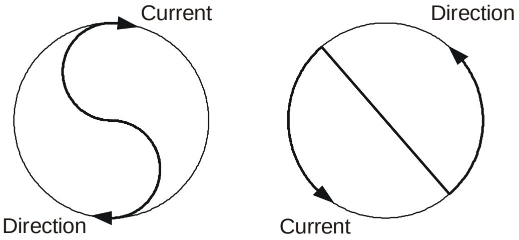
> If a coil is gripped with the right hand, and with the fingers pointing in the direction of the current, then the thumb outstretched parallel to the axis of the solenoid, points in the direction of the magnetic field inside the solenoid (ie points in the direction of the North pole).

The magnetic field associated with a solenoid and the grip is shown in Fig 6. Fig 7 illustrates a simple method of determining the polarities of the solenoid ends.

14-2 Fig 6 The Magnetic Field of a Solenoid and the Grip Rule



14-2 Fig 7 The Polarity Rule



Magnetomotive Force and Field Strength

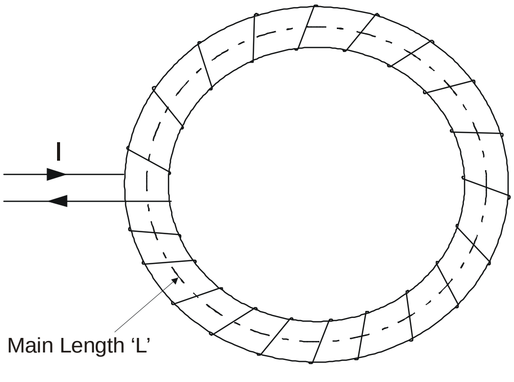
11. In an electric circuit a current is established due to the existence of an electromotive force (emf). Similarly, in a magnetic circuit, a magnetic flux is established due to the existence of a magnetomotive force (mmf). The mmf is produced by the current in the coil and its magnitude depends on the number of turns of the coil (n) and the current (I). The unit of mmf is the ampere-turn (At), however, as the number of turns has no units, the alternative unit for mmf is the ampere (A). Magnetomotive force is represented by the symbol 'Fm'. Therefore:

$$\text{mmf} = \text{Fm} = \text{nI}$$

12. An alternative quantity to express the magnetic force produced by a current is magnetic field strength. This is the magnetomotive force per unit length of the magnetic circuit and is given by the symbol H. In the magnetic circuit shown in Fig 8, where a coil of n turns is uniformly wound around a metal ring, and L is the length of the magnetic circuit, the magnetic field strength is given by:

$$H = \frac{\text{Fm}}{\text{L}} = \frac{\text{nI}}{\text{L}} \text{ amperes per metre (A/m)}$$

14-2 Fig 8 A Simple Magnetic Circuit



Permeability

13. The magnetic flux density (B) inside a current carrying coil is related to the magnetic field strength (H), since one exists as a result of the other. The ratio of these two quantities is known as permeability and may be expressed as the ease with which a magnetic flux is set up within a magnetic circuit. Permeability is represented by the symbol μ (mu) and the unit of measurement is the henry per metre (H/m). The henry is the unit of inductance and is considered in Volume 14, Chapter 3.

$$\text{Permeability } (\mu) = \frac{\text{magnetic flux density (B)}}{\text{magnetic field strength (H)}}$$

14. The magnitude of the permeability is dependent on the material placed in the centre of the coil. If the coil is placed in a vacuum the permeability is then referred to as the permeability of free space and is given the symbol μ_0. This is a constant and has the value:

$$\mu_0 = 4\pi \times 10^{-7} \text{ H/m}$$

15. If a bar of magnetic material is placed in the centre of the coil then the flux density for the same magnetic field strength is greatly increased. Under these conditions the ratio of the flux density produced by the material to the flux density in a vacuum (or air) is called the relative permeability. This is a measure of the number of times the permeability of the material is greater than that of air. Relative permeability has the symbol μ_r, and has no units, since it is a ratio of like quantities.

$$\text{Relative Permeability } (\mu_r) = \frac{\text{Flux density with magnetic core}}{\text{Flux density with air/vacuum core}}$$

The absolute permeability (μ) of the material is given by:

$$\mu = \mu_0\,\mu_r = \frac{B}{H}$$

Reluctance

16. The reluctance of a magnetic circuit is its opposition to the establishment of magnetic flux, and it may be linked, by analogy, to the resistance of an electrical circuit. Reluctance has the symbol 'S' and is the ratio of mmf (Fm) to flux (Φ). Reluctance has the unit ampere per weber (A/Wb).

$$\text{Reluctance (S)} = \frac{\text{mmf (Fm)}}{\text{flux } (\Phi)}$$

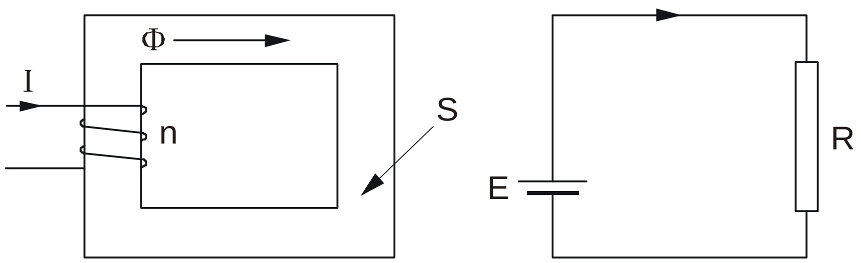
17. In Fig 9, a comparison is made between the magnetic circuit and the electric circuit. This analogy is a useful one when performing calculations on magnetic circuits, but it should be noted that, although flux may be compared to current, it does not flow. It simply exists as a result of the mmf. In the electric circuit, the emf (E) causes the current (I) to flow through the resistance (R), giving the relationship of:

$$E = IR$$

In the magnetic circuit, it is the mmf (Fm) which causes the flux (Φ) to exist in the magnetic circuit of reluctance (S). The relationship is:

$$Fm = \Phi S$$

14-2 Fig 9 Comparison between a Magnetic and an Electric Circuit



Series and Parallel Magnetic Circuits

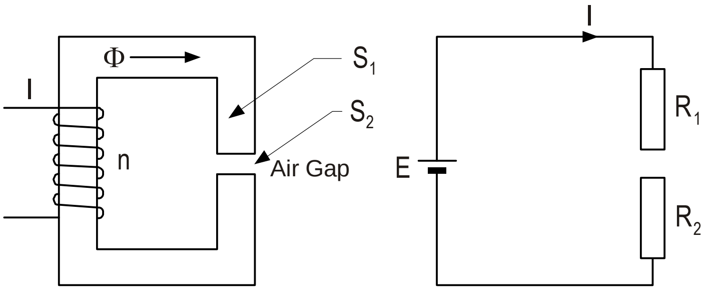
18. Fig 10 shows a simple series magnetic circuit consisting of a metal ring with an air gap. Neglecting losses, it can be assumed that the same flux (Φ) exists both in the iron ring and the air gap. The air gap has a large effect on the magnitude of the flux that exists in the circuit. If S_1 is the reluctance of the metal ring and S_2 the reluctance of the air gap, then the magneto-motive force required to maintain the flux is given by:

$$Fm = \Phi S_1 + \Phi S_2$$

This may be compared with the series electric circuit shown in the same figure.

$$E = IR_1 + IR_2$$

14-2 Fig 10 Series Magnetic Circuit



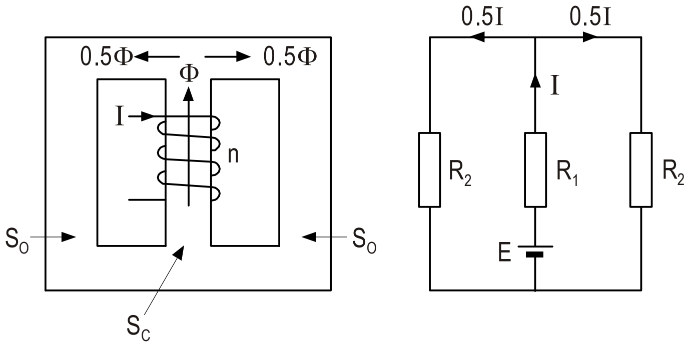
19. Fig 11 shows a parallel magnetic circuit such as is used in a certain type of transformer. The flux (Φ) which exists in the centre limb splits evenly between the two outer limbs so that each carries half of the flux, 0.5Φ. If S_C is the reluctance of the centre limb and S_O the reluctance of each outer limb, then, by considering only one of the magnetic circuits, the mmf is given by:

$$Fm = \Phi S_C + 0.5\Phi S_O$$

This equation can be compared with that of the parallel electrical circuit shown in the same figure. This time applying Kirchhoff's law and considering one loop:

$$E = IR_1 + 0.5IR_2$$

14-2 Fig 11 A Parallel Magnetic Circuit



Ferromagnetism

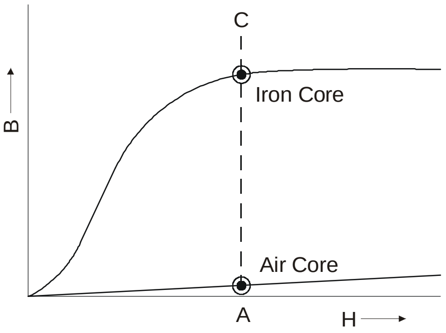
20. As stated earlier in this chapter, certain materials when magnetized produce their own magnetic field which greatly aids the applied field. These materials are known as ferromagnetic and have the ability to increase the magnetic flux density by as much as 1,000 times.

21. Ferromagnetic materials are classified as either hard or soft. Hard materials, such as cobalt steel and certain alloys, have a low relative permeability and are difficult to magnetize and demagnetize. On the other hand, soft materials, such as silicon and soft iron, have high relative permeability and are easily magnetized and demagnetized. Permanent magnets are made from materials with hard ferromagnetic properties, and areas of equipment where magnetic influences need to be minimized are usually protected using soft materials. This latter ferromagnetic feature is known as screening.

Magnetic Hysteresis

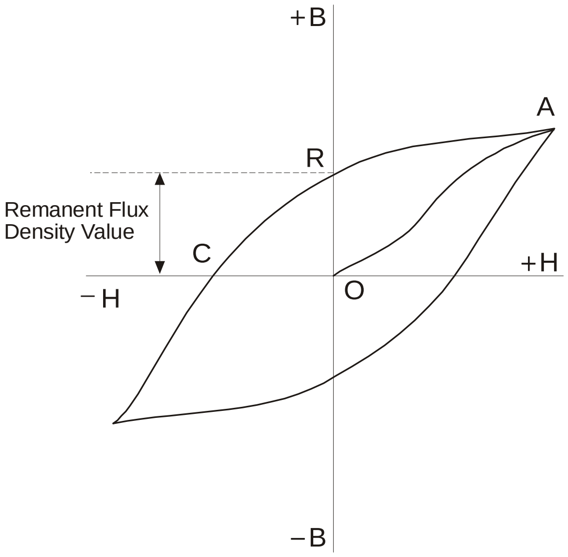
22. The usefulness of a piece of ferromagnetic material depends on its particular magnetic properties, and these can be recorded on a magnetization curve. This is a graph of flux density (B) plotted against magnetic field strength (H). A typical magnetization curve for a ferromagnetic material is shown in Fig 12; the curve for air is also shown for comparison.

14-2 Fig 12 Magnetization Curves



23. It is evident that the slope of the magnetization curve for the ferromagnet is not constant and, at the point C, increasing the magnetizing force has little effect on flux density; this is known as saturation. If the magnetizing force is now reduced to zero the material will retain some of its magnetization, which is referred to as the remanent flux density, or remanence. It is represented by OR in Fig 13.

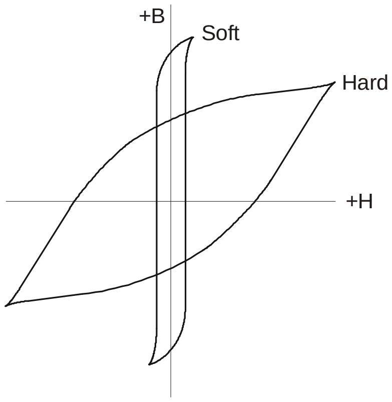
14-2 Fig 13 Magnetic Hysteresis Loop



24. If the magnetic field strength is reversed (–H), then the flux density will eventually reach zero. This field, represented by OC, is known as the coercive force or coercivity of the material. Increasing the magnetic field strength still further, in the negative sense, magnetizes the material in the opposite direction.

25. When the magnetic field strength is increased and decreased, first in one direction and then in the other, the plotted values of B and H describe a loop which is referred to as the hysteresis loop of the material. The area under the curve is proportional to the work carried out in order to complete the cycle and as such represents losses. Fig 14 illustrates the differences in the hysteresis loops for hard and soft ferromagnetic materials.

14-2 Fig 14 Hysteresis Loops for Hard and Soft Materials

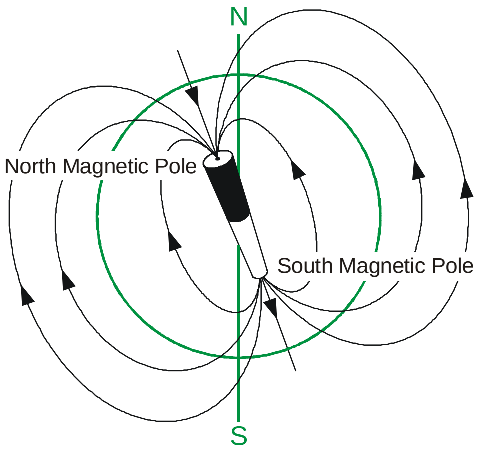


TERRESTRIAL MAGNETISM

Introduction

26. The centre of the Earth is made up of a core of magnetic material whose magnetic field of influence extends beyond the surface of the planet. The core is aligned in such a way that its magnetic poles correspond approximately to the Earth's geographical poles. In essence, the Earth can be regarded as being similar to a large bar magnet with its lines of force running from geographical South to geographical North. This field is depicted in Fig 15.

14-2 Fig 15 The Earth's Magnetic Field



Magnetic Field Components

27. At any point on the Earth’s surface, the magnetic field can be resolved into its vertical and horizontal components. The horizontal component is used extensively for navigation and direction finding by means of a compass. The angle formed between the horizontal and the actual force line is referred to as the angle of dip. The angle of dip varies between 0º at the magnetic equator and 90º at the magnetic poles.

Magnetic Variations and Anomalies

28. The Earth’s North and South magnetic poles are not coincident with the geographical poles; there is an angle between the true North and the magnetic North. The angle between them, known as variation, is a variable quantity and describes a circle around the geographical poles roughly every 1,000 years. In England, the variation is seen as a swing from about 27º W to 27º E. This long-term movement in magnetic variation is modulated by daily changes (Diurnal), annual fluctuations (opposite directions in northern and southern hemispheres) and changes which coincide with 11-year sunspot cycle, (Periodic).

29. The Earth’s magnetic field can also be disturbed by metallic objects placed within its field. In the same way that an iron rod distorts the magnetic field of a bar magnet due to induced magnetism, underground cables and shipwrecks distort the Earth’s magnetic field. These disturbances, or anomalies, can be detected using an instrument known as a magnetometer. The magnetometer principle is dealt with in Volume 14, Chapter 5.

CHAPTER 3 - ELECTROMAGNETIC INDUCTION AND INDUCTANCE

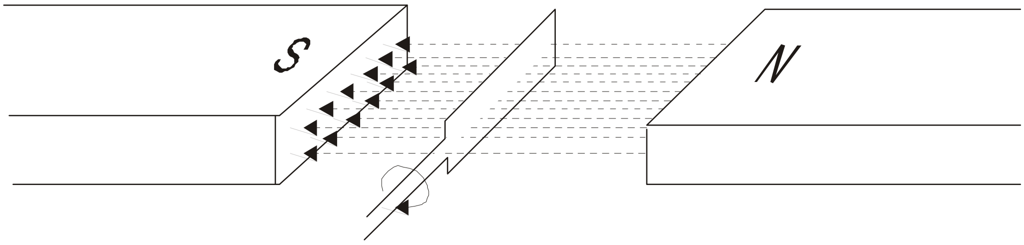
Introduction

1. The phenomenon of electromagnetic induction forms the basis of many electrical machines, such as the dc motor, dc generator, induction motor, synchronous motor and the transformer. Since these are the most frequently used machines in electrical engineering, it is important that the laws of magnetic induction and inductance are clearly understood.

Magnetism as a Source of Electricity

2. When a magnetic field is moved across a conductor, or conversely when a conductor is moved across a stationary magnetic field, an emf is induced in the conductor. This can be demonstrated by moving a magnet into and out of a coil connected to a sensitive ammeter (a device to measure current), or by rotating a loop connected to an ammeter within a magnetic field so as to cut across the field (Fig 1). In both cases relative movement of the conductor and the magnetic field will produce a reading on the ammeter, showing that a current has been created as a result of the induced emf.

14-3 Fig 1 A Simple Generator



3. It can be demonstrated that:

a. The emf is present only whilst there is relative motion between the conductor and the magnetic field.

b. The greater the relative motion, the greater the emf. Thus the rate at which the flux is changing relative to the conductor determines the magnitude of the induced emf.

c. The stronger the magnetic field, the greater the induced emf.

d. The relative motion must be such that the conductor cuts across the magnetic flux.

e. Reversing the relative motion will reverse the direction of the induced emf.

4. To obtain a large voltage, a coil with a large number of turns must be used, and the magnetic field must be a strong one with a continuous high relative motion. One way to obtain rapid continuous relative motion is to rotate a loop in the magnetic field, as in Fig 1, which represents a simple electricity generator. In practice many turns are used in place of a single loop.

Electromagnetic Induction

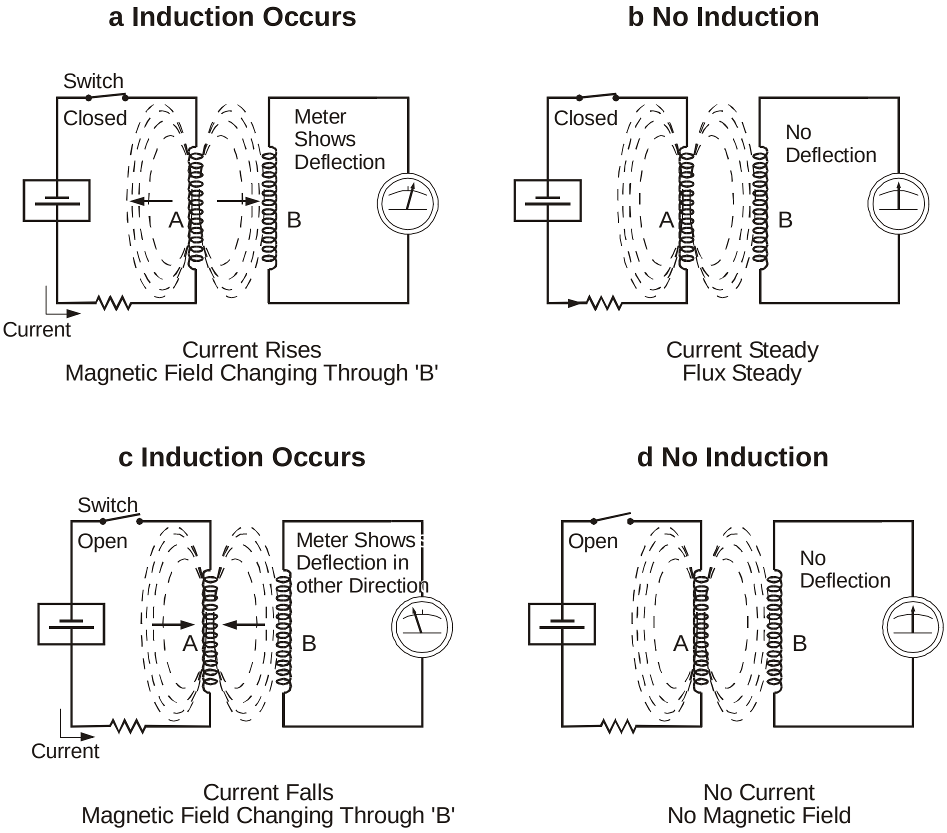
5. Two laws state the theory of electro-magnetic induction very concisely:

a. **Faraday's Law**. When the magnetic flux through a circuit is changing an induced emf is set up in that circuit, and its magnitude is proportional to the rate of change of flux.

b. **Lenz's Law**. The direction of an induced emf is such that its effect tends to oppose the change producing it.

6. **Mutual Induction**. Consider the two coils, A and B, connected as shown in Fig 2. When no current is flowing there is no magnetic field, but when the switch is closed a magnetic field expands outwards from the primary coil (A) and cuts the secondary coil (B). As there is a change of flux through the secondary coil an emf is induced in it and the meter shows a deflection. When the field stops growing there is no longer a change of flux through the secondary and no emf is induced. When the switch is opened the primary field collapses and contracts, thus flux changes in the secondary and an emf is induced again, but in the opposite direction. It is important to note that when the DC supply is steady no mutual induction occurs, but when the supply is AC induction occurs continuously.

14-3 Fig 2 Mutual Induction



a Induction Occurs

Current Rises
Magnetic Field Changing Through 'B'

b No Induction

Current Steady
Flux Steady

c Induction Occurs

Current Falls
Magnetic Field Changing Through 'B'

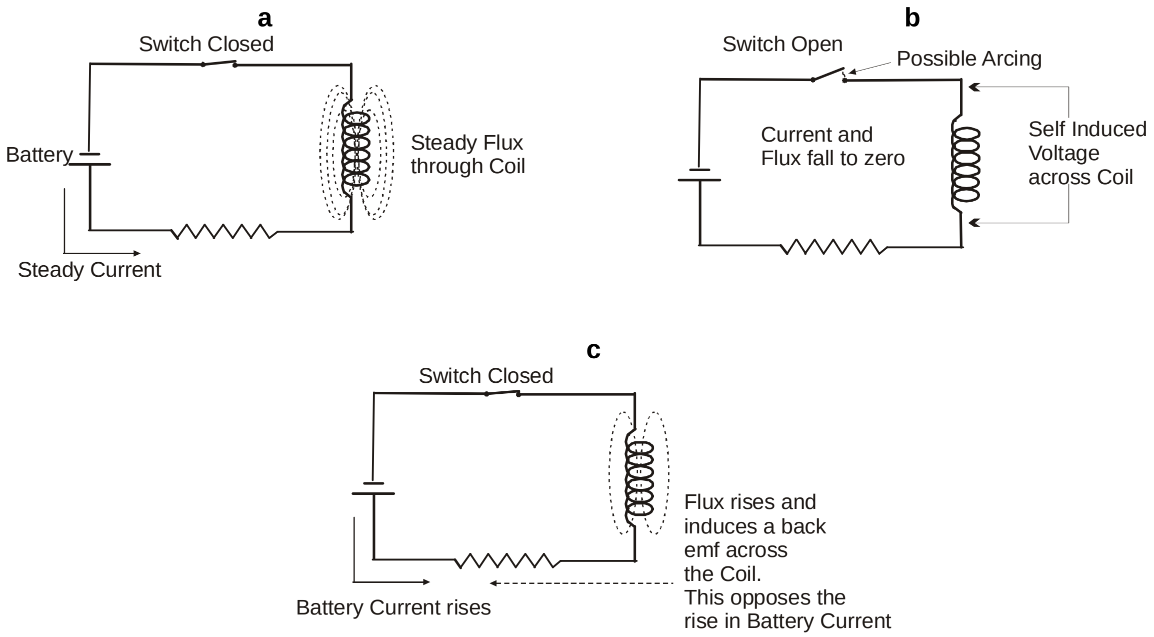
d No Induction

No Current
No Magnetic Field

7. **Mutual Inductance (M)**. The size of the induced emf depends partly on the rate at which the flux is changing. It also depends on various factors associated with the coils, these being the ratio of the number of turns of the primary and secondary coils, the relative positions of the coils, and the permeability of the medium through which the flux travels; together these are known as the mutual inductance of the coils (M).

8. **Self-Induction**. In the circuit shown at Fig 3 a magnetic field is established around a coil. When the switch is opened the field will start to collapse as the current reduces. There is therefore a change of flux through the coil, and an emf will be induced in it. The induced emf opposes the inducing emf (Lenz's Law), and thus is called the back emf. No back emf is created by a steady DC supply except at switch on and switch off, whereas an AC supply causes self-induction continuously.

14-3 Fig 3 Self-Induction


a
Switch Closed
Battery
Steady Flux through Coil
Steady Current
b
Switch Open
Possible Arcing
Current and Flux fall to zero
Self Induced Voltage across Coil
c
Switch Closed
Flux rises and induces a back emf across the Coil.
This opposes the rise in Battery Current
Battery Current rises

9. **Self-Inductance (L)**. Any circuit that has a voltage induced in it by a change of current through the circuit itself has self-inductance. The property of self-inductance is to oppose any change of current by the creation of a back emf. The size of the back emf depends partly upon the rate of change of current. If the circuit includes a coil, the size of the back emf depends also on the number of turns in the coil and the permeability of the medium around the coil, these factors associated with the coil being termed the self-inductance (L).

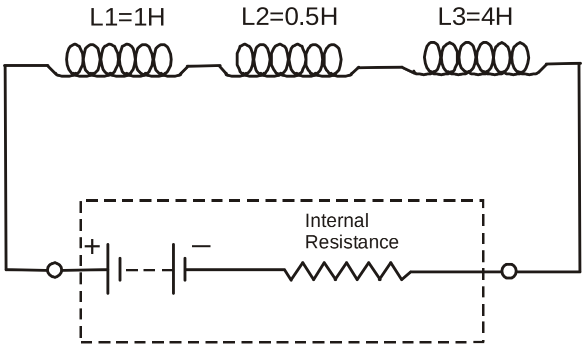
10. **Units of Inductance**. Both mutual and self-inductance are measured in the same units - the henry (H). The mutual inductance of two circuits is 1 henry when a current changing uniformly at the rate of 1 ampere per second in one circuit produces a mutually induced emf of 1 volt in the other circuit. The self-inductance of a closed circuit is 1 henry when an emf of one volt is produced when the current in the circuit is changing uniformly at the rate of one ampere per second.

11. **Inductors**. When a coil is used specifically to provide inductance and thus oppose any change in current in the circuit it is called an inductor or choke. Typical inductance values of inductors in practical use range from about 100 H down to a few microhenrys according to the application.

Inductive DC Series Circuits

12. When several inductors are connected in series (L1, L2 and L3 in Fig 4) circuit inductance is found by summing the individual values, provided that there is no mutual inductance between the coils.

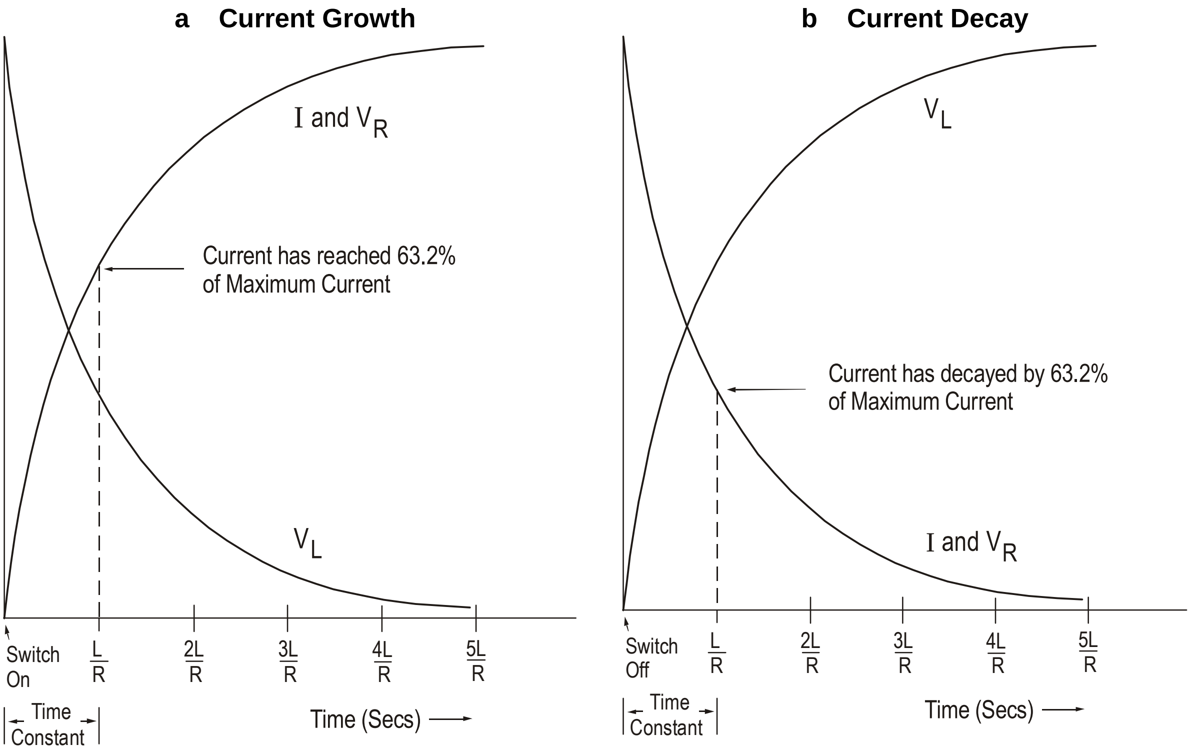
14-3 Fig 4 Inductors in Series



$$\begin{aligned} L &= L1 + L2 + L3 \\ &= 1 + 0.5 + 4 \\ &= 5.5 \text{ henrys} \end{aligned}$$

13. It was stated earlier in this chapter that the back emf created by the build-up or collapse of the magnetic field around an inductor opposes the change in current. A plot of circuit current against time takes the form of an exponential curve, rising for current increase or switch-on (Fig 5a), and falling with current decrease or switch-off (Fig 5b). It is impossible to create a purely inductive circuit because the conductor must have some resistance, and the resistive voltage, V_R, rises and falls as the circuit current rises and falls (Fig 5). On the other hand the voltage, V_L, across the inductor, falls as the current rises and rises as the current falls (Fig 5). The time taken for the current to reach particular values is discussed in para 15.

14-3 Fig 5 Growth and Decay of Current in an Inductive DC Circuit



Inductive DC Parallel Circuits

14. When inductors are connected in parallel (Fig 6), providing alternative paths for the current flow, the total inductance of the circuit, provided that the coils are completely isolated from one another such that there is no mutual inductance, is given by:

$$\frac{1}{L} = \frac{1}{L_1} + \frac{1}{L_2} + \frac{1}{L_3} + \ldots\ldots\ldots\ldots\ldots\ldots \frac{1}{L_n}$$

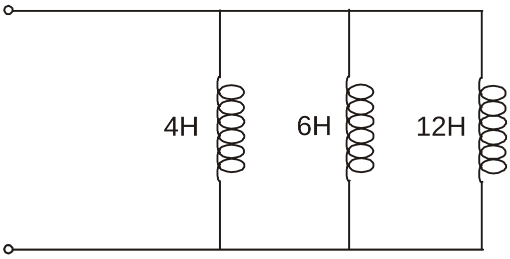
Applying this formula to the example in Fig 6:

$$\frac{1}{L} = \frac{1}{4} + \frac{1}{6} + \frac{1}{12}$$

$$= \frac{6}{12}$$

$$\therefore L = 2H$$

14-3 Fig 6 Inductors in Parallel



Inductive-Resistive Circuits, Time Constants

15. The exponential rise and fall of current in an inductive-resistive DC circuit was discussed in para 13. In theory the current in an inductive circuit can never reach a maximum or fall completely to zero. However, for all practical purposes the growth or decay of current is complete in $5\frac{L}{R}$ seconds, where L is the inductance in henrys and R the resistance in ohms. $\frac{L}{R}$ seconds is known as the time constant of an inductive-resistive circuit. The time constant is defined as the time taken for the current through an inductive circuit to rise to 63.2% of its maximum value when connected to a supply, or to fall by 63.2% of its maximum value when disconnected from a supply. The time constant is indicated in Figs 5a and 5b.

CHAPTER 4 - ELECTROSTATICS AND CAPACITANCE

Introduction

1. This chapter opens with a brief mention of electrostatics, which, as the name implies, is primarily the science of electric charges at rest. It then goes on to discuss a related phenomenon known as 'capacitance', and describes the construction and properties of devices used to introduce capacitance into the circuit ('capacitors').

ELECTROSTATICS

General

2. Electrification by friction is a common phenomenon: a comb passed through dry hair attracts the individual hairs, which then tend to stand on end, repelling one another. If a glass rod is rubbed with a piece of silk, the silk is attracted towards the rod. In this case, the silk removes electrons from the glass which is thus left with a positive charge; the electrons acquired by the silk give it an equal negative charge. If two freely suspended glass rods are treated in this way, and then brought near to each other, a mutual repulsion is evident. From these facts, the first law of electrostatics can be stated: like charges repel each other; unlike charges attract.

Coulomb’s Law

3. It can be shown that the size of the force of attraction (or repulsion) is greater between large charges than between small charges, and is greater when the charges are close together than when they are more distant. Coulomb’s law states that the force (F) between two quantities of electricity (Q_1 and Q_2), placed a distance apart (d), is proportional to the product Q_1Q_2 and inversely proportional to d^2, ie $F \propto Q_1Q_2/d^2$.

The Electric Field

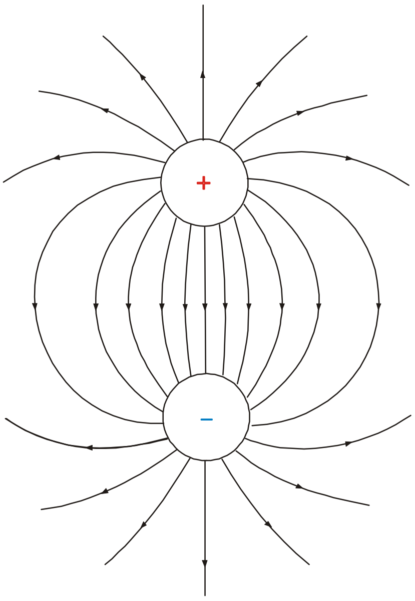
4. If an insulated metal sphere (A) is positively charged, it will repel another positively charged body (B) brought near to it; the nearer the two approach, the greater the repulsion. Work has to be done to overcome this repulsion; if the force drawing the bodies together is removed, the repulsion will force B away towards its original position (assuming A to be fixed). The system possesses potential energy, and a potential is said to exist at any point in the vicinity of A; the nearer to A, the higher the potential. A positive charge placed in the vicinity of A will tend to move away, ie from a position of higher to lower potential. The change of potential per unit distance is known as potential gradient.

5. The region around a charged body is described as an electric field. Electric lines of force (akin to lines of magnetic flux) are used to show the distribution of an electric field, each line showing the direction in which a free positive charge would tend to move if placed in that field. Fig 1 shows a typical field pattern of two oppositely charged spheres. The unit of electric flux, or charge (Q), is the coulomb, and the flux density (D) is the charge per unit area:

$$\text{Flux Density} = \frac{\text{Charge (coulombs)}}{\text{Area (square metres)}}$$

$$\text{or } D = \frac{Q}{A}$$

14-4 Fig 1 Typical Electric Field Pattern

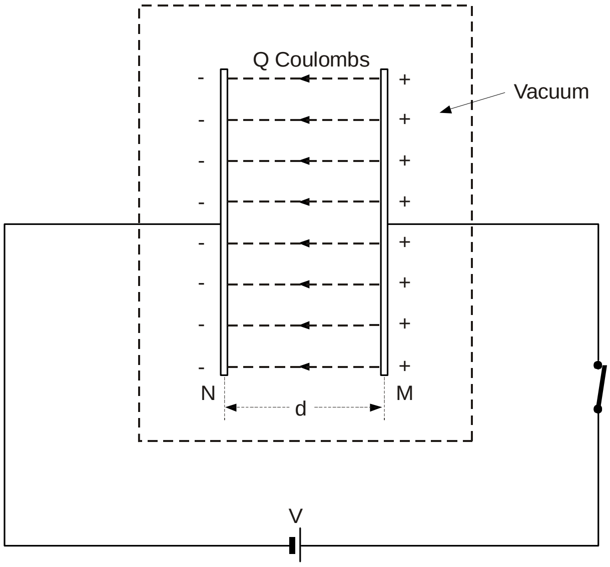


6. **Electric Field Strength**. Two plates (M and N), separated by a gap (d), are connected to a battery as shown in Fig 2. Electrons begin to move from plate M towards the positive terminal of the battery, leaving the plate M positively charged. At the same time, electrons move from the negative plate of the battery to the plate N, which acquires a negative charge. A difference of potential will exist between the plates, and the space between them is an electric field. The intensity (E) of the electric field will depend on the PD and the distance between the plates:

$$\text{Intensity} = \frac{\text{Potential Difference (volts)}}{\text{Distance Apart (metres)}}$$

$$\text{or } E = \frac{V}{d}$$

14-4 Fig 2 The Electric Field Between Two Plates



7. **Permittivity of Free Space**. The ratio of flux density (D) to the electric field strength (E) in free space is termed the 'permittivity of free space' (ε_0). Thus:

$$\varepsilon_0 = \frac{D}{E} = \frac{Q}{A} \div \frac{V}{d} = \frac{Qd}{VA}$$

But, $\frac{Q}{V} = C$ (see para 11).

$$\therefore \varepsilon_0 = C\frac{d}{A} \quad \text{and } C = \varepsilon_0 \frac{A}{d} \text{ farads}$$

The value of the constant ε_0 is 8.85×10^{-12} F/m. (F/m: Farad per metre)

CAPACITANCE

The Simple Capacitor

8. To charge a conductor negatively, electrons are added to it. Initially, the electrons move easily but, as the conductor becomes more and more negatively charged, so the electrons on the conductor repel those which try to add to their number. Eventually, the repelling force equals the charging force and no further movement of electrons takes place. The conductor is then said to be fully charged.

9. The charging process is due to the applied voltage. The charge can be increased by increasing the applied voltage, but the conductor will oppose the change because of the repulsion by the electrons it already possesses. Similar remarks apply when a conductor is positively charged.

10. It has already been shown that if, instead of a single conductor, a pair of plates, as in Fig 2, is considered, a difference of potential will exist between the plates if connected to a battery. The charge held by the combination of plates for a given applied voltage can be made greater by sandwiching a suitable insulator (a 'dielectric') between them.

11. For a given pair of plates, increasing the charge increases the voltage between the plates, but the ratio of charge to applied voltage remains constant. This ratio is known as the 'capacitance' (C) of the capacitor. The unit of capacitance is the farad; a capacitor having a capacitance of one farad if a charge of one coulomb (i.e. a charging current of one ampere flowing for one second) causes a charge of one volt in the potential difference between the plates, i.e.:

$$\text{Capacitance (farads)} = \frac{\text{Charge (coulombs)}}{\text{Potential Difference (volts)}}$$

Because the farad is a very large unit, capacitances are usually given in microfarads (μF - a millionth of a farad, ie 10^{-6} F) or picofarads (pF - a million-millionths of a farad, ie 10^{-12} F).

Factors Affecting Capacitance

12. The capacitance of a capacitor depends on the following factors:

a. **The Area of the Plates**. Capacitance increases with increase in area (A) of the opposed surfaces.

b. **The Distance Apart of the Plates**. Capacitance increases as the distance (d) between the plates decreases (ie by using a thinner dielectric), because the field then becomes more concentrated.

c. **The Material of the Dielectric**. Different dielectrics produce different capacitances in capacitors where other factors (plate area and distance between the plates) are equal. For example, substituting waxed paper for air increases the capacitance by a factor of about three. The ratio of the capacitance of a capacitor having a certain material as dielectric to the capacitance of the same capacitor having a vacuum as dielectric, is known as the 'dielectric constant', or 'relative permittivity' (ε_r) of the dielectric material.

13. It follows that the capacitance of a capacitor is given by:

$$C = \varepsilon_0 \varepsilon_r \frac{A}{d} \text{ (farads)}$$

A comparison of dielectric materials is given in Table 1. For practical purposes, air can be considered to have a dielectric constant of unity, as for free space.

Table 1 Dielectric Constants

Dielectric Material	Relative Permittivity
Dry air	1.00
Polypropylene	2.25
Polystyrene	2.50
Polycarbonate	2.80
Polyester	3.20
Impregnated paper	4 to 5
Mica	6.00
Aluminium oxide	7.50
Tantalum oxide	25.00
High-permittivity ceramic	10,000.00

Safe Working Voltage

14. The safe working voltage is the maximum DC voltage that can safely be applied to a capacitor without causing the dielectric to break down; if this voltage is exceeded the electric field becomes strong enough to break down the insulation of the dielectric, a spark occurs, and the capacitor is usually ruined. Increasing the thickness of the dielectric enables the capacitor to withstand higher voltages, but, to compensate for the greater distance between the plates, a larger plate area is necessary to maintain the capacitance value. Hence, capacitors with high voltage ratings are larger than similar types of capacitor of low rating. Note that a charged capacitor (especially one of large capacitance) can be dangerous; capacitors must be discharged before touching them.

Factors Determining the Construction of Capacitors

15. Capacitors are used for a wide variety of purposes in electronics. In their construction, three factors associated with the dielectric are important:

a. The dielectric constant, discussed in paras 12 and 13.

b. The dielectric strength.

c. Dielectric losses.

16. **The Dielectric Strength**. Dielectric strength is a measure of the PD required to break down the dielectric and is normally given as kilovolts per millimetre thickness. For a given material, the breakdown voltage will depend on:

a. **Thickness of the Dielectric**. Although not directly proportional, the thicker the dielectric, the greater the breakdown voltage.

b. **Temperature**. An increase in temperature of the dielectric produces a decrease in the breakdown voltage.

c. **Frequency**. If the applied voltage is AC, the higher the frequency the lower the breakdown voltage due to the higher temperature produced.

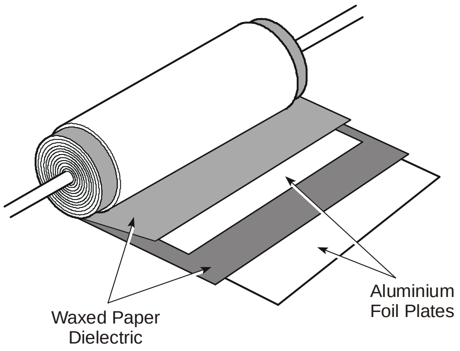
The safe working voltage of a capacitor is usually given as a DC voltage at a given temperature, eg "350V DC working, 71 ºC". When used in an AC circuit, the rms voltage applied should not exceed 0.707 of the safe working DC voltage.

17. **Dielectric Losses**. A certain proportion of the energy applied to a capacitor on charging is not available as energy on discharging. Some energy is expended in the dielectric, being lost as heating, leakage current, and through other effects. The ratio of energy applied on charging to energy available on discharge is a measure of the capacitor efficiency.

Fixed Capacitors

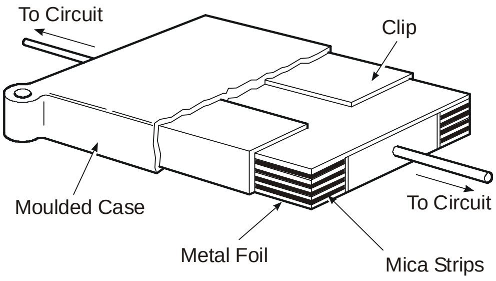
18. **Paper Type**. Capacitors with paper dielectrics are the commonest and cheapest fixed capacitors. The two plates are long strips of aluminium foil, separated by similar strips of waxed paper acting as the dielectric (Fig 3). This assembly is rolled up into a tube, which is then wax-sealed into an outer container of cardboard or metal.

14-4 Fig 3 Paper Capacitor



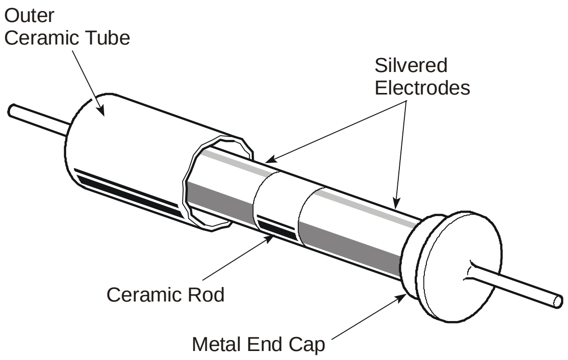
19. **Mica Type**. Mica Capacitors are high quality capacitors with several plates of metal foil interleaved with layers of mica acting as the dielectric (Fig 4). Alternate foils are joined together and connected to the two terminals. The assembly is placed inside a moulded plastic or metal case.

14-4 Fig 4 Mica Capacitor



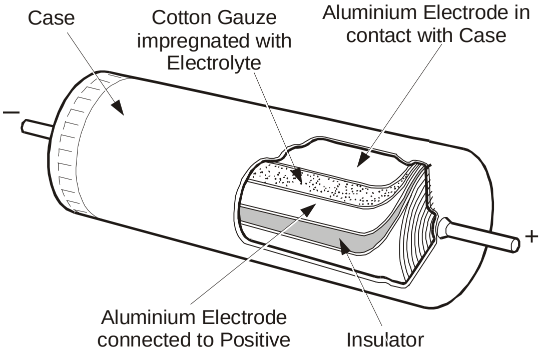
20. **Ceramic Type**. Ceramic capacitors can give very small values of capacitance. They use a ceramic disc or rod as the dielectric. Two films of silver, deposited on the ceramic, act as the plates (Fig 5).

14-4 Fig 5 Ceramic Capacitor



21. **Electrolytic Type**. For large values of capacitance, electrolytic capacitors are used, paper and mica types being too bulky. The large capacitance value is due to the very thin dielectric used. The electrodes are two long strips of aluminium foil. During manufacture, the dielectric film is formed on the positive electrode. The electrodes are separated by strips of cotton gauze impregnated with electrolyte. The whole assembly is tightly rolled (Fig 6) and mounted in a metal case. One of the aluminium electrodes provides a contact between the case and the electrolyte paste, which acts as the negative plate. The dielectric film can only be maintained if the capacitor has a DC component of voltage applied to it. The capacitor must be connected the right way round in the circuit, and any AC component must be less than the polarizing DC component.

14-4 Fig 6 Electrolytic Capacitor


Case
Cotton Gauze impregnated with Electrolyte
Aluminium Electrode in contact with Case
–
+
Aluminium Electrode connected to Positive
Insulator

Variable and Trimmer Capacitors

22. In electronic circuits, it is often necessary to be able to vary the capacitance. Variation of capacitance can be achieved by:

a. **Varying the Effective Area of the Plates**. One set of plates (the rotors) can be moved into and out of mesh with the other set (the stators), so that the effective area of overlap is changed (Fig 7a).

b. **Varying the Distance between the Plates**. The distance between the plates can be varied by a screw adjustment which squeezes the plates together (Fig 7b).

c. **Changing the Dielectric**. If two (or more) metal plates are set at a fixed distance apart in a container, then, when the container is empty, the dielectric is air. If a fluid which has dielectric properties is pumped into the container, then the capacitance will change as the level of fluid changes (Fig 7c). This principle is applied in measuring the contents of storage tanks.

The types described in sub-paras a and b are normally referred to as variable capacitors; a smaller version of that in sub-para a, and the type at sub-para b are used to make small variations of circuit capacitance, and are referred to as trimmer capacitors. The main properties and uses of the various forms of capacitor are summarized at Table 2.

14-4 Fig 7 Variable and Trimmer Capacitors

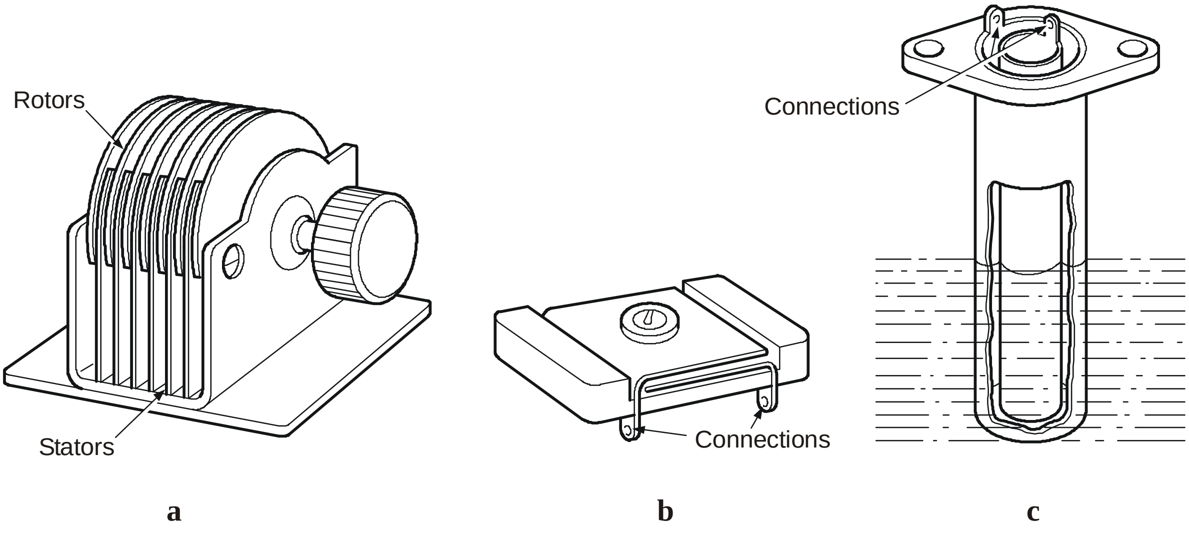


Table 2 Capacitor Summary

Type	Capacitance Values	DC Working Voltage (max)	Remarks
Paper	100 pF to 12 μF	150 kV	Used in circuits where dielectric losses are unimportant. Cheap.
Mica	50 pF to 0.25 μF	2 kV	Used in low-loss circuits. Expensive.
Ceramic	0.5 pF to 0.005 μF	500 V	Used in low-loss precision circuits where micro-miniaturization is important or where temperature compensation is required.
Electrolytic	2 μF to 32 μF	600 V	Used where losses are not important - smoothing circuits.
	32 μF to 3000 μF	50 V	A polarizing DC voltage must be operative in the circuit to prevent reverse electrolytic action.
Variable	Variable from 50 pF to 500 pF	2 kV	Used for circuit tuning.
Trimmer	Variable from 2 pF to 50 pF	350 V	Used for circuit alignment.

Energy Stored in a Charged Capacitor

23. When an initially uncharged capacitor is charged at a constant rate of I amps for t seconds, the charge (Q) equals (I × t) coulombs. During the charge, the PD across the capacitor will have risen from zero to V volts at a constant rate. Thus, the average PD during the charge is $\frac{V}{2}$ volts. The average power is the product of average PD and the current, i.e. $P = \frac{V}{2} \times I$ watts. The energy used to charge the capacitor is stored in the charged capacitor and is given by:

$$\text{Energy} = P \times t = \frac{V}{2} \times I \times t$$

$$= \frac{V}{2} \times Q$$

$$= \frac{1}{2} CV^2 \text{ joules}$$

Capacitive DC Series Circuits

24. When capacitors are connected in series (Fig 8), the inverse individual values are summed and the reciprocal found to determine the total capacitance of the circuit, i.e.:

$$\frac{1}{C} = \frac{1}{C_1} + \frac{1}{C_2} + \frac{1}{C_3} + \frac{1}{C_4}$$

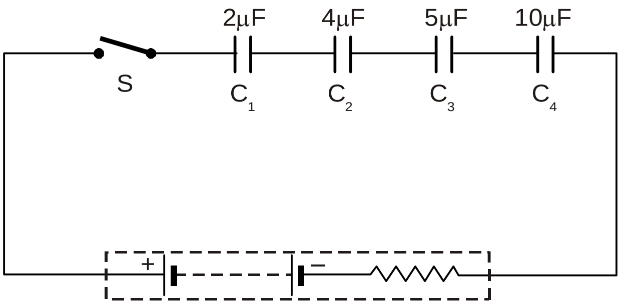
Substituting the values from Fig 8,

$$\frac{1}{C} = \frac{1}{2} + \frac{1}{4} + \frac{1}{5} + \frac{1}{10} = \frac{10}{20} + \frac{5}{20} + \frac{4}{20} + \frac{2}{20} = \frac{21}{20}$$

$$\therefore C \approx 1\,\mu F$$

The effect of the series connections is to create a 'total capacitor' of large effective distance (d) between its plates. In Fig 8, the resistor represents the circuit resistance.

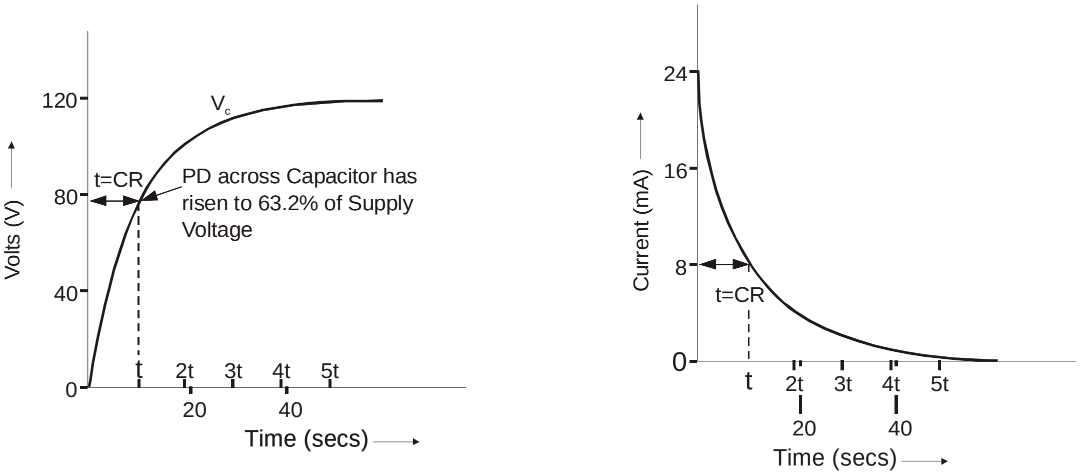
14-4 Fig 8 Capacitors in Series



25. When only the resistance of connectors is present, the time taken to charge the capacitors is negligible, but becomes significant if a large resistor is connected in series. This is discussed further in

para 27. Fig 9a shows a plot of the PD (V_c) across the capacitor against time; the equivalent current curve shown at Fig 9b is inverted with respect to the voltage curve.

14-4 Fig 9 Voltage and Current during Charges in a Capacitive DC Circuit



a Variation of V_c (120V Supply)

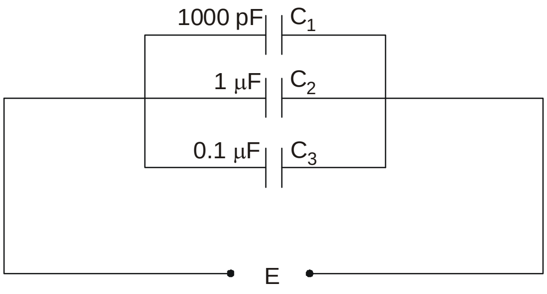
b Variation of Current

Capacitive DC Parallel Circuits

26. The total capacitance of several capacitors in parallel (Fig 10) is the sum of their individual values, ie C = $C_1 + C_2 .... + C_n$. Substituting the values from Fig 10:

C = 0.001 + 1 + 0.1 = 1.101 μF

14-4 Fig 10 Capacitors in Parallel



The capacitance is increased because the effective plate area is increased.

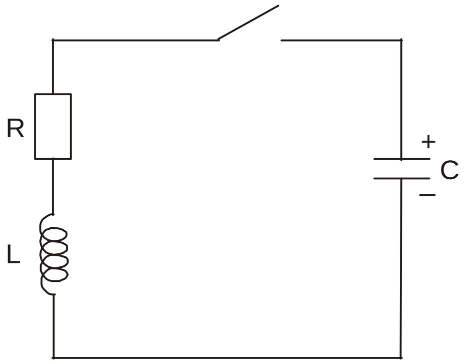
Capacitive-Resistive Circuits, Time Constants

27. The time required to charge a capacitor depends upon both its capacitance and the circuit resistance. The time (t) is equal to CR seconds (where C is in farads and R in ohms) and is known as the 'time constant' of a capacitive-resistive circuit. The time constant is defined as the time taken for the PD across a capacitor to rise to 63.2% of the maximum value of the supply voltage on charge, or to fall by 63.2% of its fully charged PD when on discharge. The time constant (t) is indicated in Figs 9a and 9b. In theory, a capacitor would take an infinitely long time to completely charge or discharge, but after a time of 5t seconds (ie 5CR) the charge or discharge is complete for all practical purposes.

The Oscillatory Circuit

28. In the electrical circuit shown in Fig 11, a charged capacitor is connected in series with an inductor of inductance L and resistance R. Closing the switch causes the capacitor to discharge and current flows in the circuit. This current flow generates a magnetic field within the coils of the inductor, and the electrostatic energy stored in the capacitor is transferred to the inductor in the form of electromagnetic energy.

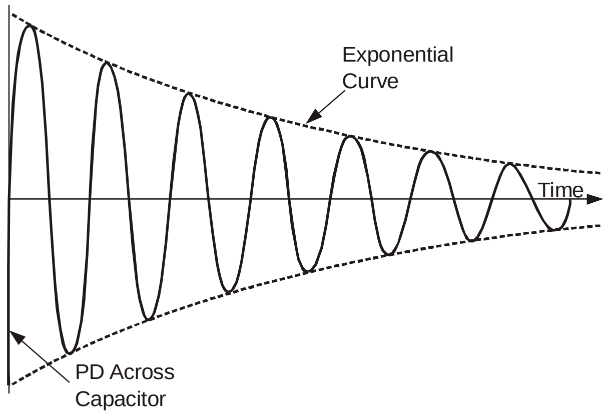
14-4 Fig 11 The Oscillatory Circuit



29. Once the capacitor is completely discharged, the magnetic field within the inductor starts to collapse and the collapsing (changing) flux tends to keep the current flowing. The capacitor finally charges up once again, but this time with the reverse polarity to the original charge. When the current has fallen to zero, the capacitor begins to discharge again and the same sequence of events is repeated. This exchange of energy between two parts of the circuit is called 'oscillation', and the period between two successive instants at which the conditions in the circuit are similar is called a 'cycle'.

30. Since the circuit contains resistance, a small amount of energy is dissipated in the form of heat during each cycle. Therefore, each successive charge across the capacitor is smaller in magnitude than the previous one. This effect is known as 'damping' and the decay in amplitude due to damping follows an exponential curve. This is depicted graphically in Fig 12.

14-4 Fig 12 Oscillatory Damping due to Circuit Resistance



31. For any given oscillatory circuit, there is a value of resistance which will prevent oscillations taking place. This resistance, known as 'critical damping resistance', is given by the following equation:

$$\text{Critical Damping (R)} = 2\sqrt{\frac{L}{C}}$$

If R equals or exceeds this value, then the nature of the discharge of the capacitor is that of a simple CR combination referred to in para 26. If R is much lower than the critical damping value, then the frequency of oscillation is approximately:

$$\text{Resonant Frequency } (f_r) = \frac{1}{2\pi\sqrt{LC}}$$

CHAPTER 5 - MEASURING INSTRUMENTS

Introduction

1. In all fields of science and engineering instruments are required in order to measure the quantities being used; electrical engineering is no exception to this. Associated with most electrical circuits are the basic quantities of voltage, current, power and resistance. The instruments which measure these are the voltmeter, ammeter, wattmeter, and ohmmeter respectively.

CURRENT AND VOLTAGE

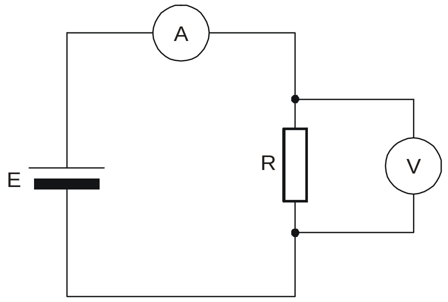
Ammeters and Voltmeters

2. An ammeter is used to measure current flow through a circuit and therefore the instrument must be connected in series with the circuit, as shown in Fig 1. It is important that the current being measured is unaffected by the instrument's internal resistance, and for this reason ammeters are designed to have a low electrical resistance. This being the case, an ammeter should never be connected across a supply voltage, since there would be nothing to limit the current flow and the instrument would be damaged.

3. A voltmeter is designed to measure electromotive force and potential difference. The potential difference across a resistor is measured by connecting the meter as shown in Fig 1. In order to ensure the minimum disturbance in any circuit voltmeters are required to pass very little current and therefore have a high internal resistance.

4. There are a number of ways that an electric current or voltage can cause the needle of a measuring instrument to be deflected. The most common method utilizes the magnetic field of an electric current to produce the deflection. In this case it is important to realize that regardless of the units being measured the principle of operation remains the same; ie current through the instrument causes needle deflection. The internal resistance, high or low relative to the circuit under test, determines whether the device is a voltmeter or an ammeter.

14-5 Fig 1 Ammeter and Voltmeter Connections in a Circuit



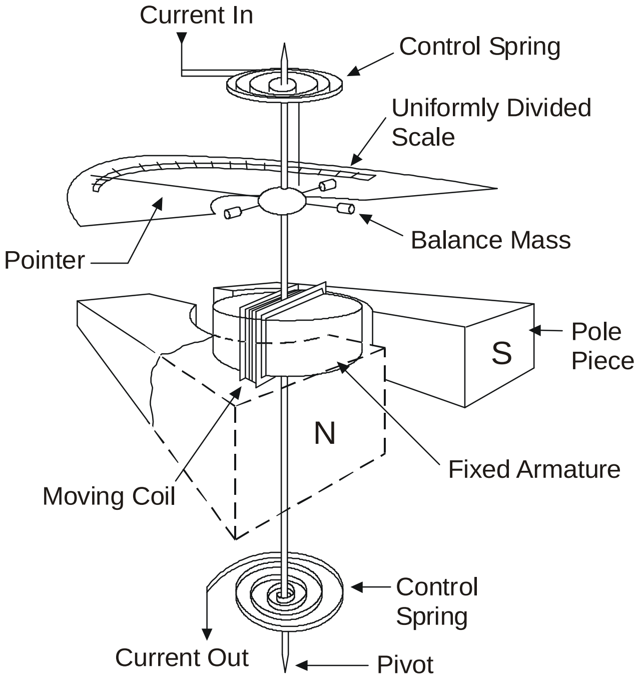
Moving Coil Instruments

5. The operation of a moving coil instrument depends on the reaction between the current in a moveable coil and the field of a fixed permanent magnet. This reaction is discussed in more detail in Volume 14, Chapter 6, under the topic of DC motors.

6. The moving coil is wound around an aluminium former and suspended between the curved poles of a magnet, as shown in Fig 2. The current to be measured is fed through the moveable coil via two

coiled springs. A magnetic field is set up around the moveable coil which reacts with that of the magnet and produces a turning force (torque). The coil will turn and apply tension to the coiled springs until the restoring torque of the springs is equal to the force produced by the coil. The amount of angular movement is proportional to the applied current and is indicated by a pointer on a scale.

14-5 Fig 2 A Simple Moving Coil Instrument



7. In order to prevent the pointer overshooting and then oscillating before settling down at the correct reading, damping of the movement must be provided. This is achieved through the aluminium former supporting the coil. When the coil moves through the magnetic field, small circulating currents (eddy currents) are set up in the former which produce a restraining torque while the coil is in motion and thus damp out oscillation.

8. The main features of a moving coil instrument are:

a. It measures DC only.

b. A linear scale.

c. High sensitivity.

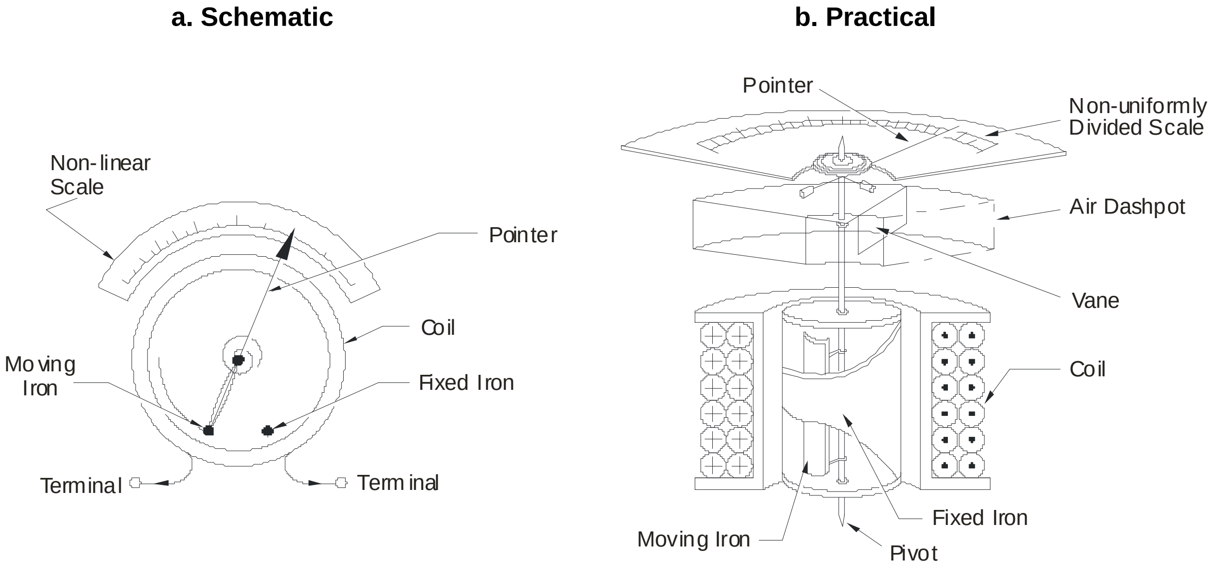
d. It may be used as a DC ammeter with a shunt (para 11).

e. It may be used as a DC voltmeter with a multiplier (para 12).

Moving Iron Instruments

9. The repulsion moving iron instrument is basically two pieces of iron placed inside a coil, as shown in Fig 3a. One of the pieces is fixed while the other is attached to a pointer and free to move. When current is passed through the coil, a magnetic field is produced which magnetizes both irons in the same sense. Since like poles repel each other, the free iron is forced away from the fixed iron. As in the case of the moving coil instrument, a coiled spring at the pivot point of the pointer provides the restoring torque.

14-5 Fig 3 A Moving Iron Instrument



a. Schematic

b. Practical

10. In the practical instrument, Fig 3b, the moving iron is rectangular in shape, while the fixed iron is in the shape of a tapered scroll. This is done to improve the scale linearity due to the non-linear forces between two irons of similar shape. Instrument damping is achieved by means of a vane and air dashpot. The main features of this type of instrument are:

a. It measures DC and AC.

b. A non-linear scale.

c. It may be used as an ammeter, but shunts should not be used, due to their non-linearity.

d. It may be used as a voltmeter, directly or with multipliers.

e. It is accurate only for frequencies below about 200 Hz.

Shunts and Multipliers

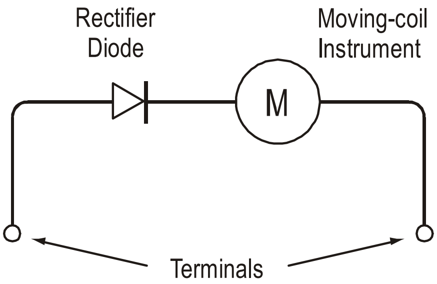
11. In general, meters are constructed to have a high sensitivity and therefore give full-scale deflection (fsd) of the pointer with small currents of the order of a few micro-amps. In order to measure currents of a higher magnitude, resistors, known as shunts, are connected across the meter in order to by-pass the excess current. The meter may now be calibrated to read an apparent fsd which is equal to the inherent fsd plus the excess current through the resistor.

12. If a meter is required to read voltages then a resistor has to be connected in series with the instrument across which most of the applied voltage is developed. The voltage across the instrument itself is small and the current flowing through it is still only of the order of micro-amps. This type of resistor is referred to as a multiplier.

Rectification for AC Measurements

13. A rectifier is a device which allows current to flow in one direction only, very much like a non-return valve in a fluid system. When such a device is connected in series with a moving coil instrument, as shown in Fig 4, a rectifier allows the instrument to be used for the measurement of AC voltages and currents.

14-5 Fig 4 Rectifier and Moving Coil Instrument



14. The instrument indicates the average value of the current flowing through it, and thus gives a reading proportional to the average value of the half-wave rectified waveform. Meters of this type give accurate readings for AC voltages and currents up to about 20 kHz, but at low frequencies the pointer tends to fluctuate.

RESISTANCE

Volt-amp Method

15. The value of a resistor (R) may be determined by voltage (V) and current (I) measurements using instruments arranged as shown in Fig 1. By applying Ohm's law to the values obtained the resistance may be calculated.

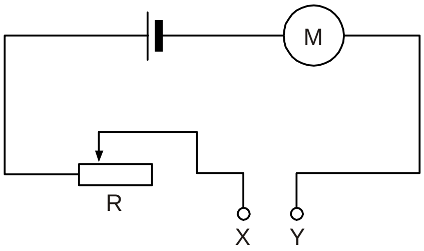
$$R = \frac{V}{I} \text{ ohms}$$

The Ohmmeter

16. The disadvantage of the volt-amp method of determining resistance is the need for two meters, a voltmeter and an ammeter. This problem is overcome with the introduction of a single meter known as the ohmmeter. Basically, the ohmmeter uses the same movement as the voltmeter and the ammeter, but the zero reading is at the opposite end of the scale.

17. The circuit of a basic ohmmeter is shown in Fig 5. It comprises a moving coil ammeter (M), a variable resistor (R) and a battery. Prior to any measurement being taken the terminals X and Y are shorted together and the variable resistor adjusted to give full-scale deflection of the needle; this corresponds to zero resistance. Any resistance placed between the two terminals will cause less current to flow in the circuit and result in less deflection of the needle. The amount of deflection is therefore inversely proportional to the value of the resistance.

14-5 Fig 5 A Simple Ohmmeter



The Wheatstone Bridge Method

18. Bridge methods of measuring component values are used where high accuracy is most important. The most straightforward bridge circuit is the Wheatstone bridge. It is less convenient than an ohmmeter, but far more accurate.

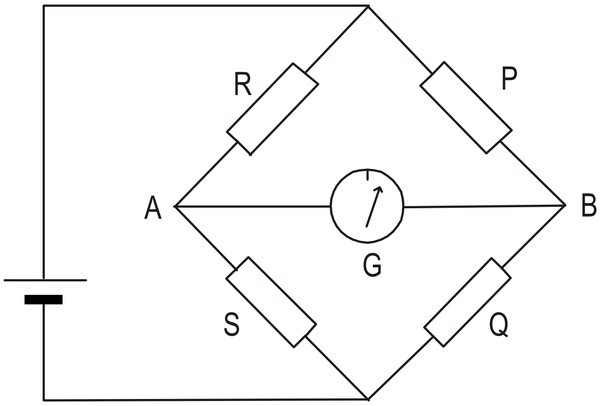
19. The basic configuration consists of a four limb resistance bridge, as shown in Fig 6, where R is the resistance to be measured, and resistors P, Q and S are variable standard resistors.

20. The variable resistors (P, Q and S) are adjusted until no current is registered in the galvanometer connected between points A and B (a galvanometer is a very sensitive DC current measuring instrument using the same principle as the moving coil instrument). Under these conditions the bridge is said to be in balance and the ratio of P to Q will then equal that of R to S.

Therefore, $\frac{P}{Q} = \frac{R}{S}$

$$R = \frac{P}{Q} \times S$$

14-5 Fig 6 A Wheatstone Bridge



POWER

Volt-amp Method

21. Referring back to the circuit shown in Fig 1, power in the resistor may be found by measuring the voltage across the resistor and the current through it. The power (P) is then given by the equation:

$$P = VI$$

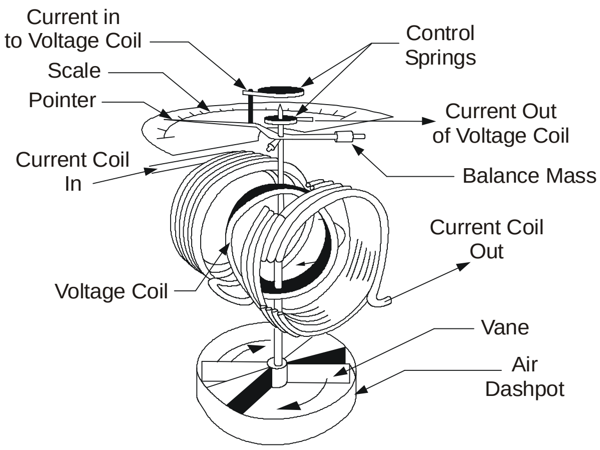
22. This method may be used to give accurate measurement of power in a resistive load in AC and DC circuits. Where loads are not purely resistive, ie they contain inductance and capacitance, the foregoing method of determining power gives incorrect results; this is due to the phase differences between voltage and current. Under these conditions a wattmeter is used.

Wattmeters

23. The moving coil meters discussed in the earlier paragraphs of this chapter rely on permanent magnets to provide the fixed magnetic field. However, this magnetic field can be created by passing current through fixed current coils. These additional coils are the basis of wattmeter design.

24. A wattmeter consists of two fixed low resistance current coils and a high resistance voltage coil which is free to move as shown in Fig 7. The amount of pointer deflection will be proportional to the current flowing through the load and the voltage across it; ie power.

14-5 Fig 7 A Wattmeter

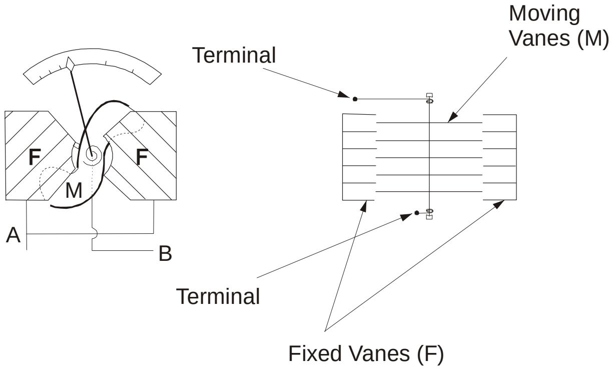


OTHER COMMON INSTRUMENTS

The Electrostatic Voltmeter

25. The electrostatic voltmeter is essentially a variable capacitor whose moving plates are suspended by a spring or wire and are free to move, as shown in Fig 8. When a voltage is applied, the mutual attraction between the fixed and moving plates causes the latter to turn until restrained by the restoring force of the spring. Since no current flows after the initial charging of the plates, the electrostatic voltmeter consumes no power and has a very high internal resistance-ideal for voltage measurement. The chief disadvantage of this type of meter is its insensitivity over the low voltage ranges.

14-5 Fig 8 The Electrostatic Voltmeter

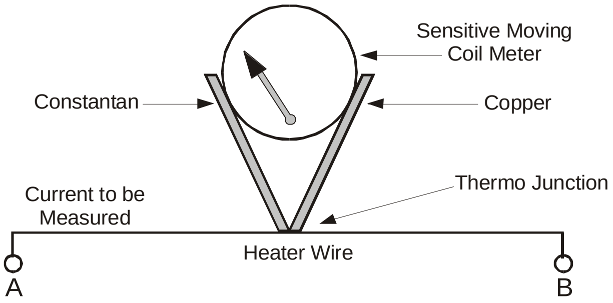


The Thermo-ammeter

26. When two dissimilar metals are in contact with each other an electromotive force (emf) is produced at the point of contact. The magnitude of this emf depends on the metals used and the temperature of the junction. This principle is used in the thermo-ammeter, an instrument which can be used for AC measurement at both low and high frequencies.

27. In the simplified diagram, shown in Fig 9, current passing through the wire heats the junction of the thermo-couple. The resulting emf can then be measured using a standard moving coil instrument. In addition to being a most useful instrument for high frequency current measurements, the thermo-ammeter principle is used extensively in engine temperature monitoring.

14-5 Fig 9 The Thermo-ammeter (Principle of Operation)



Digital Measuring Instruments

28. The instruments discussed so far indicate their readings by the movement of a pointer over a scale and are known as 'analogue' instruments. An alternative form of display presents the reading as a sequence of digits, and instruments with this form of display are referred to as 'digital' instruments. Digital measuring instruments have the advantage that the measured quantity is displayed in numerical form and, in comparison to the moving pointer type of instrument, they are quicker and easier to read. When used to measure voltages, digital instruments are usually referred to as DVMs (digital voltmeters).

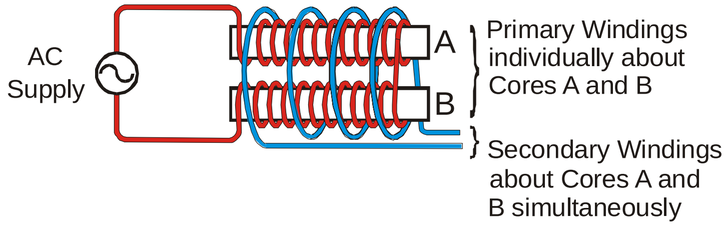
29. The reading is produced by a voltage measuring analogue-to-digital converter. The binary output from the circuit is applied to a decoder which drives the meter display. When the input voltage is of the order of millivolts, it is normally amplified before being measured. For varying voltages, a latch circuit is included to hold the display steady at the last value while further samples of the input are taken. With some extra internal components, currents and resistances can be measured, thus allowing the various multimeter tasks to be undertaken.

Flux Meters

30. The need to measure the strength of magnetic fields in science and industry has resulted in the production of magnetometers of many different types. Early magnetometers consisted of spring-loaded magnets in which field strength was measured by the extension of the spring. These were soon superseded by electronic systems, one of the more common being the fluxgate magnetometer.

31. The fluxgate magnetometer depends for its operation on the rapid AC magnetization of a pair of saturable cores. Exciter coils are wound round each of the cores and are in phase opposition. The output sum is taken via a secondary winding which is wound round the outside of both cores. This is depicted in Fig 10.

14-5 Fig 10 Fluxgate Core Windings


AC
Supply
A
B
Primary Windings individually about Cores A and B
Secondary Windings about Cores A and B simultaneously

32. Under external flux free conditions, the magnetic flux induced by the AC excitation in one core cancels out the flux induced in the other; the resultant being a zero output. The introduction of an external magnetic field has the effect of introducing magnetic bias into the system and causes one of the cores, depending on sense, to saturate. Core saturation results in the flux in one core having a greater influence than that in the other. The summing effect of the output coil produces AC current, of twice the input frequency, which is proportional in magnitude to the external magnetic field. This may be measured using a moving coil meter and rectification.

33. Fluxgate magnetometers are capable of detecting rapidly varying magnetic fields, and are ideally suited to the detection of submarines, and sunken objects such as ships or mines.

CHAPTER 6 - DC MACHINES

Introduction

1. While a conductor is moving in a magnetic field in such a way that the magnetic flux linkage is changing, there is in that conductor an induced emf which at any instant is proportional to the rate of change of flux linkage (Volume 14, Chapter 3). If this emf is applied to a closed circuit an electric current will be established. Any machine which produces such an emf is known as a generator. The DC generator (or dynamo) is a mechanically driven machine in which conductors rotating in a magnetic field generate a DC output voltage.

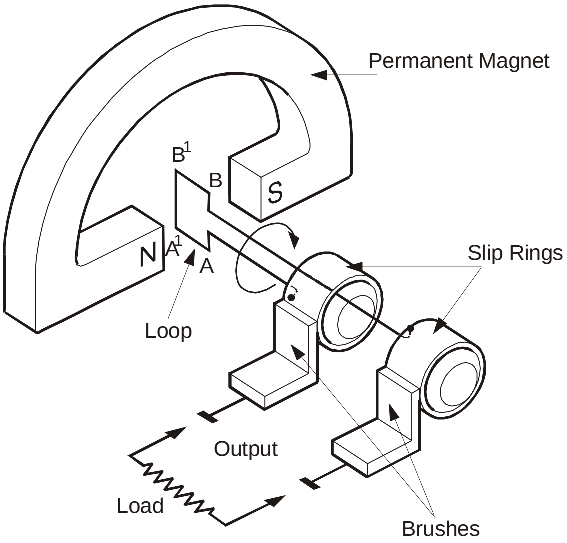
2. An electric motor converts electrical energy into mechanical energy, its action being the reverse of that of the generator. There is little difference in the design of generators and motors; providing the connections are suitable a DC machine can be used as either a generator or a motor.

The Simple Generator

3. The simplest form of generator consists of a single loop of wire which can be rotated freely in the space between the poles of a permanent magnet. Connection is made to the external circuit (or 'load') by brushes pressing on two slip rings connected to the ends of the loop (Fig 1).

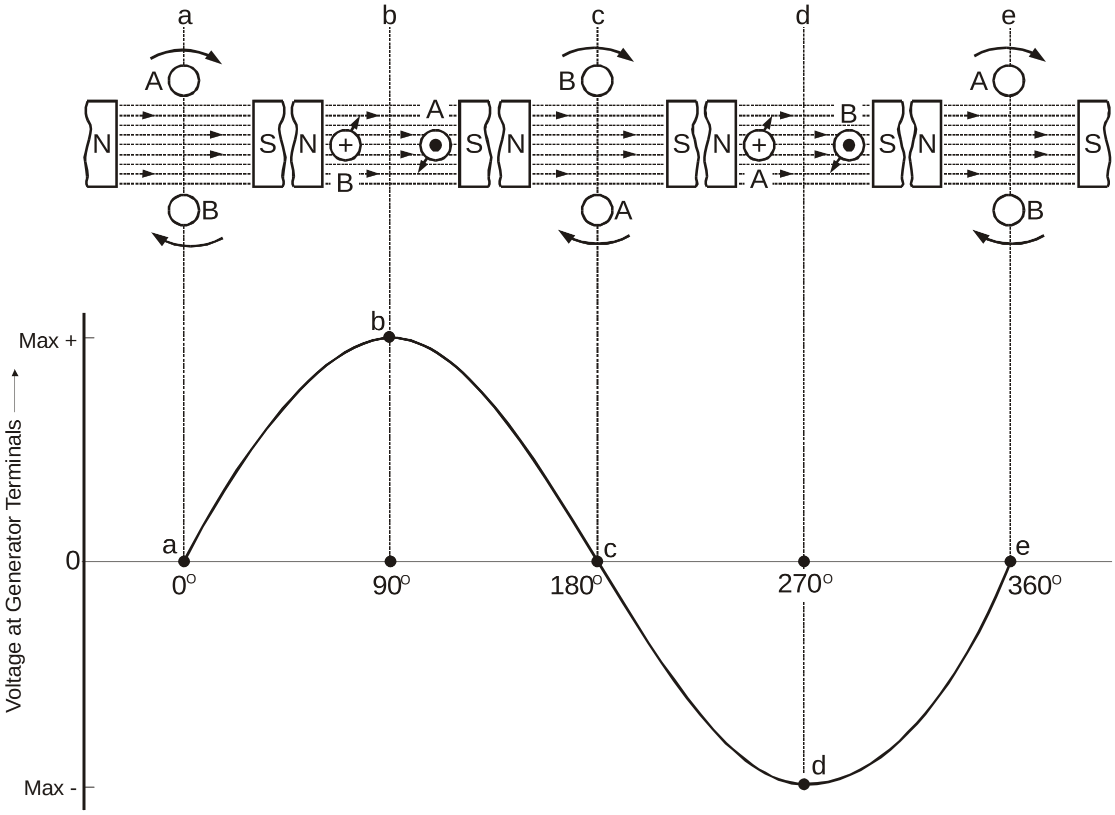
4. While the loop is rotating the magnetic flux linkage is changing, and an emf will be induced in each of the straight sides A – A^1 and B – B^1. The direction of the induced emfs given by Fleming's Right-hand Rule (see para 8) is such that the two emfs in series are additive and combine to establish a current when the load is connected.

14-6 Fig 1 The Simple Generator



5. The magnitude of the induced emf is proportional to the rate of change of flux linkage; thus, assuming the speed of rotation to be constant, the induced emf will at any instant depend on the position of the loop in the magnetic field. This is illustrated in Fig 2, which represents the view from the slip rings end of the loop.

14-6 Fig 2 Values of Induced Voltage for One Revolution of the Coil



a. In position 'a' Fig 2, the conductors A and B are moving parallel to the lines of magnetic flux. In a very small period of time dt about the instant shown in 'a', the change of flux linkage (dΦ) is zero. Thus the rate of change of flux linkage $\frac{d\Phi}{dt}$ is zero, and since E (the back emf) = $-\frac{d\Phi}{dt}$ volts, the emf induced in the loop at this instant is zero.

b. Position 'b' shows the conductors cutting the flux at right angles. The rate at which the flux linkage is changing is now a maximum and the emf induced in the loop is a maximum, the direction being given by Fleming's Right-hand Rule. At this position the emf is arbitrarily assumed to be maximum in a positive direction.

c. Position 'c' represents a position where the rate of change of flux linkage is again zero and no emf is induced in the loop.

d. Position 'd' is similar to position 'b' except that the side of the loop which was previously moving downwards (side A) is now moving upwards and vice versa. The rate of change of flux linkage is again a maximum and Fleming's Right-hand Rule will confirm that in relation to 'b' the emf is now a maximum in a negative direction.

e. Position 'e' is identical with position 'a'.

6. At other positions of the loop, the rate of change of flux linkage is intermediate between zero and maximum and so, therefore, is the emf. Thus during one complete revolution of the loop the voltage at the terminals of the generator will vary in the manner shown in the graph of Fig 2. This shows one cycle of alternating voltage.

7. The frequency of the alternating voltage in cycles per second, or hertz, at the generator terminals depends on the speed of rotation and on the number of pairs of poles in the field magnet system. Thus:

$$\text{frequency } f = \frac{pN}{60}\,(\text{Hz})$$

where p = number of pairs of poles and N = revolutions per minute

Hence the frequency of an alternating voltage produced by a 4-pole AC generator running at 1,500 rpm is:

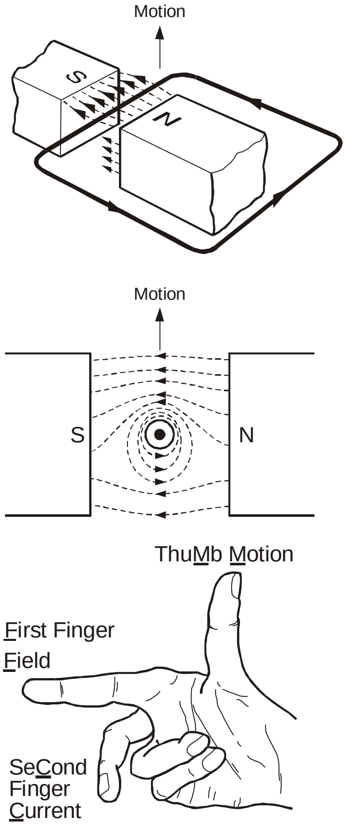
$$f = \frac{pN}{60} = \frac{2 \times 1{,}500}{60} = 50\,\text{Hz}$$

DC GENERATORS

Fleming's Right-hand Rule

8. The right-hand rule for generators is a method of remembering the direction of the voltage induced in a conductor moving in a magnetic field. The 'direction of voltage' indicates the direction in which conventional current would flow in a closed circuit as a result of this voltage. The right-hand rule (Fig 3), sometimes called 'the geneRIGHTer rule', requires the thumb, the first finger and the second finger of the RIGHT hand to be held at right angles to each other. With the thu**M**b pointing in the direction in which the conductor has been **M**oved, and the **F**irst finger in the direction of the magnetic **F**ield (N to S), the se**C**ond finger indicates the direction in which conventional **C**urrent would flow in the conductor; this in turn gives the direction of the induced voltage.

14-6 Fig 3 Fleming's Right-hand Rule

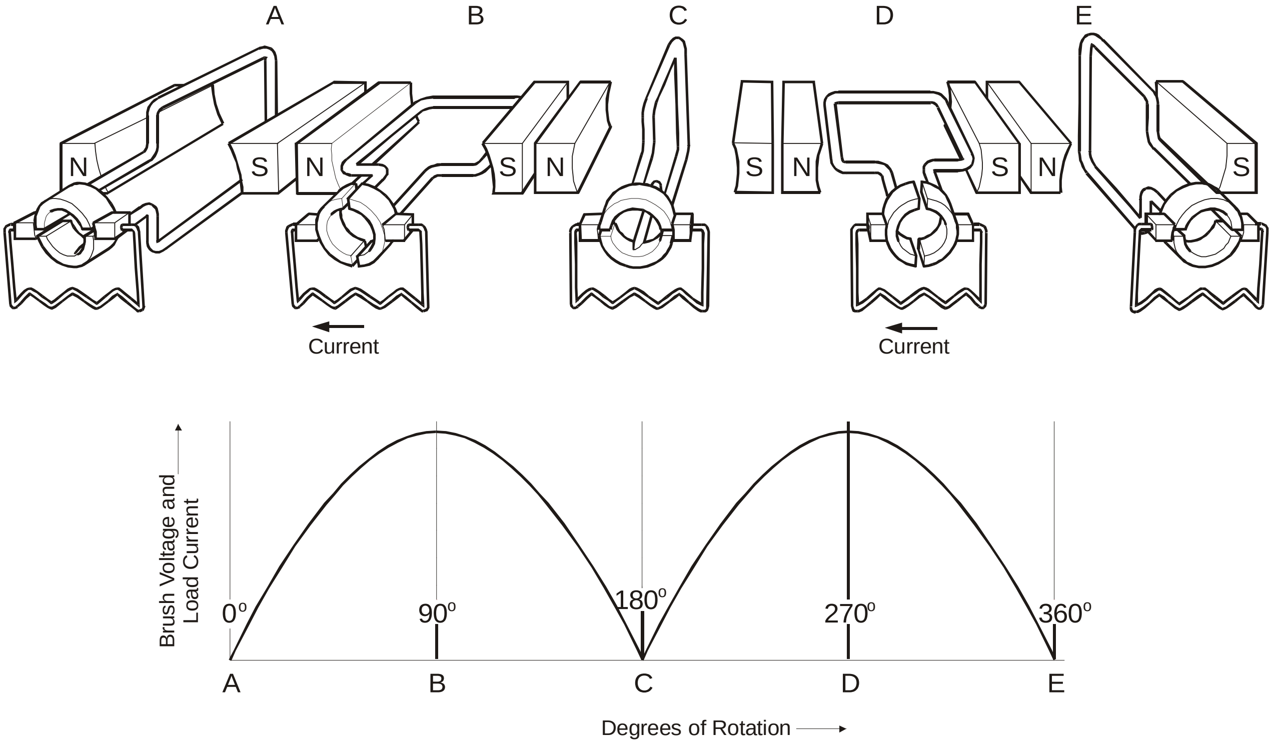


Production of DC

9. The alternating voltage produced in the loop reverses its polarity every time the loop goes through the 0º and 180º positions, because at these points the conductors forming the loop reverse their direction of travel across the magnetic field. To obtain DC from the generator, the AC produced by the rotating loop must be converted to DC. A possible method is to have a switch connected across the output in such a way that the connections to the load are reversed every time the polarity of the voltage in the loop changes. The commutator is a device which does this, switching automatically as the loop rotates, thus maintaining the same direction of current in the load.

10. **The Commutator**. A simple commutator for a single-loop generator consists of the two halves of a slip ring insulated from each other and from the shaft which carries them and the loop. Each end of the loop is connected to a segment of the commutator, and the load is connected to the loop by brushes bearing on opposite sides of the commutator as shown in Fig 4. As the loop rotates, an alternating voltage is induced in it; a fluctuating DC output is derived from this alternating voltage by using a commutator. Since the commutator rotates with the loop, the brushes bear on opposite segments during each half cycle (compare B and D of Fig 4). The left-hand brush is always in contact with that segment which is negative, and the right-hand brush with that segment which is positive. The changeover from one segment to the other takes place at the instants when the voltage induced in the loop is zero, at A, C and E. The current in the external circuit is therefore always in the same direction, it is a unidirectional current, and this is the first step towards obtaining a true DC output such as we get from a battery. The variations in brush voltage and external circuit current during one complete revolution of the loop are also shown in Fig 4. Note that within the rotating loop itself an alternating voltage is produced; the commutator is merely a switch that reverses the direction of alternate half-cycles in the output leads.

14-6 Fig 4 Production of DC by Commutator Action



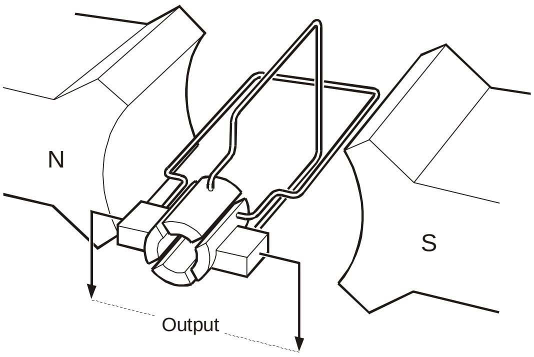
Improving the DC Output

11. The voltage at the brushes and the current in the external circuit of a single-loop DC generator fall to zero twice during each revolution. The voltage wave-form is improved, so that it becomes smoother and of greater amplitude, by:

a. Increasing the numbers of loops, which necessitates additional segments on the commutator.

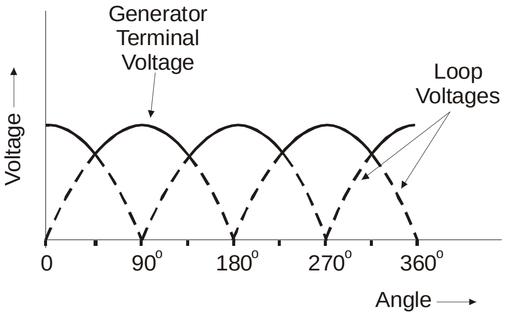
b. Creating stronger magnetic fields.

14-6 Fig 5 A Two-loop Generator



12. **Use of More Than One Loop**. When just one loop is added, thus giving two loops at right angles to each other (as shown in Fig 5), the output voltage using a commutator with four segments becomes that shown in Fig 6. The ripple that remains may be made smoother by the use of additional symmetrically placed loops, and the use of a commutator with a correspondingly large number of segments.

14-6 Fig 6 Two-loop Generator Output



13. **The Creation of Stronger Magnetic Fields**. Greater effective strength of the magnetic fields is obtained by:

a. Using electromagnets instead of permanent magnets. The pole pieces are made of soft iron and are magnetized by a current flowing through the coils wound round them. These are known as the field windings.

b. Winding the rotating loops, which should be made of many turns of wire, on a soft iron core to concentrate the magnetic field through the loops. This assembly is known as the armature. The coils, are, of course, insulated from the iron core.

c. Decreasing the air gap between the poles and armature by shaping them to fit each other.

d. Increasing the number of poles. Four pole generators are quite usual; heavy duty and larger machines use more.

Variation of the Magnitude of the Induced EMF

14. From Faraday's and Lenz's laws the emf induced in a conductor moving through a magnetic field is given by:

$$E = -n\frac{d\Phi}{dt} \text{ volts}$$

where $\frac{d\Phi}{dt}$ represents the rate of change of flux, and n the number of turns of the coil. The emf of a generator therefore depends on:

a. The number of conductors connected in series between the brushes.

b. The speed of rotation.

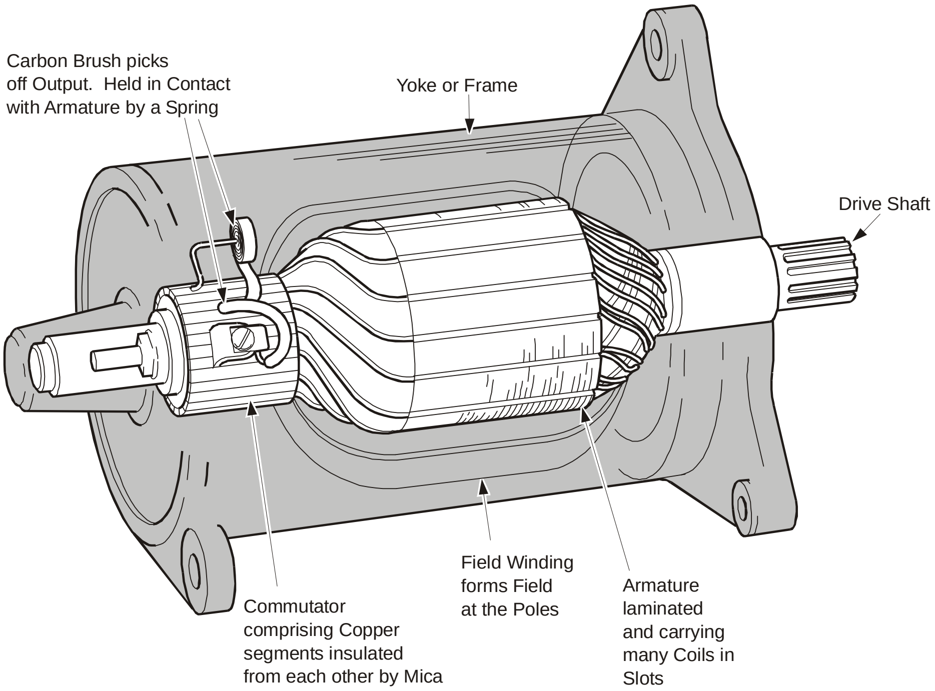
c. The magnetic flux density.

15. In practice the number of conductors is fixed and the speed is nominally constant, so control of the emf must be obtained by variation of the magnetic flux density. Using permanent magnets the flux density is constant, but by substituting electromagnets the emf can be controlled by varying the current in the magnet windings.

Practical DC Generator

16. A small DC generator is illustrated in Fig 7. The mainframe or yoke is the main chassis of the generator and it also serves to complete the magnetic circuit between the pole pieces. The pole pieces are laminated to reduce eddy current losses (see para 25), and the field windings are mounted on the pole pieces as shown. The end housings contain the bearings for the armature which rotates at high speed.

14-6 Fig 7 A Typical DC Generator


Carbon Brush picks off Output. Held in Contact with Armature by a Spring
Yoke or Frame
Drive Shaft
Field Winding forms Field at the Poles
Armature laminated and carrying many Coils in Slots
Commutator comprising Copper segments insulated from each other by Mica

17. The armature is made up of a shaft, armature core, armature windings and the commutator. The armature core is laminated to reduce eddy current losses, the armature winding resting in slots cut in the core but insulated from it. The commutator is made up of copper segments insulated from each other, and from the shaft, by mica. The ends of the armature windings are soldered to their appropriate commutator segments.

18. **The Brush Assembly**. The output emf is picked off from the commutator by the brushes. They are made of some form of carbon which is self-lubricating and which causes very little wear of the commutator. In addition their relatively high resistance minimizes a tendency for sparking to occur as the brushes pass from one commutator segment to the next. The carbon pieces are mounted in brush holders and are held against the commutator by tensioned springs. The holders are bolted to brush-rockers which can be adjusted to move a few degrees in either direction so that the tendency for sparking to occur may be further reduced. The brushes are connected to the external circuit by copper wires.

Classification of DC Generators

19. DC generators are usually classified according to the method by which the magnetic circuit of the machine is energized. The recognized classes are:

a. Permanent magnet generators.

b. Separately-excited generators.

c. Self-excited generators.

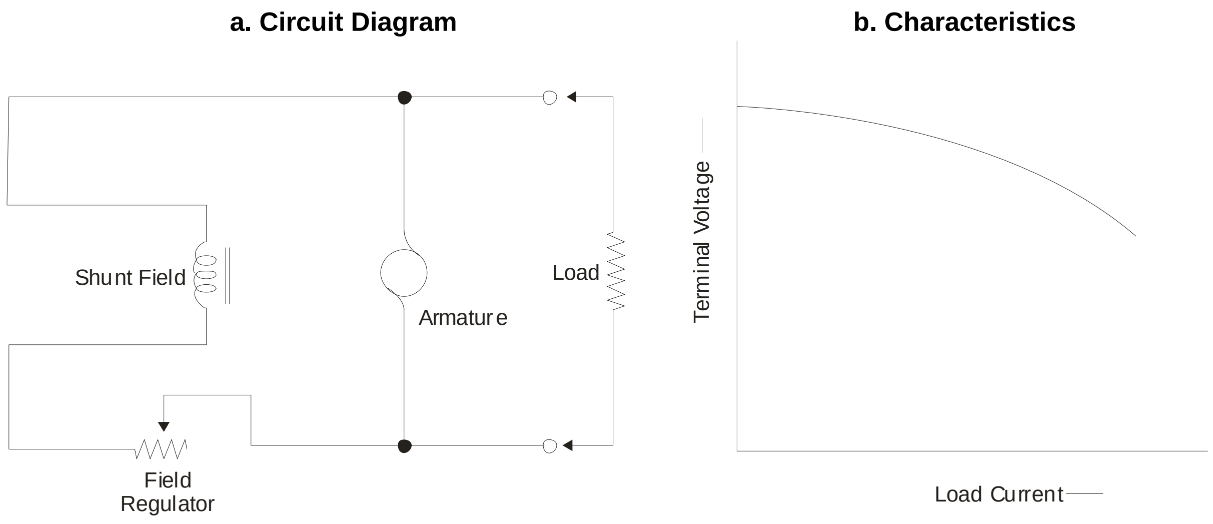
20. **Permanent Magnet Generators**. The permanent magnet type of generator is usually a simple, small machine because of the low strength of the magnetic field. The PD across the terminals decreases slightly as the loading increases, the generator having what is termed a falling external voltage characteristic.

21. **Separately-excited Generators**. The magnetic field for the separately-excited type of generator is obtained from electromagnets which are excited by an external DC source. The exciter supply can be regulated, varying the magnetic field strength, and thus the terminal PD. The terminal PD falls with an increase in loading.

22. **Self-excited Generators**. The power required to excite the electromagnets in the self-excited type of generator is generated by the machine itself. This type of generator may be further classified as follows:

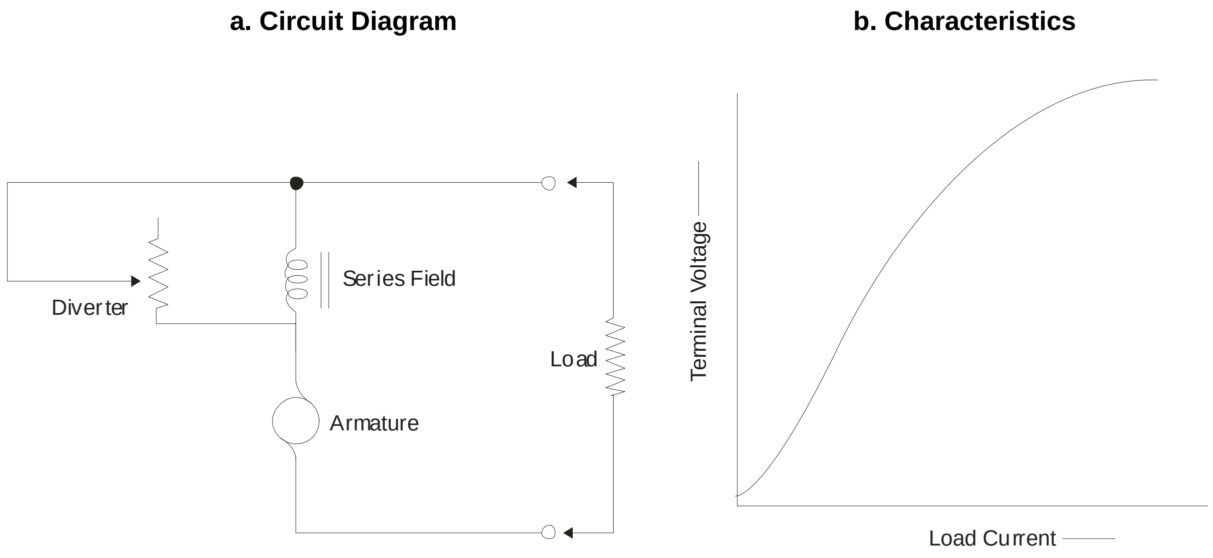
a. The shunt-wound type, where the field winding is connected directly across the armature. Provision is made for regulating the exciting current by a variable resistor connected in series with the field winding as shown in Fig 8a. Thus control over the terminal PD is obtained. In a shunt generator, as the load draws more current from the armature, the terminal voltage decreases because there is a greater potential drop in the armature and hence a lower terminal PD, and thus field current decreases. The characteristic curve for a shunt generator thus starts at a maximum and then falls as shown in Fig 8b. Note that the generator is excited even with the load disconnected.

14-6 Fig 8 Shunt Generator



b. The series-wound type where the field winding is connected in series with the armature and the load. The terminal PD may be controlled by a diverter (a variable resistor connected in parallel with the field winding so as to adjust the current in the field coils). The circuit is shown at Fig 9a. If the load is disconnected no current flows through the field coils; the only voltage induced is that due to residual magnetism. With a load connected, current flows in the armature, the field strength increases and the terminal voltage rises. This continues as the load draws more current until the magnetic field reaches saturation. The characteristic curve for a series generator is shown at Fig 9b.

14-6 Fig 9 Series Generator



c. The compound-wound type, which is a combination of the types in sub-paras a and b, and may be constructed so as to give an almost constant voltage output under all loads.

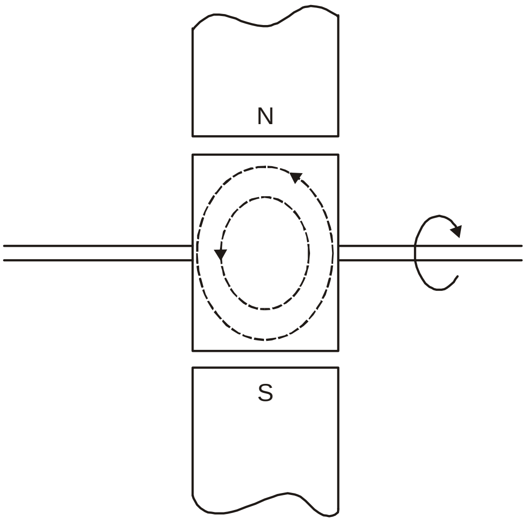
Generator Losses and Efficiencies

23. There are four types of losses associated with DC generators: copper losses (para 24), hysteresis loss (see Volume 14, Chapter 2, para 23), eddy current loss (para 25), and friction losses (which are self-explanatory).

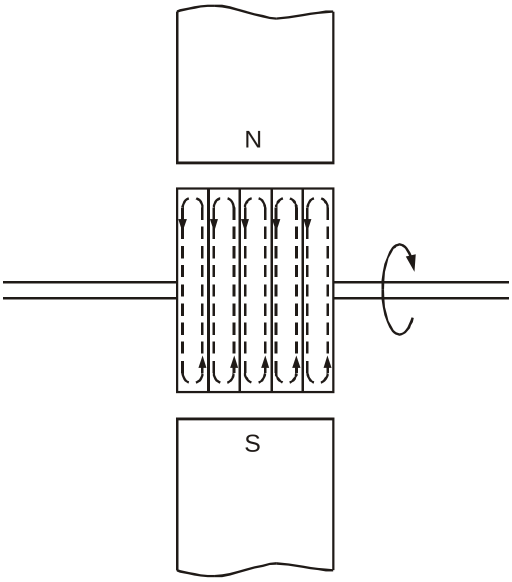
24. **Copper Losses**. Another name for copper losses is electrical losses. They are a power loss (I^2R watts), caused by the resistance of the conductors. Usually they increase with loading.

25. **Eddy Current Losses**. If the armature core were made of solid iron, as shown in Fig 10 for a two-pole machine, and rotated, eddy currents would be set up in the core. The eddy currents are caused by emfs which are generated in the iron in exactly the same way as emfs generated in conductors wound on the iron core (Fleming's Right-hand Rule will give the current direction). These currents cause two losses: one through the heat generated, and the other by creating magnetic fields which oppose the pole magnetic fields. The losses are minimised by splitting up the solid iron core, ie by laminating it, so that the eddy currents are split also (see Fig 11). A five-lamination core suffers about 1/25th the loss of a solid core (loss $\propto I^2$).

14-6 Fig 10 Production of Eddy Currents



14-6 Fig 11 Splitting Eddy Currents by Laminations



26. **Efficiencies**. Due to the losses mentioned in the previous paragraph the efficiency of a DC generator in converting mechanical energy into electrical energy varies with the type of machine. Three definitions of efficiency follow:

a. **Mechanical Efficiency**. The electrical power in the armature is equal to the applied mechanical power less the hysteresis, eddy current and friction losses. Thus:

$$\text{mechanical efficiency (\%)} = \frac{\text{power in armature}}{\text{power supplied}} \times 100$$

b. **Electrical Efficiency**. The electrical power output is equal to the power in the armature less the copper losses. Thus:

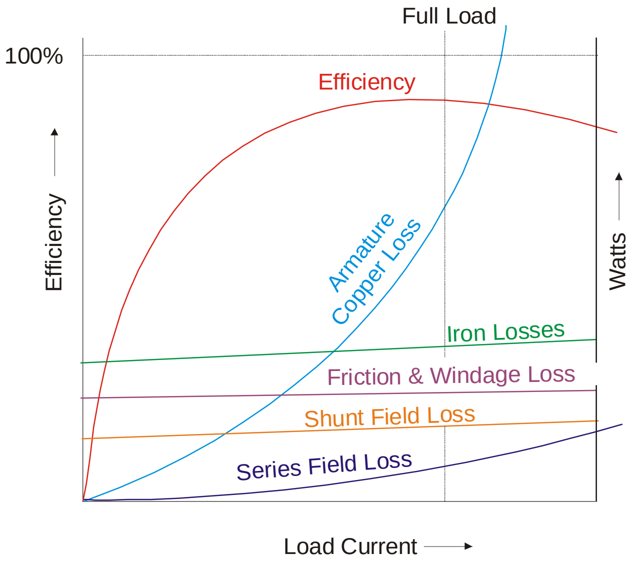
$$\text{electrical power output efficiency (\%)} = \frac{\text{power output}}{\text{power in armature}} \times 100$$

c. **Commercial Efficiency**. The commercial efficiency is the one most commonly used to describe the overall efficiency of a machine and:

$$\text{commercial efficiency (\%)} = \frac{\text{electrical power output}}{\text{mechanical power input}} \times 100$$

A graph of losses and efficiency for a compound generator is shown at Fig 12.

14-6 Fig 12 Losses and Efficiency of a Compound Generato

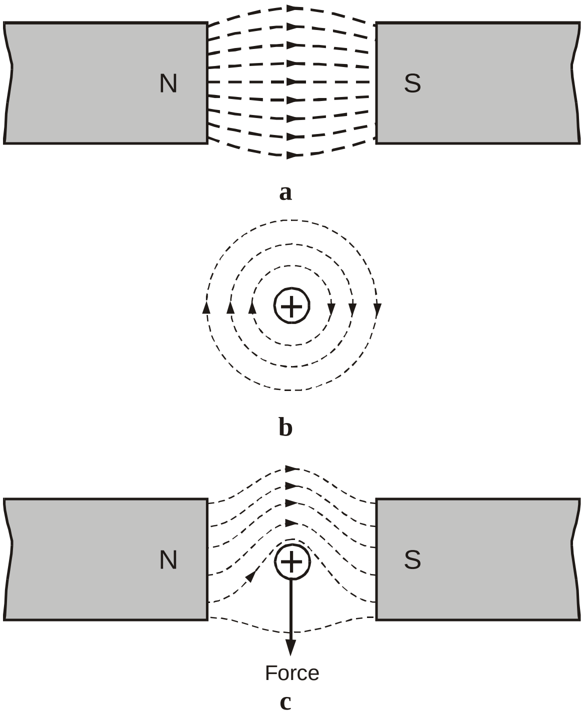


THE DC MOTOR

General Principles

27. An electric motor is a machine for converting electrical energy into mechanical energy, its function being the reverse of that of the generator. The conversion of the electrical energy relies on the fact that an electric current flowing through a conductor placed in a magnetic field causes the wire to move (if free to do so). The lines of flux of the two fields interact as shown in Fig 13 and, in this case, a force is created downward on the conductor.

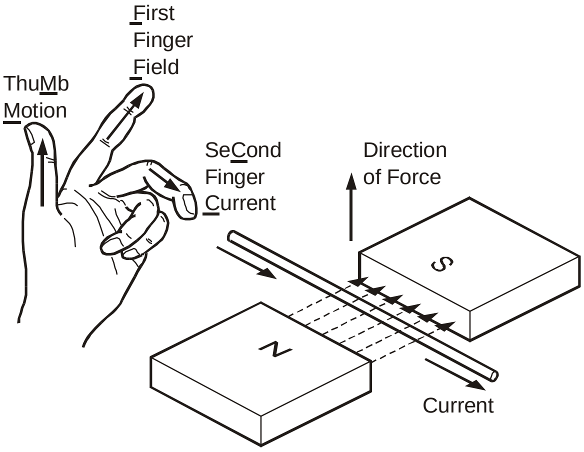
14-6 Fig 13 Force on a Current Carrying Conductor



Fleming's Left-hand Rule

28. The direction in which a current carrying conductor will move when placed in a magnetic field can be determined from Fleming's Left-hand Rule (see Fig 14). If the first finger, the second finger, and the thumb of the LEFT hand are held at right angles to each other, then with the **F**irst finger pointing in the direction of the **F**ield (N to S), and the se**C**ond finger in the direction of the **C**urrent in the conductor, the thu**M**b will indicate the direction in which the conductor tends to **M**ove.

14-6 Fig 14 Fleming's Left-hand Rule

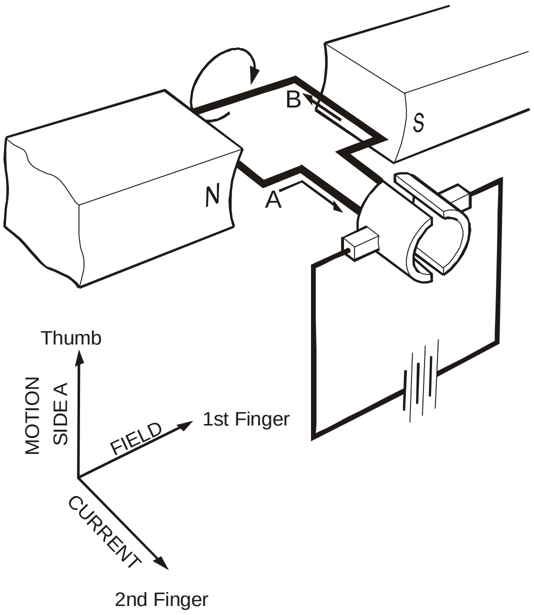


The Simple Motor

29. A motor depends for its operation on the force exerted upon current-bearing conductors situated in a magnetic field. Consider a simple permanent magnet motor connected to a battery as shown in Fig 15. By applying Fleming's Left-hand Rule it will be seen that side A of the loop (under the N pole) tends to move upwards, while side B of the loop tends to move downwards. The forces acting on the two sides of the loop are thus cumulative in their effect, and tend to turn the loop in a clockwise direction.

30. As the sides of the loop pass through the neutral position, the commutator reverses the connections of the supply to the loop, and the current in the loop is consequently reversed. It follows that the force acting on side A will now be downwards, and on side B upwards. The mechanical force on the loop is thus continued in the original direction, and rotation continues so long as the supply is connected.

14-6 Fig 15 DC Motor Principle



The Practical Motor

31. A single-loop motor is of little use and improvements like those made to a simple generator described in paras 11, 12 and 13 are incorporated. The direction of rotation of the motor depends upon the direction of the current in the armature coils and also upon the direction of the magnetic field. If one of these is reversed the direction of rotation is reversed; reversing both together has no effect.

The Speed of a Motor

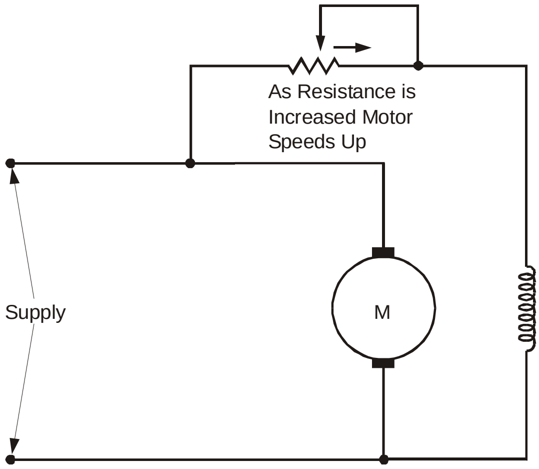
32. **The Back emf**. When the armature in the motor rotates an emf is induced in the conductors by generator action. The direction of the emf is such as to oppose the motion or other cause producing it (Lenz’s law), that is it opposes the supply voltage. This emf (called the back emf) is proportional to the field strength and the speed of rotation of the armature, but can never be as great as the supply input voltage.

33. **Speed with Variable Load**. The back emf developed in the armature of a DC motor when it is running determines the current in the armature, and makes the motor a self regulating machine in which speed and armature current are automatically adjusted to the mechanical load. At small values of load the shaft torque exceeds the load torque; the armature therefore accelerates and gives rise to a large back emf. The back emf cuts down the armature current, thus reducing the shaft torque, until eventually a state of balance between the two torques is obtained and the speed is stabilized. With increasing load the load torque is increased, exceeding the shaft torque and causing a fall in armature speed. Reduced armature speed results in reduced back emf and increased armature current; the increase in armature current produces an increase in shaft torque restoring torque balance and stabilizing the speed again. The variation of speed with armature current (ie with mechanical load) is known as the speed characteristic of the motor.

34. **Control of Motor Speed**. Assuming constant load, there are two methods commonly used to vary the speed of a DC motor:

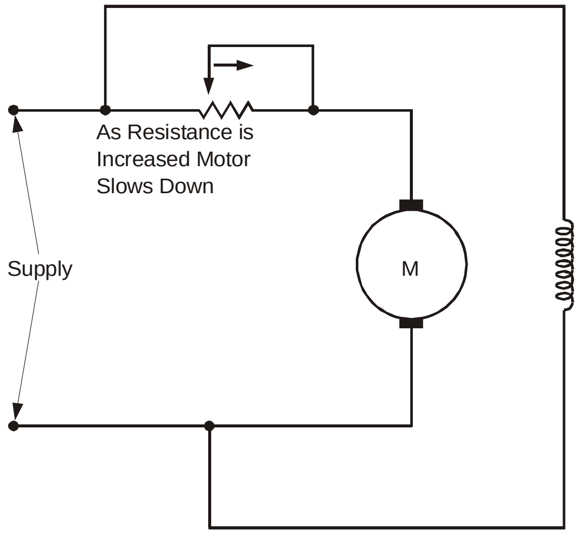
a. **Field Control** (Fig 16). By weakening the main flux of a motor the back emf is reduced, increasing the effective voltage and the armature current. The increased armature current gives rise to an increased shaft torque, causing the motor to accelerate until the back emf, rising with increased speed, restricts the armature current and shaft torque to restore the balance of shaft and load torques. At this point the speed of the motor will stabilize. Conversely, an increase in field strength will cause a reduction in speed. The minimum speed will be obtained with full field excitation, this being known as the normal speed of the machine.

14-6 Fig 16 Speed Varied by Controlling Field Strength



b. **Armature Control** (Fig 17). By reducing the voltage across the armature of a motor the effective voltage is reduced, with a corresponding reduction in armature current and shaft torque. The excess of load torque over shaft torque causes the motor to slow down to a point where the reduced back emf permits sufficient armature current to produce a state of balance between the two torques. At this point the speed of the motor will stabilize.

14-6 Fig 17 Speed Varied by Controlling Armature Current



Types of Motors

35. DC motors are classified according to the method by which the field is excited:

a. The majority of motors are comparable to self-excited generators, ie the armature winding and field winding are supplied from a common source. Their speed and load characteristics vary according to the method of connecting the field winding to the armature, and as a class, they are capable of fulfilling most requirements.

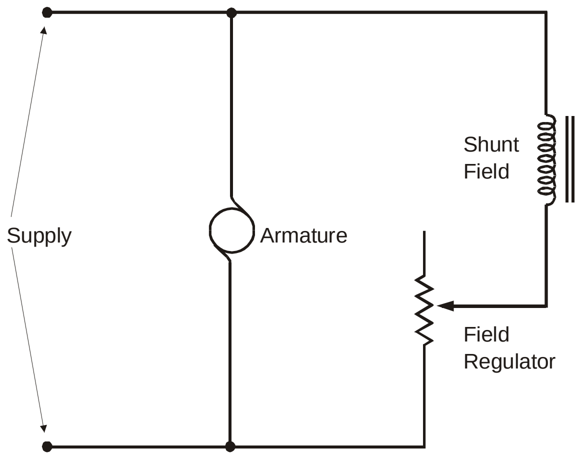
b. Separately-excited motors are used only for special purposes where the more normal types are unsuitable.

c. Permanent magnet motors are employed for certain special purposes, eg in small control systems.

Only the self-excited types will be described in detail.

36. **Shunt-wound Motors**. The field winding of the shunt-wound motor is connected in parallel with the armature (Fig 18); it is thus directly across the supply and must be of relatively high resistance to restrict the current through it. The speed of the motor drops only slightly as the load increases and can be considered constant for many applications. The torque develops proportionally with armature current. The initial torque is small because of the restricted armature current, and shunt motors should be started on light load or no load. Its constant speed makes this type of motor suitable for lathes, drills, and light machine tools.

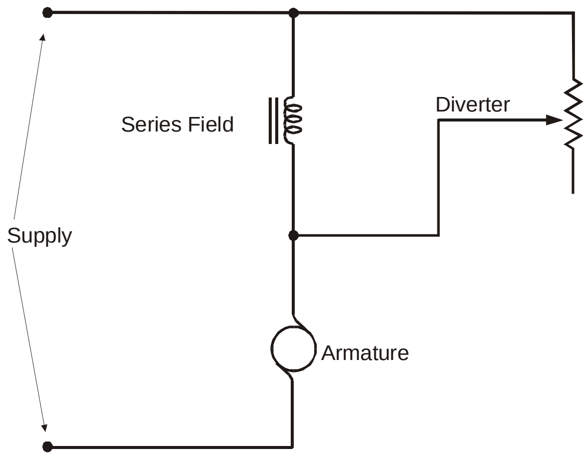
14-6 Fig 18 A Shunt-wound Motor



37. **Series-wound Motors**. The armature and the field winding of the series-wound motor are in series with each other across the supply (Fig 19). Thus the currents in each of them are the same or (as in Fig 19) proportional to each other. The speed decreases with increase in load, and these motors should always be started under load to prevent over-speeding at start-up. The torque increases rapidly as the load is increased, the starting torque being high. Due to its high starting torque a series-wound motor is used for engine starting and traction work.

38. **Compound-wound Motors**. By arranging for part of the field winding to be in series with the armature and part in parallel with it, the large starting torque of the series motor and the almost constant speed of the shunt can be reproduced in the compound-wound motor.

14-6 Fig 19 A Series-wound Motor



Motor Starters

39. Due to the low resistance of the armature, extremely high currents would be experienced in the armature until its movement caused sufficient back emf to limit the current to an acceptable level. To prevent the high current burning out the armature windings, a starter resistance is inserted in series with them. The resistance is then progressively reduced as the back emf increases. Starter resistors are usually used with heavy motors; the armatures of small motors have sufficiently high resistance to limit the current.

Rotary Transformers

40. The rotary transformer is a combination of a DC motor and generator. It consists of a single field system and a single armature, on which are wound the motor and generator windings. A DC input is transformed into a different DC output, a typical example being an input of 24V 7A and an output of 1,200V l00 mA. The power output is always less than the power input.

CHAPTER 7 - ALTERNATING CURRENT THEORY

Introduction

1. It was stated earlier that a direct current is one in which the electron flow is in one direction only through the circuit, whereas an alternating current is one in which the electrons flow first in one direction then the other, changing direction at regular, very short time intervals.

2. The national electrical power supply network delivers power, the mains, to consumers as an AC. Two advantages of AC over DC in this case are:

a. With AC, voltages can be readily stepped up or stepped down by transformers (which will be discussed later), whereas with DC conversion from one voltage to another is cumbersome and expensive.

b. With AC, the loss of power in transmission is minimized. The power carried by a supply line is EI; the power lost in transmission due to the resistance of the cables is I^2R. By using a high voltage and a small current high power can be carried with little loss.

3. For the mains supply, the AC voltage generated at the power station is stepped up to a high value for transmission over the national grid, and then stepped down at sub-stations to the nominal mains supply voltage (usually 240 V) for domestic lighting, heating, etc.

AC WAVEFORMS

Terms Used

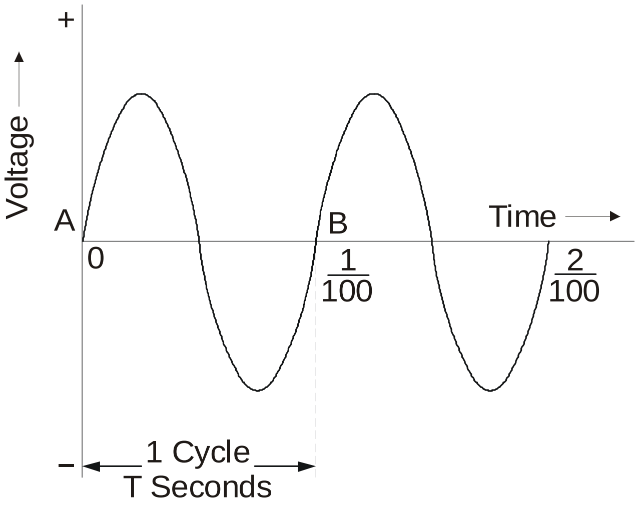
4. The production of an alternating emf by a simple generator was discussed in Volume 14, Chapter 6. Fig 1 is a graph of an alternating voltage, showing voltage against time. Some of the terms associated with such a waveform are:

a. **Cycle**. A complete sequence of positive and negative values, such as that from A to B.

b. **Period (or Periodic Time)**. The duration of one cycle, denoted by the symbol T. In Fig 1, T = 1/100 second.

c. **Frequency**. The number of cycles completed in one second, denoted by the symbol f. It will be apparent that f=1/T cycles per second. In Fig 1, since T=1/100 sec, f = 100 cycles per second, i.e. 100 hertz (Hz). Low frequency AC currents are used in the supply of electrical power. The AC mains supply in Britain has a frequency of 50 hertz; aircraft generally use 400 hertz. In radio, AC currents with frequencies ranging up to hundreds of megahertz are used.

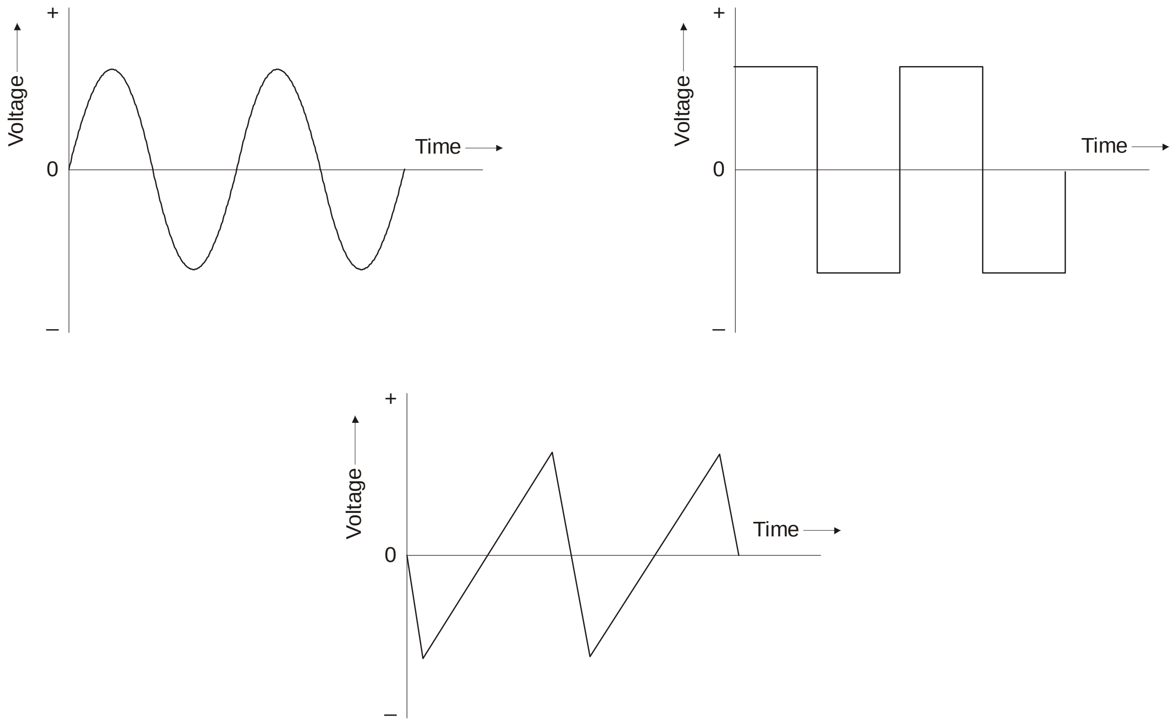
14-7 Fig 1 Alternating Voltage



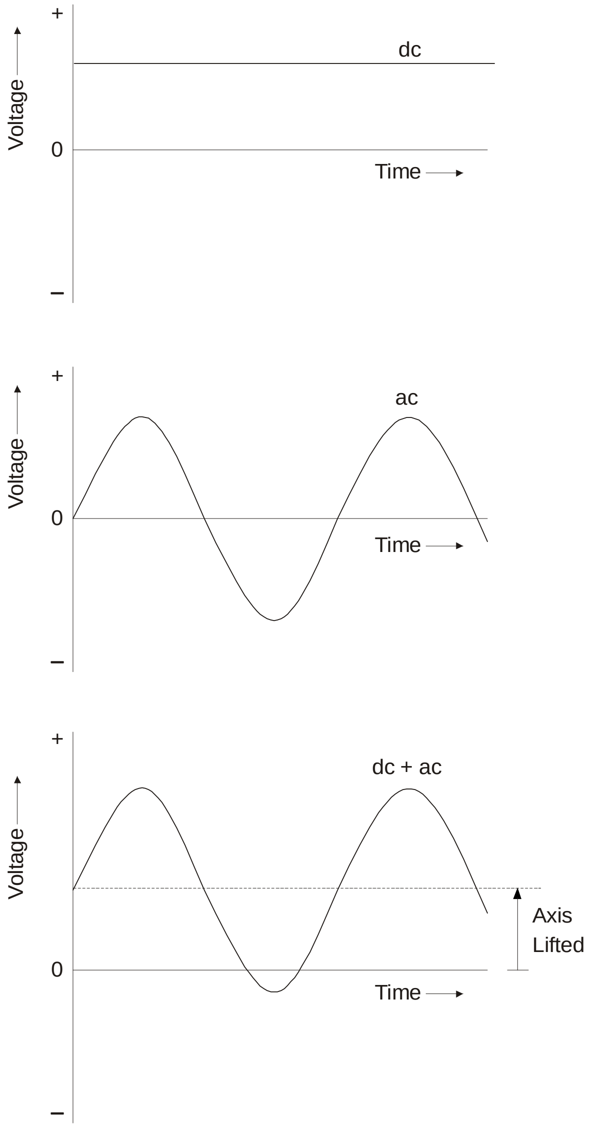
Waveform

5. The shape of the AC voltage or current graph is known as the waveform; in Fig 1, it is a sine curve. Fig 2 shows three voltages with the same frequency but different waveforms. In some circuits, AC and DC are both present, and add together in such a way that the AC is superimposed on the DC. The result is that the AC axis shifts from the zero line (Fig 3). Note that if the resulting voltage never falls below zero, the current is called a pulsating DC current.

14-7 Fig 2 AC Waveforms



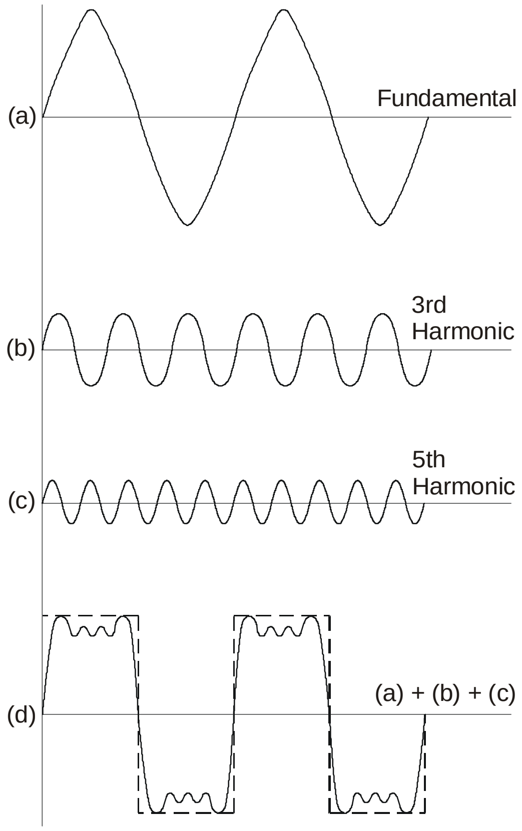
14-7 Fig 3 AC Superimposed on DC



Square Waveform

6. It is of interest here to note how a square wave is produced. Fourier's theorem states that any recurrent wave-form of frequency f can be resolved into the sum of a number of sinusoidal wave-forms having frequencies f, 2f, 3f etc; the number of sine waves may be finite or infinite. The frequency f is known as the fundamental or first harmonic, and frequencies 2f, 3f, 4f, etc as the second, third, fourth etc harmonics. The amplitude of a harmonic is inversely proportional to its frequency. The square wave consists of a fundamental sine wave and all its odd harmonics up to infinity. Fig 4 shows the result of taking harmonics up to the fifth, and it will be seen that this gives a good approximation to a square wave.

14-7 Fig 4 Composition of a Square Wave

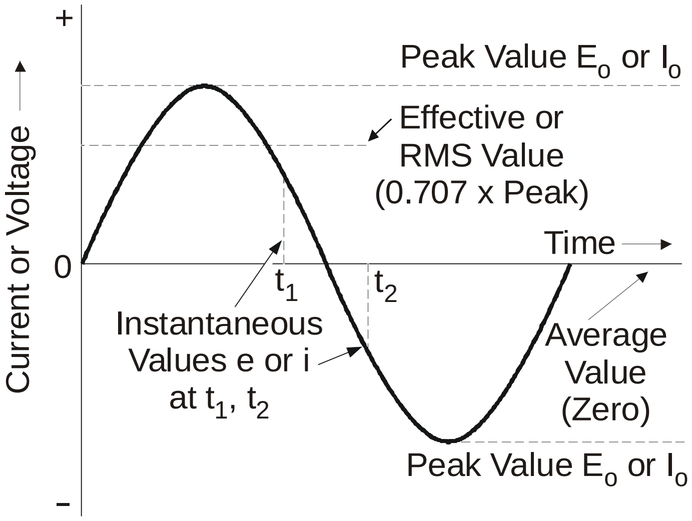


AC VALUES

General

7. Several things can be meant by the value of an alternating quantity such as an AC current; the various terms used to make it clear which meaning is intended are defined in the following paragraphs and illustrated at Fig 5.

14-7 Fig 5 AC Values


+
0
–
Current or Voltage
Peak Value E_o or I_o
Effective or RMS Value (0.707 x Peak)
Time
t_1
t_2
Instantaneous Values e or i at t_1, t_2
Average Value (Zero)
Peak Value E_o or I_o

Instantaneous Value

8. The value of an alternating quantity is continuously changing, its instantaneous value being that at any given instant of time. Instantaneous current is denoted by the symbol 'i', and instantaneous voltage by 'e' (Fig 5).

Peak Value

9. The maximum value of an alternating quantity (either positive or negative) reached during a cycle is called the peak value. Peak current is denoted by I_0 (or I_{max}) and peak voltage by E_0 (or E_{max}).

Average or Mean Value

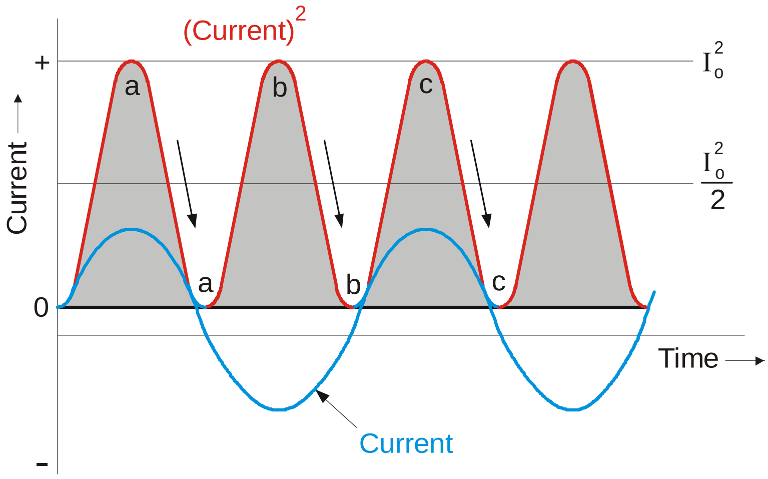
10. The average value of an alternating quantity over a number of complete cycles is zero. In Fig 5, the area between the current curve and the time axis represents the quantity of electricity that has passed during the interval concerned. As the curve is symmetrical about the time axis (provided that as in Fig 5 it represents AC with no DC present), the quantity of electricity that passes one way during the first half exactly equals that passing the other way during the second half of any cycle. The net transfer of electricity, and hence the average current, is zero.

Effective or Root-Mean-Square (rms) Value

11. The heating effect of an electric current is independent of the direction of the current, and the effective value of an alternating current is 1 ampere if it has the same heating effect as an unvarying direct current of 1 ampere.

12. It was shown in Volume 14, Chapter 1, when considering power, that the rate of dissipation of energy as heat in a resistor depended upon the square of the current or voltage, ie $P = I^2R = V^2/R$ watts. This applies also to AC. In Fig 6, the square of the sinusoidal current is plotted against time, giving a curve which is always positive and is symmetrical about the halfway line $\frac{I_0^2}{2}$. The energy dissipated over any time interval is proportional to the shaded area beneath the curve of (current)^2. In this graph, crest a will fit into trough a, crest b into trough b, and so on, so that the area beneath the curve of (current)^2 equal to the area beneath the horizontal line drawn midway between the square of the peak current and the time axis. Thus, the heating effect of the current plotted at Fig 6 is the same as that of a direct current of value $\sqrt{(I_0^2/2)}$, or $I_0/\sqrt{2}$.

14-7 Fig 6 Effective or rms Value



13. This value is known as the effective value (since it is that value which has the same heating effect as an equivalent direct current), or as the root-mean-square (rms) value (since it is obtained by finding the mean value of the square of the current and then taking the square root). If I_{rms} is the rms value of alternating current in a circuit, it is related to its peak value by the expression:

$$I_{rms} = \frac{I_0}{\sqrt{2}} = 0.707\, I_0 \text{ amps}$$

Also, $I_0 = I\sqrt{2}$ = 1.414 I (amps)

14. Alternating currents and voltages are usually measured by their rms values. For instance, the normal AC mains supply has a voltage of 240 V; this is an rms value, the peak voltage being $240\sqrt{2}$ which equals 340 V. Ammeters and voltmeters for AC are normally calibrated in rms values.

REPRESENTATION OF SINUSOIDAL WAVEFORM

General

15. Alternating currents and voltages have so far been represented by graphs. In the following paragraphs this method of representing sinusoidal wave-forms will be examined in more detail, and brief mention will be made of two other methods often encountered: representation by trigonometrical equations and by vectors.

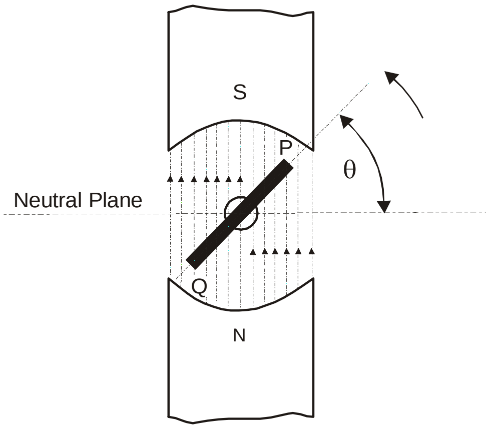
Graphical Representation

16. Consider a conductor loop (PQ) rotating in an anti-clockwise direction at constant speed in a uniform magnetic field, as shown in Fig 7. When the loop is in the neutral plane, at right angles to the magnetic field (θ = 0º), the emf induced in the loop is zero. When the loop is at right angles to the neutral plane (θ = 90º) the emf is a maximum (E_0), its value depending on the speed of rotation and on the flux density. At intermediate points, the induced emf is between zero and its maximum value and depends on the angle which the loop makes to the neutral plane. The emf induced in the loop at any instant is proportional to the sine of the angle (θ) through which the loop has rotated from the neutral plane, i.e.:

$$e = E_0 \sin\theta \text{ (volts).}$$

For example, when the loop is in the neutral plane, θ = 0º, sin θ = 0, and e = E_0sin θ = zero. When the loop is at right angles to the neutral plane, θ = 90º, sin θ = 1, and e = E_0sin θ = E_0 volts.

14-7 Fig 7 Dependence of emf on Sin θ



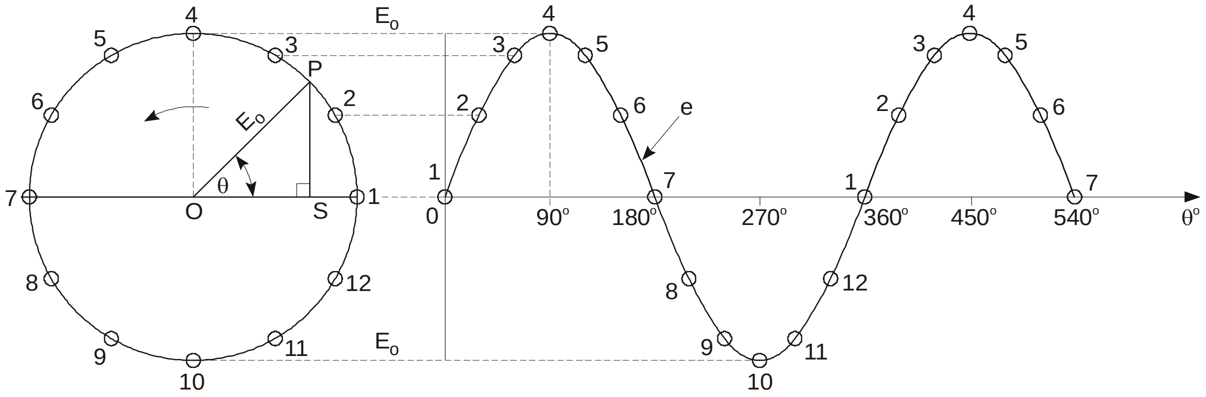
17. To plot a curve showing the instantaneous value of the emf for all values of the angle θ, consider Fig 8. The line OP is assumed to rotate about the point O in an anti-clockwise direction, its length representing the maximum value of the emf (E_0) to any convenient scale. A horizontal line is drawn through the centre of rotation (O) and, along this, a scale of degrees of rotation is marked. Now:

$$\sin\theta = \frac{PS}{OP}$$

$$\therefore\ PS = OP\sin\theta = E_0\sin\theta$$

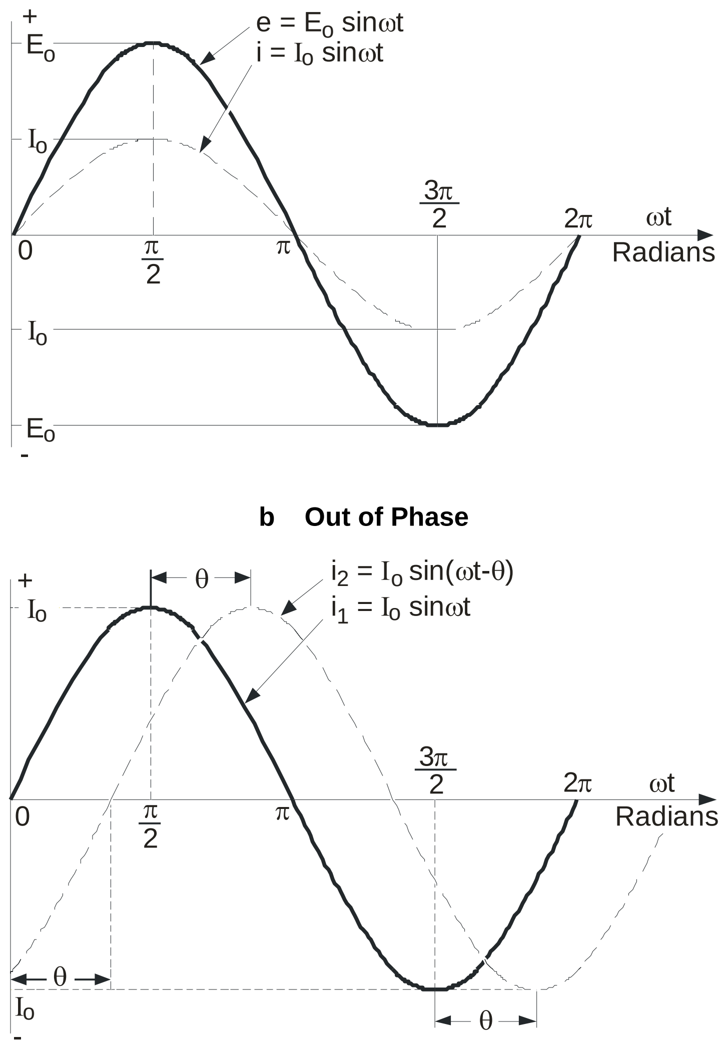
Thus, the vertical line PS represents, to scale, the instantaneous emf 'e' = $E_0\sin\theta$. As OP rotates, the length of the line PS varies, and plotting this against the various values of the angle θ on the horizontal axis, the graph of the instantaneous emf is obtained. This curve is termed a sine curve, and any quantity which varies in this manner is said to have a sinusoidal waveform.

14-7 Fig 8 Plotting a Sine Wave



18. **Phase Difference**. When two alternating quantities of the same frequency pass through corresponding points in a cycle at the same instant of time they are said to be in phase with each other (Fig 9a). If they pass through corresponding points at different instants of time, there is a phase difference between them and one is said to be leading or lagging the other by a certain phase angle. For example, in Fig 9b, i_1 is leading i_2 by θ radians (or i_2 is lagging by θ radians). Thus, i_1 reaches its maximum value θ radians before i_2.

14-7 Fig 9 Phase Difference by Graphs



a In Phase

b Out of Phase

Trigonometrical Representation

19. **Angular Velocity**. Normally the size of angles is expressed in degrees. When dealing with rotation another unit of measurement, the radian, is used. A radian is the angle subtended at the centre of a circle by an arc of the circumference equal in length to the radius of that circle. As the total length of the circumference is 2πr and a complete revolution is 360°.

$$360° = 2\pi \text{ radians, or } 1 \text{ radian} = 57.3°$$

20. **Instantaneous Values of e and i**. Using radians, expressions for the instantaneous values of e and I, which are related to time, can be derived. In one revolution, the loop passes through 360° or 2π radians. If it rotates at f revolutions per second it passes through 2πf radians per second; this is termed the angular velocity of the loop and is denoted by the Greek letter ω. Thus:

angular velocity = ω (radians per sec)

After an interval of t seconds from the commencement of rotation the loop has rotated through an angle θ equal to 2πft = ωt radians, and the emf at this instant is:

$$e = E_0 \sin\theta$$

$$= E_0 \sin\omega t \text{ (volts)}$$

Similarly, with an alternating current in a circuit, the value of the current at any instant 't' is:

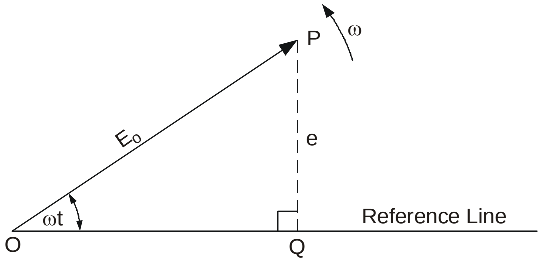
$i = I_0 \sin \omega t$ (amps)

21. **Phase Difference**. The information in the graph of Fig 9b can also be conveyed by trigonometrical equations. For instance, the pair of equations $i_1 = I_0 \sin \omega t$ and $i_2 = I_0 \sin (\omega t - \theta)$ indicate that the two currents i_1 and i_2 have the same amplitude I_0 and the same angular velocity, ω, but i_2 is lagging i_1 by θ radians.

Vectorial Representation

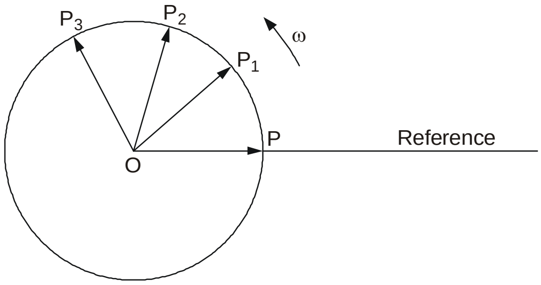
22. The trigonometrical representation of voltages and currents, described in para 20, can be depicted by vectors. In the special application of vectors to alternating quantities, the vector is assumed to be fixed at its origin but free to rotate at the frequency of the alternations, the length of the vector being equal to the peak value of the alternating quantity. For example, consider a voltage $e = E_0 \sin \omega t$. This voltage may be represented by a vector of length corresponding to E_0 rotating in an anti-clockwise direction with reference to a datum or reference line, with angular velocity ω radians per second. This is shown in Fig 10 where, after time 't' seconds, the vector has rotated through an angle of ωt radians. The instantaneous value of e is shown to the same scale as the vector E_0 by the line PQ (see para 17).

14-7 Fig 10 Vector Representation of AC



23. While the voltage E_0 throughout one cycle could be represented by a complete series of lines OP, OP_1, OP_2, etc (Fig 11), it is usually sufficient to show one position of the vector, for its rotation with angular velocity ω is understood. It is usual to take the instant t = 0 for the voltage $e = E_0 \sin \omega t$ as the reference line.

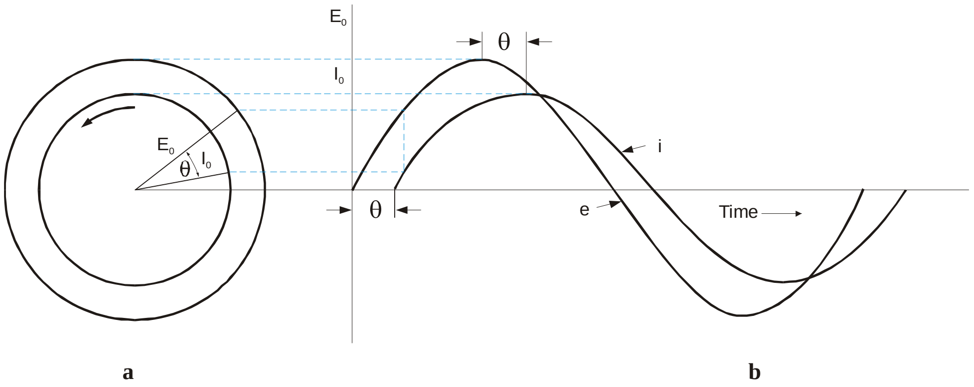
14-7 Fig 11 Vector Convention



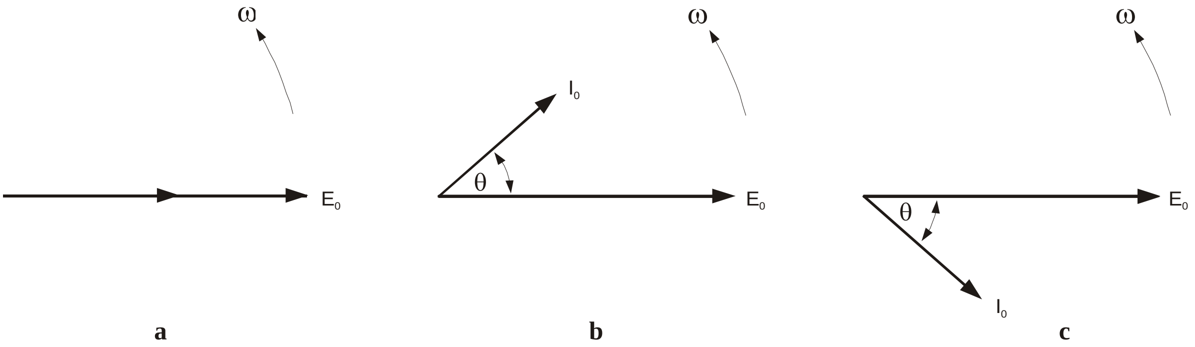
24. Figs 12a and 12b both show the phase difference between a current and its related voltage, but the vector representation is simpler to construct and use. In Fig 13, let E_0 represent the peak value of an alternating voltage, and I_0 the peak value of the current in a circuit. Then, if the voltage and current are in phase, E_0 and I_0 may be represented by vectors which are coincident in direction and rotate at equal angular velocity ω radians per second. E_0 and I_0 are therefore represented as in Fig 13a. In

practice I_0 may be alternatively in phase with, or lead, or lag the voltage E_0. In Fig 13b, I_0 is shown as leading by angle θ, and in Fig 13c as lagging by angle θ with reference to E_0.

14-7 Fig 12 Phase Difference by Graphs and Vectors

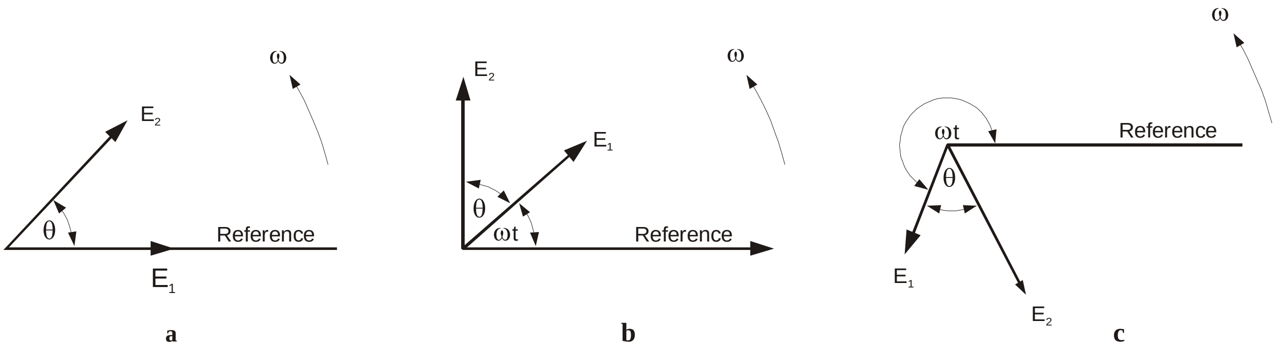


14-7 Fig 13 Phase Difference by Vector



25. **Reference Vectors**. Consider two voltages $e_1 = E_1 \sin \omega t$ and $e_2 = E_2 \sin (\omega t + \theta)$. These two voltages can be represented by vectors of length E_1 and E_2 respectively, rotating with the same angular velocity ω radians per second and displaced from each other by an angle θ. At the instant t = 0, E_1 is lying along the reference vector and E_2 is leading by an angle θ, as shown in Fig 14a. Since both vectors are rotating at the same angular velocity, they maintain the same relative positions as shown in Figs 14b and 14c. In AC problems, it is sufficient to consider vectors in one position only and it is usual to take the instant t = 0 as depicted in Fig 14a when one of the vectors lies along the reference line.

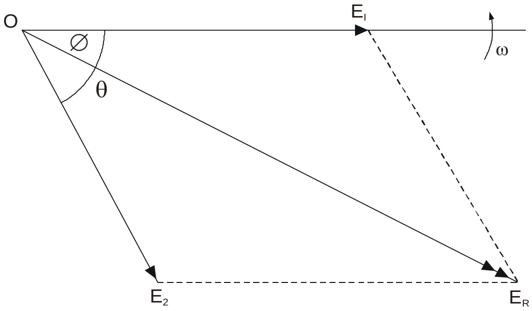
14-7 Fig 14 Reference Vectors



26. **Resultant of Two Voltages**. A single voltage, equivalent to any two alternating voltages acting across the same circuit at the same time, can be found by the parallelogram rule for finding the resultant of two vectors. Thus, in Fig 15, E_1 and E_2 are the vectors representing two voltages

$e_1 = E_1\sin \omega t$ volts and $e_2 = E_2\sin (\omega t - \theta)$ volts respectively. The resultant voltage is represented by the diagonal of the parallelogram drawn outwards from O, ie the vector E_R. This resultant vector rotates at the same angular velocity ω as E_1 and E_2 and with reference to E_1 lags by an angle Ø. The resultant voltage is, therefore, $e_R = E_R\sin (\omega t - Ø)$ volts.

14-7 Fig 15 Resultant of Two alternating Reference Vectors



IDEAL COMPONENTS IN CIRCUITS

Introduction

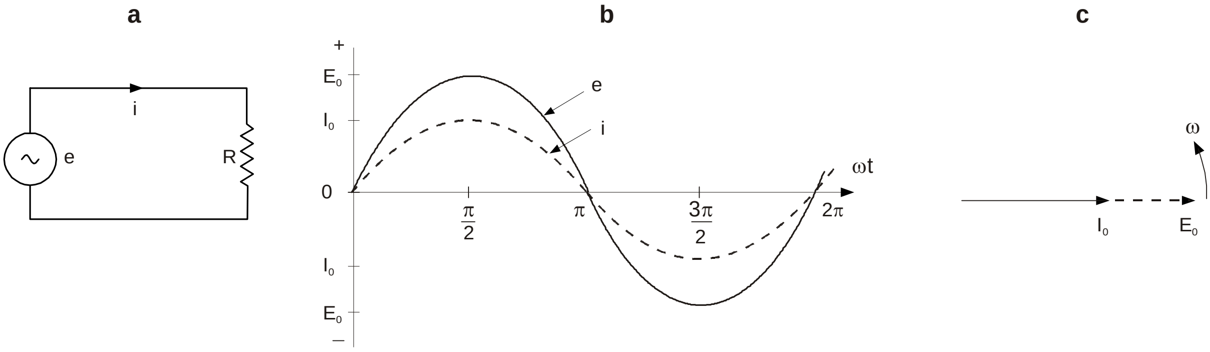
27. In the next chapter, the combined effects of resistance, inductance, and capacitance in AC circuits are examined in some detail. In this chapter a simpler approach is adopted, the assumption being made that resistance, inductance and capacitance exist separately; as has already been stated in discussing DC circuits this is an idealized approach, inductors for example having some resistance, and wire-wound resistors some inductance.

28. When purely resistive components are used in AC circuits, Ohm's Law, Kirchoff's Laws, and the usual circuit rules apply exactly as in DC circuits, noting that in general cases rms values of current and voltage are used.

Pure Resistive Circuits

29. When an alternating voltage is applied across a resistance (Fig 16a), the waveform of the voltage is a sine wave. As Ohm's Law applies at any instant, the current flow follows the same waveform, increasing and decreasing with the voltage and changing direction with it (Fig 16b). The voltage and current are said to be in phase (sine waves being in phase when they have the same frequency, and pass through zero together in the same direction, although they do not necessarily have the same amplitude).

14-7 Fig 16 A Resistive AC Circuit



30. **Current**. Since Ohm's Law applies at all times to a purely resistive circuit, the rms value of current is given by:

$$I_{rms} = \frac{E_{rms}}{R}$$

31. **Power**. The power in AC circuits is the average value of all the instantaneous values for a complete cycle. The instantaneous power is the product of e and i at that instant. If the multiplying process is carried out over a complete cycle, the power curve at Fig 17 is obtained (note that as e and i are both negative together, their product is always positive). The total power over one cycle is represented by the area beneath the power curve; the same area lies beneath the straight-line AB drawn midway between maximum and minimum values of power. The average power over a complete cycle is thus half the peak power. Note that:

a. The power waveform has twice the frequency of the supply.

b. Maximum power $(P_0) = E_0 \times I_0$

c. Average power $P_{average} = \frac{E_0 \times I_0}{2} = \frac{E_0}{\sqrt{2}} \times \frac{I_0}{\sqrt{2}}$

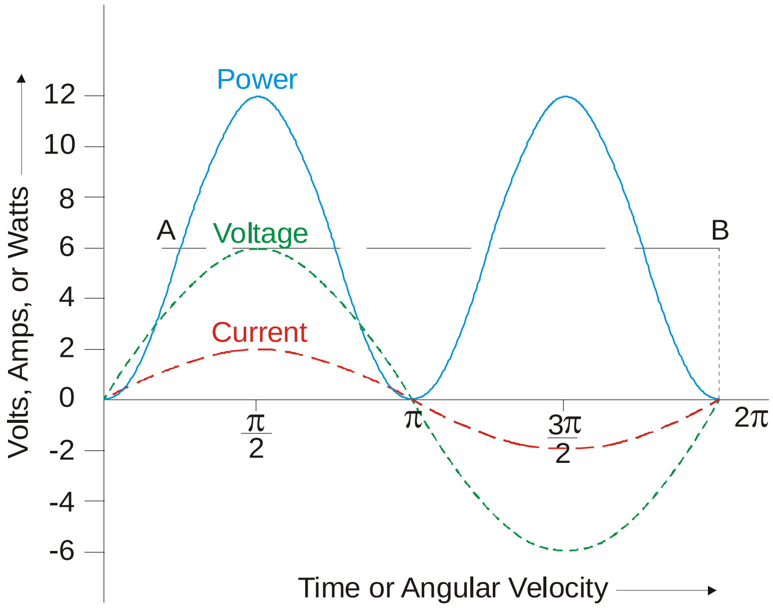
$= E_{rms} \times I_{rms}$

d. Also, since $E_{rms} = I_{rms} \times R$,

$$P_{average} = I^2_{rms} \times R$$

$$= \frac{E^2_{rms}}{R}$$

14-7 Fig 17 Power Waveform - Resistive AC Circuit



32. **Deduction**. The observations made in paras 29 and 30 can be deduced using the trigonometrical approach:

Applied voltage (e) = $E_0 \sin \omega t$ volts

From Ohm's Law,

$$\text{current (i)} = \frac{e}{R} = \frac{E_0 \sin \omega t}{R} \quad \text{amps}$$

These equations show that i is a sinusoidal current in phase with e and having the same frequency f = ω/2π Hz. The peak value of i occurs when e = E_0 volts (ie when sin ωt = 1), thus E_0 /R = I_0 amps. Thus:

a. The current and voltage are in phase.

b. The ratio $\frac{E_0}{I_0} = \frac{E_{rms}}{I_{rms}} = R$ ohms

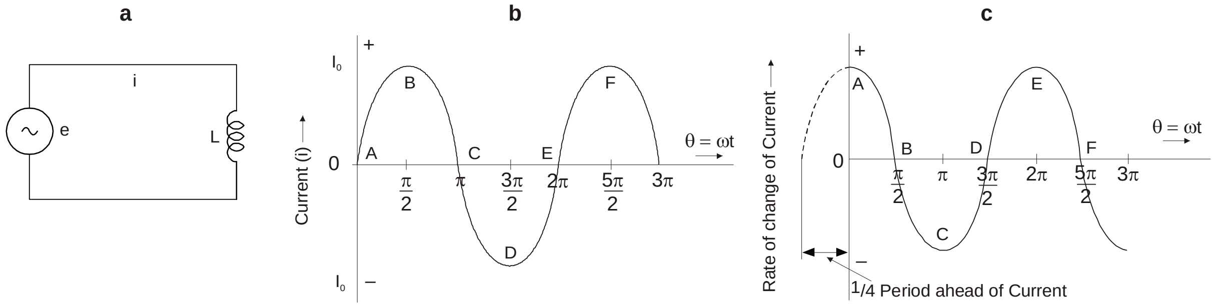
Fig 16c portrays the current and voltage vectorially.

Pure Inductive Circuits

33. It was shown in Volume 14, Chapter 3 that a change of flux through a coil induces a voltage in it, and that a change of flux through a coil can be produced merely by varying the current in it. In steady DC circuits the current changes only when the circuit is being switched on and off; in AC circuits the current is continuously changing, and thus a voltage is induced in a coil all the time AC supply is connected. It was noted that the voltage induced (the back emf) acts in opposition to the applied voltage, thus opposing any current change.

34. The back emf induced in a coil by a changing current is given by e = –L di/dt, where di/dt is an expression meaning the rate of change of current with respect to time and L is the inductance of the coil. Fig 18b shows the curve for a sinusoidal current i = I_0 sin ωt which is established in the circuit of Fig 18a. The slope of the curve is horizontal at the points B, D and F, when θ = π/2, 3 π/2, 5 π/2. Therefore, the rate of change of current is zero at these instants, giving points B, D, and F in Fig 18c. The curve has maximum and equal gradients at the points A, C and E of Fig 18b, when θ = 0, π and 2π. At these instants, the rate of change has a maximum value. At A and E, the current is positive-going, and the rate of change has a maximum positive value; at C, the current is negative-going and the rate of change has a maximum negative value. Joining the points A, B, C, D, E, and F by a curve as in Fig 18c gives the curve for the rate of change of current. Such a curve is itself a sine wave leading the current waveform by π/2 radians or 90º.

14-7 Fig 18 Rate of Change of Current - Inductive AC Circuit



35. **Voltage and Current Phases**. For a pure inductance the back emf is e = –L di/dt volts; it is in anti-phase to the rate of change of current (by virtue of the negative sign). The applied emf is, however, equal and opposite to the back emf (Kirchoff's second law) and is given by e = +L di/dt volts; it is, therefore, in phase with the rate of change of current. Hence, since the rate of change of current leads the current by π/2 radians, the applied emf is a sinusoidal wave leading the current by π/2 radians. Therefore, in a pure inductance the current and voltage are 90º out of phase, the current lagging the applied voltage by π/2 radians or 90º. This is shown graphically in Fig 19 and vectorially in Fig 20.

36. **Inductive Reactance**. In a pure resistance the ratio of voltage to current gives the resistance R. In a pure inductance the ratio of voltage to current is:

$$\frac{E_0}{I_0} = \frac{E}{I} = \omega L = 2\pi f L \ \text{(ohms)}$$

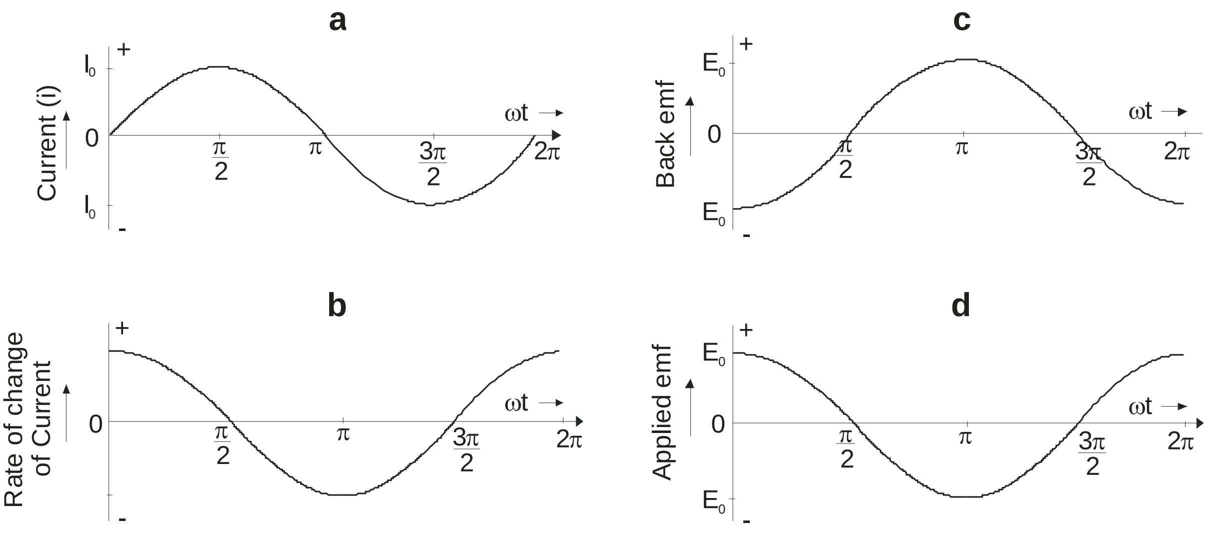
Thus, $I = \frac{E}{2\pi f L}$ (amps)

Hence, in a circuit having inductance only, the current is directly proportional to the applied voltage and inversely proportional to the frequency and the inductance. The opposition offered by pure inductance to the establishment of a current is termed the inductive reactance. The term reactance is used instead of resistance because, as can be seen from the formula, the opposition depends upon frequency. Inductive reactance is denoted by the symbol X_L and is expressed in ohms. Thus:

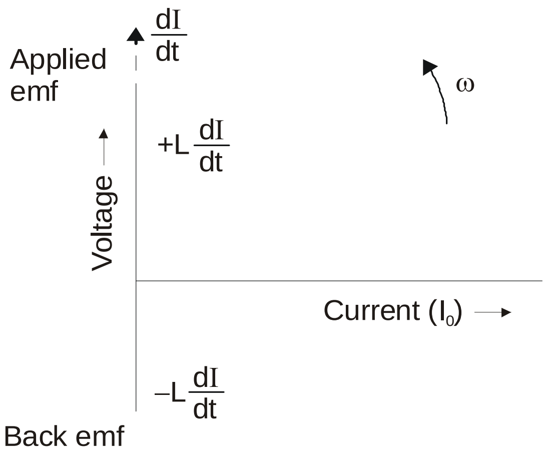
X_L = E/I = 2πfL,

where X_L = inductive reactance in ohms,
f = frequency in hertz, and
L = inductance in henrys.

14-7 Fig 19 Phase Relationships in an Inductive AC Circuit



14-7 Fig 20 Vector Relationships in an Inductive AC Circuit



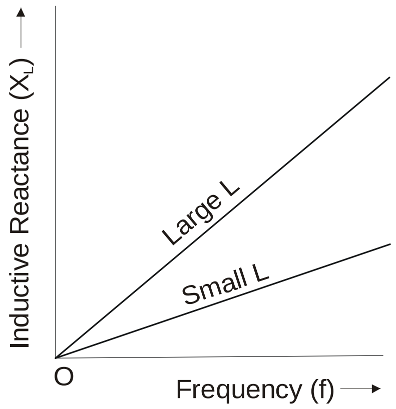
37. **The Reactance Sketch**. The reactance of an inductance is proportional to ω and so directly proportional to the frequency f, since ω = 2 πf. A graph of inductive reactance against frequency for two different values of inductance is at Fig 21. This graph is known as a reactance sketch, and for a pure inductance is a straight line through the origin at O. Thus, for a 1 mH inductance:

when f is 50 Hz, $X_L = 2\pi \times 50 \times 10^{-3} = 0.314\ \Omega$

when f is 1 kHz, $X_L = 2\pi \times 10^3 \times 10^{-3} = 6.28\ \Omega$

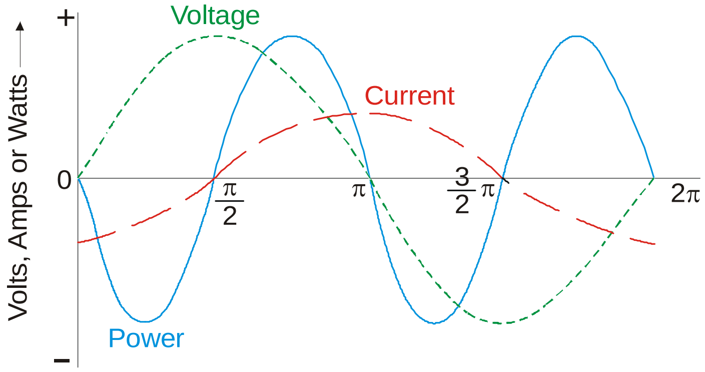
when f is 1MHz, $X_L = 2\pi \times 10^6 \times 10^{-3} = 6{,}280\ \Omega$

14-7 Fig 21 Variation of Inductive Reactance with Frequency



38. **Power**. The power curve for a purely inductive circuit is derived as usual by multiplying together corresponding values of current and voltage to find the instantaneous power and plotting the products (Fig 22). It will be seen that because current and voltage are not in phase the power is sometimes positive and sometimes negative, that the frequency of the power curve is twice that of the supply, that positive power equals negative power, and that true power equals zero. Meanings can be given to the terms positive and negative power. Power is reckoned positive when energy is being taken from the generator to be stored in the magnetic field of the inductor, and negative when power is being extracted from this magnetic field and returned to the generator.

14-7 Fig 22 Power Curve of a Purely Inductive Circuit



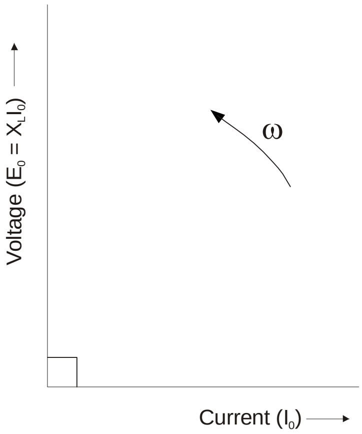
39. **Conclusions**. In a purely inductive circuit:

a. The current lags the applied voltage by 90º (π/2 radians).

b. The ratio of voltage to current gives the inductive reactance.

These two statements are illustrated by the vector diagram at Fig 23.

14-7 Fig 23 Phases of Current and Voltage in a Pure Inductance



Pure Capacitive Circuits

40. In Fig 24:

a. A pure capacitance is shown in Fig 24a connected across an alternating voltage supply.

b. The applied voltage is e = E0sin ωt, represented graphically at Fig 24b.

c. From the general expression Q = CV (where Q is the charge in coulombs, C the capacitance in farads, and V the applied emf in volts), the charge on the capacitor at the time t is given by q = CE0sin ωt, represented graphically at Fig 24c, the charge being in phase with the applied voltage.

d. The current i at time t is the rate of change of charge (dq/dt) at that instant. Using the arguments of para 34, the curve for the rate of change of charge can be shown to be itself sinusoidal, leading the curve for charge (and therefore the curve for applied voltage) by π/2 radians or 90º (Fig 24d).

41. The conclusion that the curve for voltage lags behind the curve for current is expected if one considers that when the supply voltage is applied to a capacitive circuit the current rises to a maximum value immediately. However, as the voltage builds up on the plates of a capacitor so the opposition (reactance) increases, thus delaying the voltage rise.

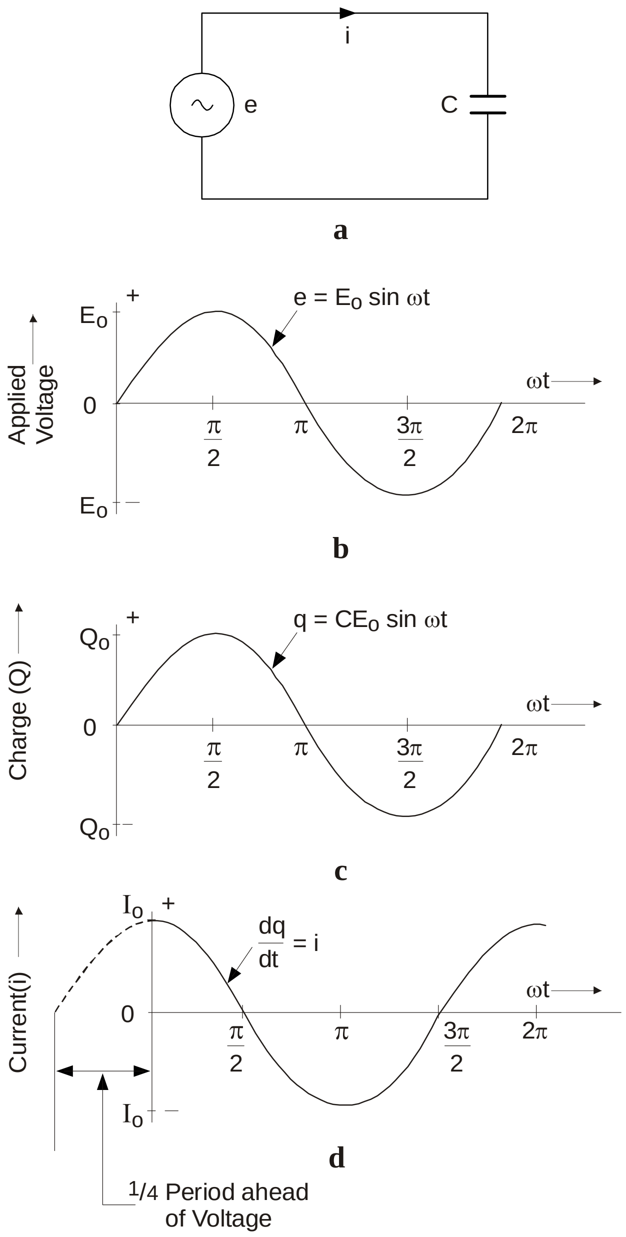
42. In a pure capacitance the ratio of voltage to current is:

$$\frac{E_0}{I_0}=\frac{E}{I}=\frac{1}{\omega C}=\frac{1}{2\pi\ fC}$$ or, transposing

$$I = E2\pi fC$$

That is, in a circuit having capacitance only, the current is directly proportional to the applied voltage, the frequency, and the capacitance.

14-7 Fig 24 Phase Relationship in a Capacitance



43. **Capacitive Reactance**. It will be recalled, from Volume 14, Chapter 4, that capacitance opposes any change in the value of voltage applied to a capacitor, thus when the applied voltage is alternating the capacitor presents an opposition at all times. This opposition is called capacitive reactance (X_C), and, as with inductive reactance, the term reactance is used rather than resistance because the opposition to the change depends upon frequency.

44. Capacitive reactance is expressed in ohms, and is given by:

$$X_C = \frac{E}{I} = \frac{1}{2\pi fC}$$

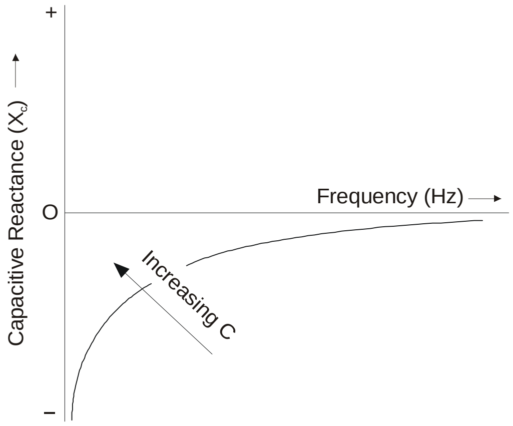
Thus, capacitive reactance is inversely proportional to frequency. A graph of capacitive reactance against frequency is shown at Fig 25. The reactance of a capacitance is assumed negative (opposite to that of an inductance) and is normally given as $X_C = 1/\omega C$; it decreases as the capacitance increases, and as the frequency increases. Thus, for a 2μF capacitor:

when f is 50 Hz, $X_C = \frac{1}{2\pi fC} = \frac{10^6}{2\pi \times 50 \times 2}$ = 1,600 Ω

when f is 1 kHz, $X_C = \frac{10^6}{2\pi \times 10^3 \times 2}$ = 80 Ω

when f is 1 MHz, $X_C = \frac{10^6}{2\pi \times 10^6 \times 2}$ = .08 Ω

14-7 Fig 25 Variation of X_C with Frequency



45. **Power**. The power curve for a purely capacitive circuit is derived as usual by multiplying together corresponding values of current and voltage to find the instantaneous power and plotting the products. The power curve is similar to that derived in Fig 22 for an inductive circuit, but, as in this case current leads voltage, inverted. Again, the total power is zero.

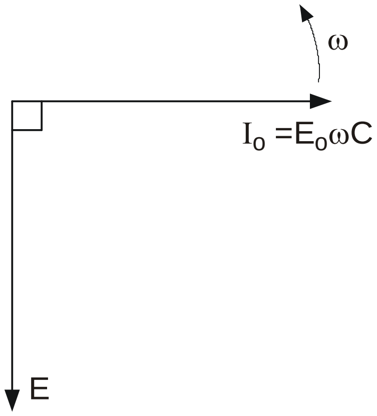
46. **Conclusions**. In a purely capacitive circuit:

a. The current leads the voltage by 90º (π/2 radians).

b. The ratio of voltage to current gives the capacitive reactance X_C.

These two statements can be shown by drawing a vector diagram as at Fig 26.

14-7 Fig 26 Phases of Current and Voltage in a Pure Capacitance



Summary

47. The relative phase of current and voltage in a capacitor and in an inductor can be remembered from the word 'CIVIL', where C is the capacitance, I the current, V the voltage, and L the inductance. CIVIL then indicates that in a capacitor the current leads the voltage; in an inductor, the current lags the voltage.

CHAPTER 8 - AC CIRCUITS

Introduction

1. In the last chapter ideal components in AC circuits were examined. In this chapter components in combination as they actually occur in circuits are discussed.

2. Circuits containing combinations of resistance, inductance and capacitance in series and in parallel are considered in detail, and the phase and magnitude relationships of circuit voltages and currents are illustrated for various typical circuits.

Notes:

1. Unless otherwise specified it should be assumed that rms values of current and voltage are intended.

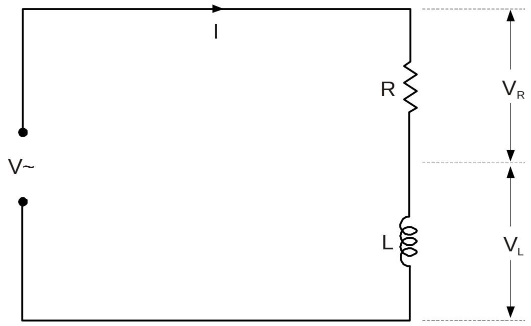
2. As this is a general theoretical treatment, units are not always specified. It should be assumed that quantities are in compatible units, viz current in amperes, emf and PD in volts, frequency in hertz, resistance in ohms, inductance in henrys, and capacitance in farads.

SERIES CIRCUITS

Resistance and Inductance in Series

3. In Fig 1 a sinusoidal alternating voltage of rms value V volts and frequency f hertz is applied across a circuit consisting of resistance R ohms and inductance L henrys connected in series. A corresponding rms current I amperes is established in the circuit. With components in series, the current is the same at all points in the circuit, and, for this reason, the current vector is taken as the reference when constructing a vector diagram.

14-8 Fig 1 R and L in Series



4. The vector diagram for this circuit is shown in Fig 2 and is constructed in the following manner:

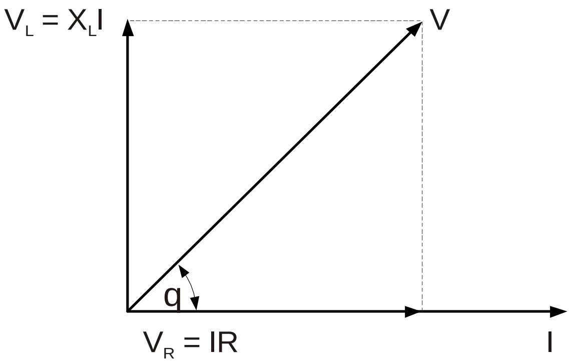
a. The current vector I is drawn to form the reference line.

b. The voltage drop across R is the product of the current and the resistance. It is in phase with the current and has an rms value $V_R = IR$. This vector is drawn to any convenient scale in line with the current vector.

c. The voltage across L is the product of the current and the inductive reactance X_L. It leads the current by 90º and has an rms value $V_L = IX_L$. This vector is drawn to the same scale as V_R, leading the current by 90º.

d. The applied voltage V is found by applying the analogous parallelogram of forces rule. This voltage is seen to lead the current by an angle θ (the phase angle).

14-8 Fig 2 Phase Relationships, R and L in Series



5. **Application of Pythagoras' Theorem**. The magnitude and phase of the applied voltage V is found by considering the triangle ABC in Fig 3. This is a right-angled triangle to which Pythagoras' theorem applies. Thus, from Fig 3:

$$V^2 = V_R^2 + V_L^2, \text{ and } V = \sqrt{V_R^2 + V_L^2}$$

Alternatively, substituting for V_R and V_L,

$$V^2 = I^2R^2 + I^2X_L^2$$

$$\therefore V = I\sqrt{R^2 + X_L^2}$$

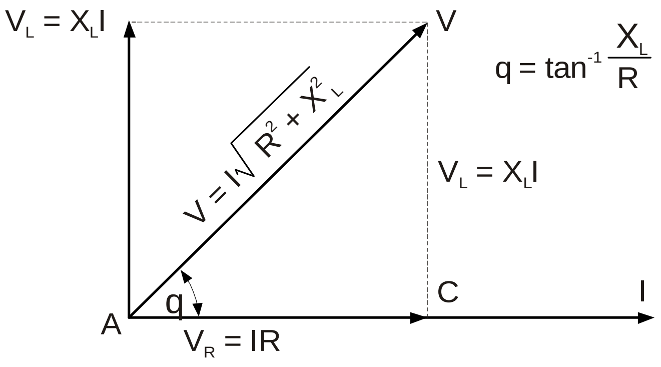
Also, V leads I by an angle θ where:

$$\tan\theta = \frac{V_L}{V_R} = \frac{X_L}{R}$$

$$\therefore \theta = \tan^{-1}\frac{\text{Re actance}}{\text{Re sistance}}$$

(Where $\tan^{-1}$ means "the angle whose tangent is given by".)

14-8 Fig 3 Magnitude and Phase of V

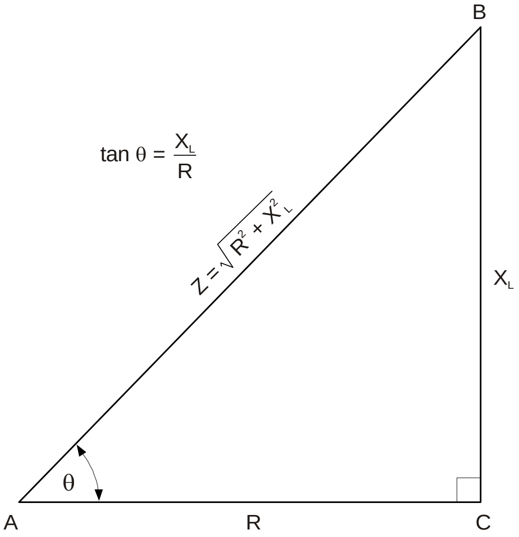


6. **Impedance**. The total opposition to current in a circuit containing resistance and reactance in combination is termed impedance (symbol Z) and is measured in ohms. For the whole circuit I = V/Z. Using Z = V/I, which applies to any circuit, for a series circuit containing resistance and inductance where the voltage V is $I\sqrt{R^2 + X_L^2}$ the impedance is:

$$Z = \frac{V}{I} = \sqrt{R^2 + X_L^2}$$

7. **The Impedance Triangle**. Although Z is composed of both resistance and reactance, it is not merely the sum of these, as the phases of current and voltage in each of them are different. An impedance triangle can be constructed from the vector diagram of Fig 2 by dividing each voltage vector by I. The result is shown in Fig 4, where the base represents the resistance, the perpendicular the reactance, and the hypotenuse the impedance. The tan of the phase angle Z is given by XL/R.

14-8 Fig 4 Impedance Triangle, R and L in Series



8. **Summary**. In a circuit containing resistance and inductance in series:

a. The applied voltage V = $\sqrt{V_R^2 + V_L^2}$ volts.

b. The magnitude of the current I = $\frac{V}{Z} = \frac{V}{\sqrt{R^2 + X_L^2}}$ amperes.

c. The phase of the applied voltage relative to the current is found from tan θ =X_L/R, tan θ being positive indicating that V leads I.

d. The impedance of the circuit is Z = $\sqrt{R^2 + X_L^2}$ ohms, where X_L (the inductive reactance), equals 2πfL (i.e. ωL).

Resistance and Capacitance in Series

9. With a current of rms value I amperes established in the circuit of Fig 5, components of the applied voltage V appear across R and across C. The vector diagram (Fig 6) is constructed as follows:

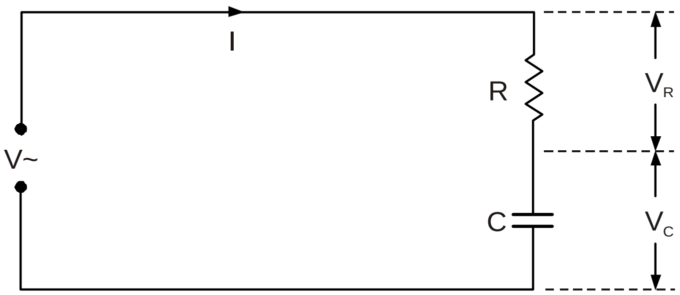
a. The current I is common to R and C and the vector I is the reference vector.

b. The voltage across R is V_R = IR, and is in phase with I.

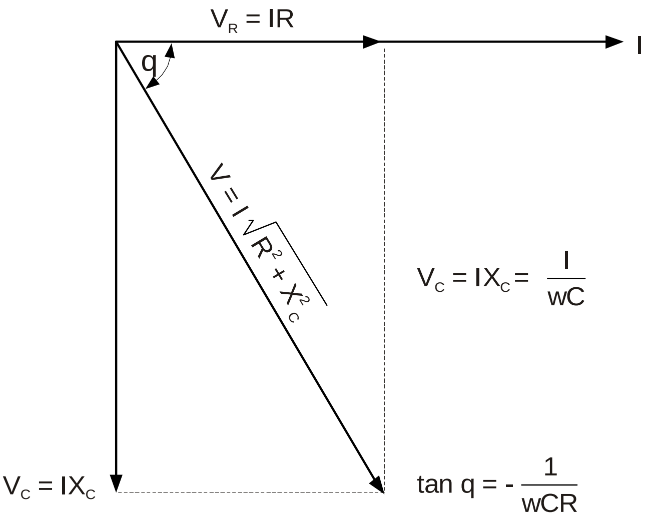
c. The voltage across C is V_C = IX_C = –I/ωC, and it lags the current by 90º. Its magnitude decreases as the frequency increases.

d. The resultant applied voltage V is found by applying the analogous parallelogram of forces rule. This voltage is then seen to lag the current by an angle θ.

14-8 Fig 5 R and C in Series



14-8 Fig 6 Phase Relationship, R and C in Series



10. **Application of Pythagoras' Theorem**. Applying Pythagoras' theorem to Fig 6:

$$V^2 = V_R^2 + V_C^2, \text{ and } V\sqrt{V_R^2 + V_C^2}$$

Or, substituting:

$$V^2 = I^2R^2 + I^2X_C^2, \text{ and } V = I\sqrt{R^2 + X_C^2}\ .$$

Also, V lags I by an angle θ where:

$$\tan\theta = \frac{V_C}{V_R} = \frac{V_C}{IR} = \left(-\frac{1}{\omega CR}\right)$$

$$\therefore\ \theta = \tan^{-1}\left(-\frac{\text{reactance}}{\text{resistance}}\right)$$

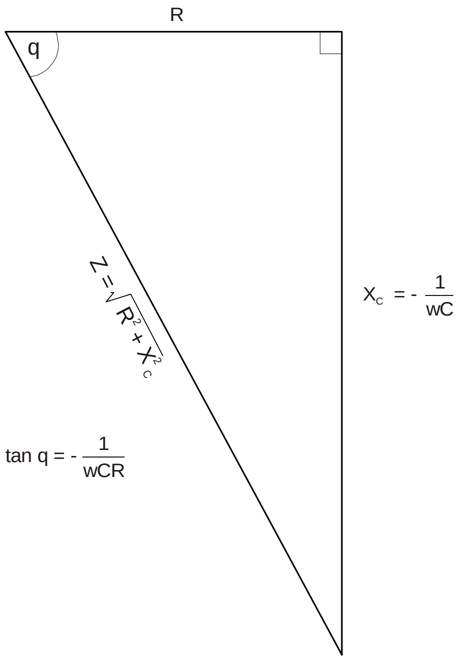
Note that tan θ is negative in a capacitive circuit, indicating that V lags I.

11. **Impedance**. The impedance of this circuit is:

$$Z = \frac{V}{I} = \sqrt{R^2 + X_C^2}$$

The corresponding impedance triangle is shown in Fig 7.

14-8 Fig 7 Impedance Triangle, R and C in Series



12. **Summary**. In a circuit containing resistance and capacitance in series:

a. The applied voltage $V = \sqrt{V_R^2 + V_C^2}$ volts.

b. The magnitude of the current is $I = \frac{V}{Z} = \frac{V}{\sqrt{R^2 + X_C^2}}$ amperes.

c. The phase of the current relative to the applied voltage is found from tan θ = X_C/R. The current leads the voltage in a capacitive circuit.

d. The impedance of the circuit is: Z = $\sqrt{R^2 + X_C^2}$ ohms, where X_C, the capacitive reactance, equals 1/2πfC.

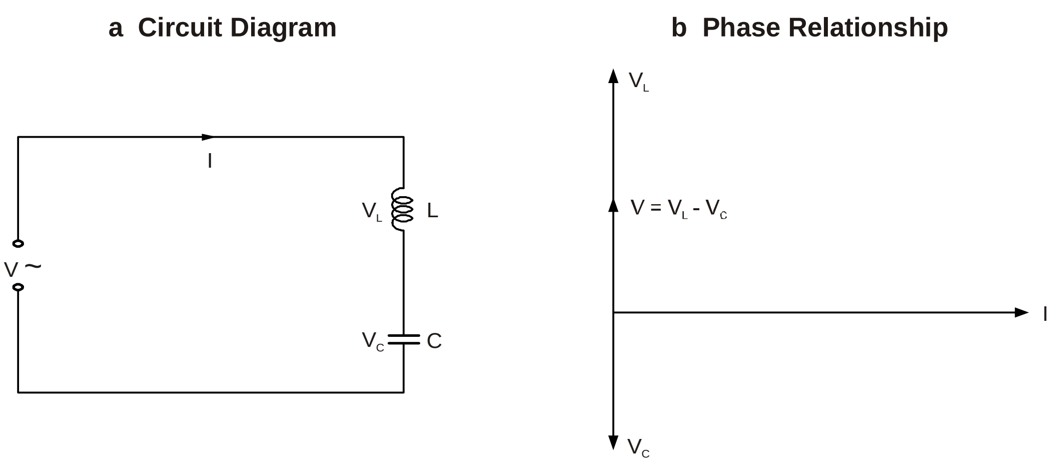
Inductance and Capacitance in Series

13. In an AC circuit containing only L and C in series (Fig 8a), the voltage V_L across the inductance leads the current by 90º, and voltage V_C across the capacitance lags the current by 90º. Thus V_L and V_C are 180º out of phase as shown in Fig 8b, and oppose each other so that the total applied voltage V is their difference. In Fig 8 V_L is taken to be larger than V_C and V is leading I by 90º. However, if the frequency of the applied voltage and values of L and C are such that $X_C > X_L$, then $V_C > V_L$ and V lags I by 90º.

14. Ohm's law applies to each part of the circuit and to the whole circuit: $V_L = IX_L$; VC = IX_C; V = IZ. Since V_L and V_C oppose each other, X_L and X_C act in opposite directions and the impedance of the

circuit is their difference. Thus Z = $X_L - X_C$, or $X_C - X_L$, depending on which is the larger. If $X_L > X_C$, then Z is an inductive reactance. If $X_C > X_L$, then Z is a capacitive reactance.

14-8 Fig 8 L and C in Series



Resistance, Inductance and Capacitance in Series

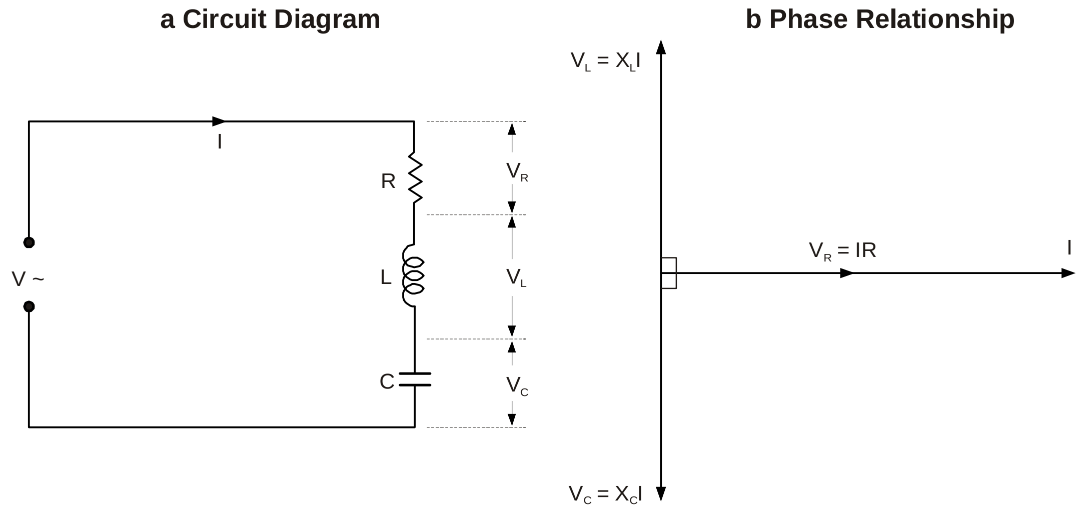
15. Fig 9a shows a circuit consisting of R ohms, L henrys and C farads connected in series across an alternating supply of frequency f Hz. With a current of rms value I amperes established in the circuit, voltages (in volts) across each component are as follows:

a. Across R, V_R = IR and is in phase with I.

b. Across L, $V_L = IX_L$ and leads I by 90º. Its magnitude increases with frequency.

c. Across C, $V_C = IX_C$ and lags I by 90º. Its magnitude decreases as the frequency increases.

14-8 Fig 9 R, L and C in Series



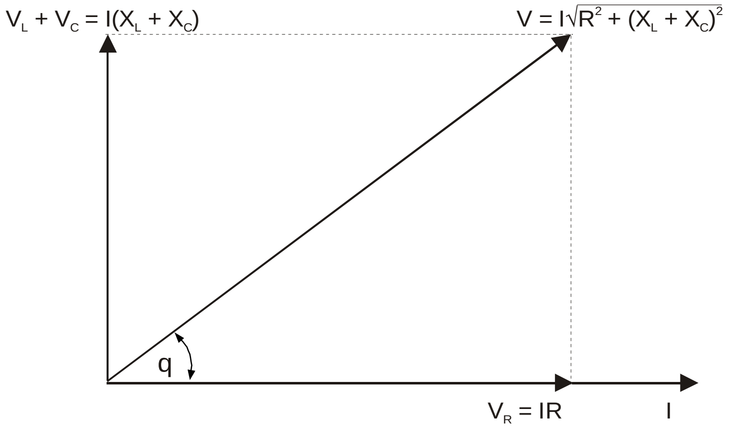
16. The resultant vector diagram is constructed as shown in Fig 9b, the current I, which is common to all components, being the reference vector. By Kirchoff's second law the vectorial resultant of V_R, V_L, and V_C must equal the applied voltage V. To obtain this:

a. Add the two vectors $V_L = IX_L$ and $V_C = IX_C$ to give a single resultant vector $I(X_C + X_L)$. Since the capacitive reactance is negative with respect to the inductive reactance and is given by:

$X_C = -\frac{1}{\omega C}$, the resultant vector is $I\left(\omega L - \frac{1}{\omega C}\right)$.

b. Apply the rule of the parallelogram of forces to vector V_R and the combined vector $(V_L + V_C)$ to obtain the resultant applied voltage vector V. This is shown in Fig 10.

14-8 Fig 10 Resultant Applied Voltage, R, L and C in Series



17. **Application of Pythagoras' Theorem**. Apply Pythagoras' theorem to Fig 10:

$$V^2 = V_R^2 + (V_L + V_C)^2 \text{ and } V = \sqrt{V_R^2 + (V_L + V_C)^2}$$

Or, substituting:

$$V^2 = I^2R^2 + (IX_L + IX_C)^2 \text{ and } V = I\sqrt{R^2 + (X_L + X_C)^2}$$

$$\therefore V = I\sqrt{R^2 + \left(\omega L - \frac{1}{\omega C}\right)^2}$$

In this case, V leads I by an angle θ where:

$$\tan\theta = \frac{V_L + V_C}{V_R} = \frac{(X_L + X_C)}{R}$$

$$\therefore \tan\theta = \frac{(\omega L - 1/\omega C)}{R}$$

$$\theta = \tan^{-1}\left(\frac{\text{effective reactance}}{\text{resistance}}\right)$$

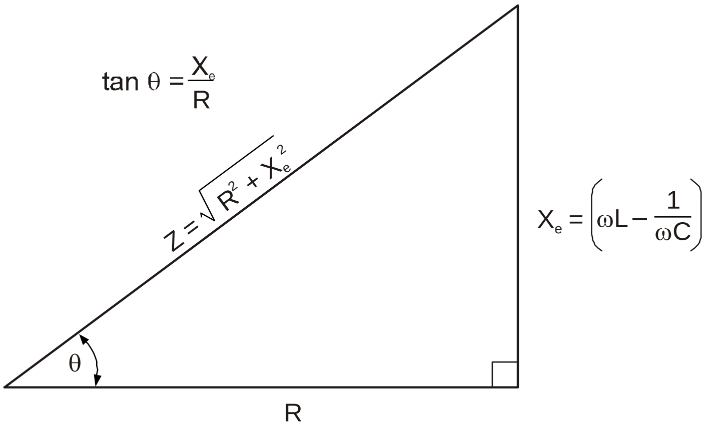
The factor tan θ is positive in this instance since X_L is greater than X_C.

18. **Impedance**. The impedance of this circuit is:

$$Z = \frac{V}{I} = \sqrt{R^2 + (X_L + X_C)^2}$$

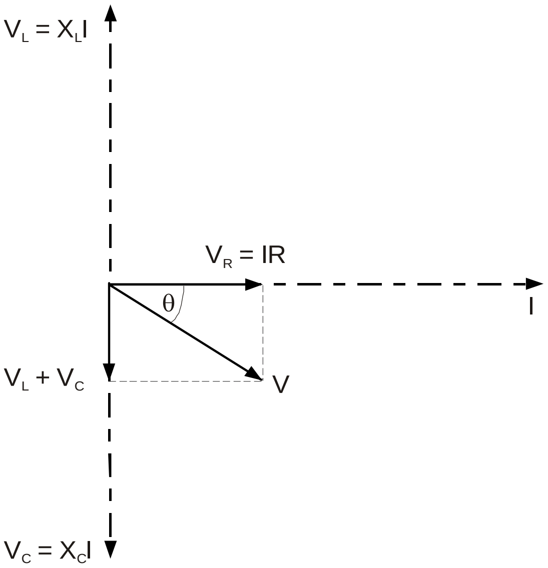
The corresponding impedance triangle is shown in Fig 11. It should be noted that X_L is a positive reactance and X_C is a negative reactance so that it is, in fact, their numerical difference which gives the magnitude of the effective reactance (X_e) in the circuit.

14-8 Fig 11 Impedance Triangle, R, L and C in Series


$\tan\theta = \frac{X_e}{R}$
$Z = \sqrt{R^2 + X_e^2}$
$X_e = \left(\omega L - \frac{1}{\omega C}\right)$
θ
R

19. **Variation of Frequency**. A variation in the frequency of the supply has an effect on the results obtained. The vector diagrams of Figs 9b and 10 apply when the frequency of the supply voltage is such that X_L is greater than X_C. As the frequency is reduced, X_L is reduced and X_C is increased. Thus the vector $V_L = IX_L$ is correspondingly reduced in amplitude, and the vector $V_C = IX_C$ increased. The condition where the frequency of the supply voltage is such that X_C is greater than X_L is shown in Fig 12, in which it is seen that the resultant voltage V now lags the current I. Thus the applied voltage V either leads or lags the current I depending on whether V_L is greater than V_C (ie $X_L > X_C$), or V_C is greater than V_L (ie $X_C > X_L$). If the voltage leads the current (tan θ positive) the circuit is 'inductive'; if the voltage lags the current (tan θ negative) the circuit is 'capacitive'.

14-8 Fig 12 Effect of Variation of Frequency



20. **Variation of L or C**. Similar results to those in the preceding paragraphs are obtained if the frequency of the supply is fixed and the value of L or C is altered. With a constant value of ω (ie of $2\pi f$) the inductive reactance ωL can be altered by varying the value of L. Similarly, a variation in C will alter the capacitive reactance ($1/\omega C$.) In this way, X_L or X_C can be varied to give resultant vector diagrams similar to Figs 10 and 11.

21. **Summary**. In a circuit containing resistance, inductance and capacitance in series:

a. The magnitude of the applied voltage is $V = \sqrt{V_R^2 + (V_L \sim V_C)^2}$, noting that V_C is negative.

b. The magnitude of the current is $I = V/Z = V/\sqrt{R^2 + (X_L \sim X_C)^2}$.

c. The phase of the applied voltage relative to the current is found from tan $\theta = (\omega L - 1/\omega C)/R$, or $(X_L \sim X_C)/R$ noting that X_C is negative, and the angle is either positive or negative depending on the relative values of X_L and X_C.

d. The impedance of the circuit is Z = $\sqrt{R^2 + (X_L \sim X_C)^2}$, noting that X_C is negative.

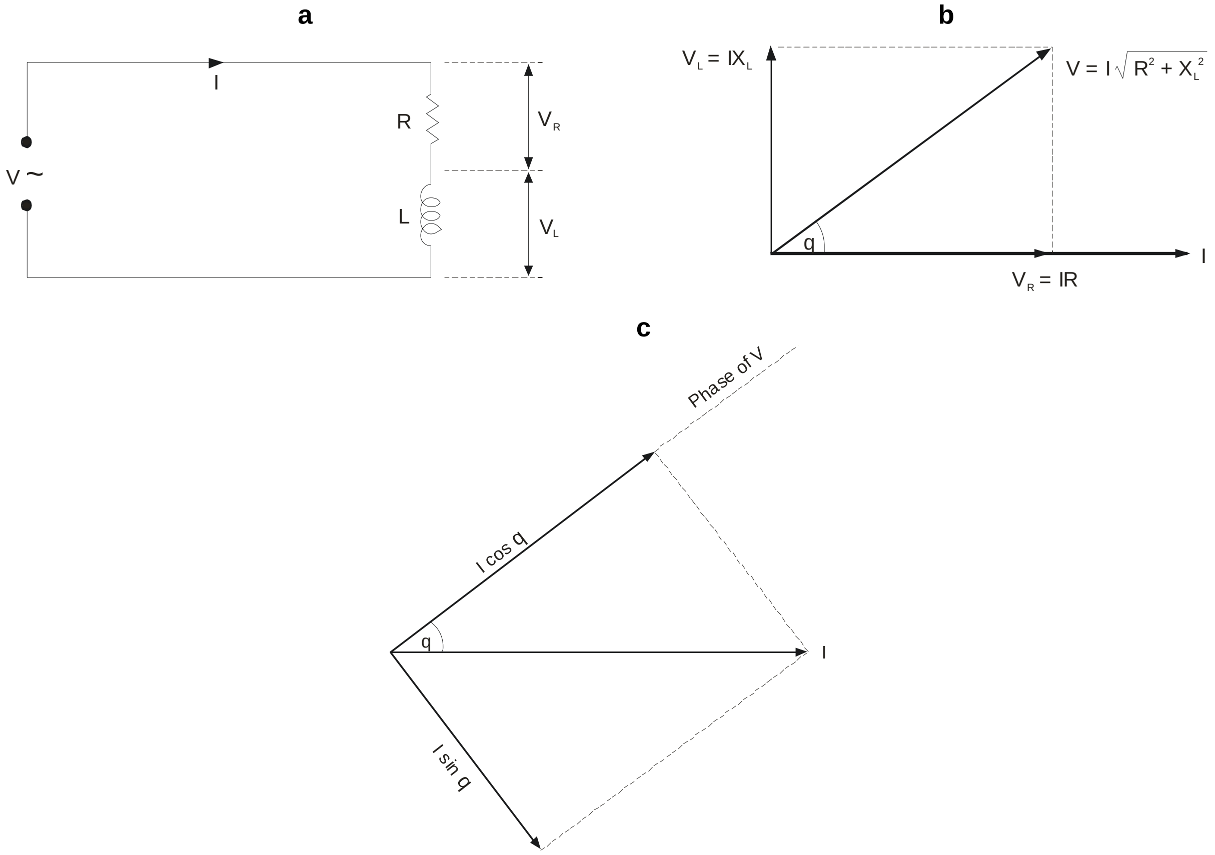
Power

22. When an alternating voltage is applied across a pure resistance R, current and voltage are always in phase. The power developed in R is VI watts, where V and I are rms values.

23. In a coil which is considered to be a pure inductance, there is no dissipation of energy as heat when the coil is connected across an alternating supply of voltage. The current and voltage are 90º out of phase, and the energy supplied ($LI^2/2$) during that part of the cycle when the current is growing to its peak value is stored in the magnetic field, and restored to the source as the current decays to zero. Similarly, in a pure capacitance the current and voltage are 90º out of phase, and the energy supplied ($CV^2/2$) to the capacitance during that part of the cycle when the voltage is building up to its peak value is stored in the electric field, and restored to the source as the voltage falls to zero. Thus, when an alternating voltage is applied across a pure reactance, either inductive or capacitive, no power is developed at all and the current is said to be 'wattless'.

24. In an actual circuit containing both resistance and reactance, there is a phase difference (θ) between the current and the voltage. The vector diagram for the circuit of Fig 13a is given in Fig 13b, from which it is seen that the current I lags the applied voltage (V) by an angle θ. It was shown in Volume 14, Chapter 7 that a vector I, at an angle θ to the reference, can be resolved into two components at right angles to each other in terms of I and θ. This is shown in Fig 13c. The component of current I which is in phase with the applied voltage (termed the in-phase component of I) is Icos θ; the out-of-phase component of I at right angles to the applied voltage is Isin θ. The former current develops a power of VIcos θ watts. The latter develops no power at all since it is 90º out of phase with the applied voltage; for this reason it is termed the wattless component of current.

14-8 Fig 13 Working and Wattless Components of Current



25. **Power Factor**. The factor cos θ is known as the 'power factor' of a circuit. The product VI is that of the rms values of voltage and current and is termed the voltamperes (VA) or the apparent power in a circuit. The true power in a circuit, where V and I are out of phase, is given by the product of the applied voltage and the in-phase component of current. Thus:

$$P = VI\cos\theta$$

$$\cos\theta = \frac{P}{VI}$$

$$\therefore \text{power factor} = \frac{\text{true power}}{\text{apparent power}}$$

The power factor may be found by measuring the true power with a wattmeter, and the apparent power with a voltmeter and ammeter. Referring again to Fig 13b it is seen that:

$$\text{voltage } V = I\sqrt{R^2 + X_L^2} = IZ.$$

$$\therefore \cos\theta = \frac{IR}{IZ} = \frac{R}{Z} = \text{power factor}$$

This gives an alternative expression for the power factor in terms of the components in a circuit. Summarizing, for a series AC circuit:

a. True power = power actually consumed in the resistance of a circuit = VI cos θ.

b. Apparent power = volts × amperes = VI.

c. Power factor = $\frac{\text{true power}}{\text{apparent power}} = \cos\theta = \frac{R}{Z}$.

d. When V and I are in phase, θ = 0, cos θ = 1, and true power = apparent power.

e. When V and I are 90º out of phase, θ = 90º, cos θ = 0, and true power = zero.

26. **Adjustment of Power Factor**. In a pure resistance, the power factor cos θ = R/Z = 1. In a pure reactance, it is zero. In a practical circuit, it lies between these extremes. The value of power factor required by a circuit depends on the purpose of the circuit. Where the power developed must be kept to a minimum, the power factor should be as nearly zero as possible (ie R should be small). Where a large power is required to be developed, eg electric fires, the power factor should be as near to unity as possible (i.e. X should be small).

27. **Power Factor of One**. Where a power factor of unity is required, the out-of-phase component of current must be zero (ie the reactive component must be zero). When the circuit is inductive, the current will lag the voltage by an angle θ, and the power factor will be less than unity. This can be adjusted by inserting a capacitive component of such value that it balances the inductive component. The current and voltage will then be in phase and the power factor will be unity.

28. **Power Losses in Components**. The power factor of an inductor or a capacitor should, in theory, be zero. That is, these components should be pure reactances. This cannot apply in practice however, since power losses are developed in the components. Most of the losses listed below increase with the frequency of the supply. They occur mainly in high frequency circuits, and because of this, are more important in radio engineering than in power engineering.

29. **Losses in Inductors**. Power losses occur in inductors because of:

a. Eddy current in the cores.

b. Hysteresis in the cores.

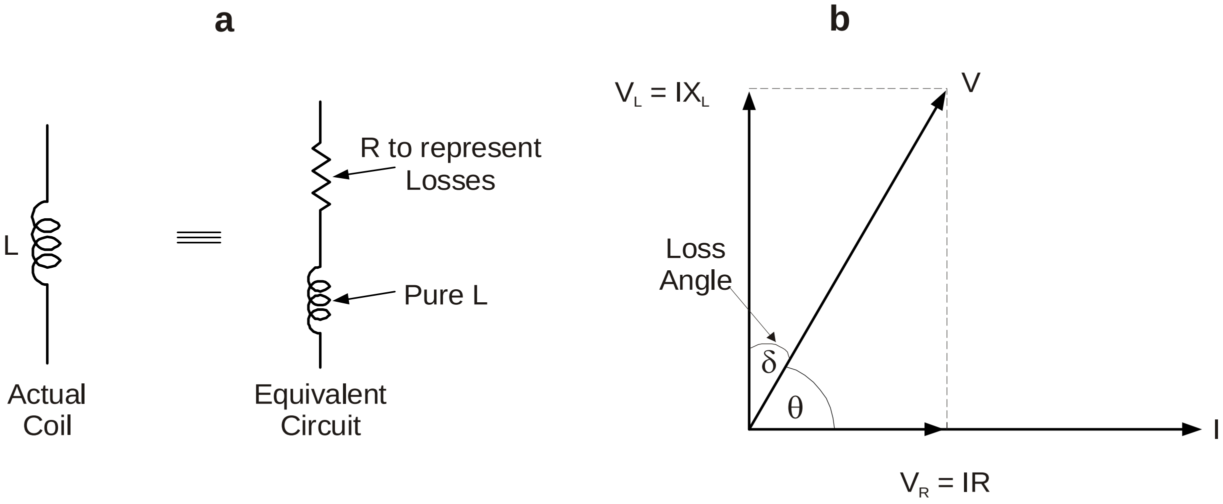
c. Skin effect in the coils.

d. Proximity effect of conductors in the near vicinity.

e. Ohmic resistance of the windings.

All of these losses can be grouped together by supposing that the inductor has in series with it a resistance R ohms, such that the power loss is I^2R watts, where I amperes is the rms value of the current. An inductor can thus be represented as L henrys and R ohms in series, where L is a pure inductance and R represents the losses (Fig 14a). The vector diagram is then as shown in Fig 14b, from which it is seen that I no longer lags V by 90º but by an angle θ, this being less than 90º by δ (the loss angle). The inductor has, therefore, a power factor of cos θ.

14-8 Fig 14 Power Factor of a Coil



30. **Losses in Capacitors**. Power losses occur in capacitors because of:

a. Leakage through the dielectric.

b. Dielectric hysteresis or absorption.

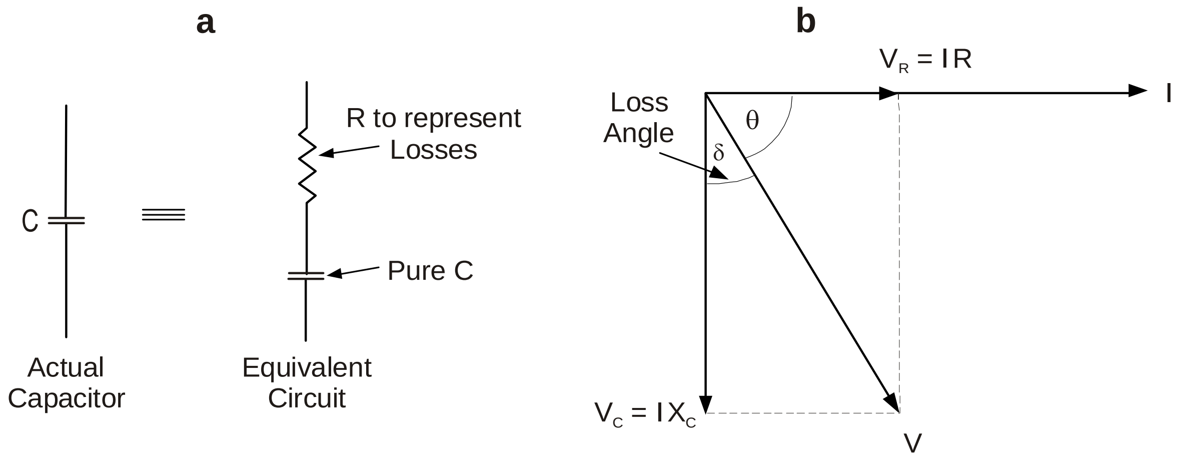
c. Brush discharge.

d. Ohmic resistance of the connecting leads.

e. Skin effect in the plates and connecting leads.

Remarks similar to those for the inductor apply to the capacitor, and a capacitor can be represented as C and R in series, where C is a pure capacitance and R represents the losses (Fig 15a). The vector diagram is then as shown in Fig 15b, from which it is seen that I leads V by an angle θ, this being less than 90º by the angle δ (the loss angle). A capacitor has, therefore, a power factor cos θ. In practice it is nearer zero than that of an inductor and in normal circuits of capacitor and inductor in series, most of the losses are in the inductor.

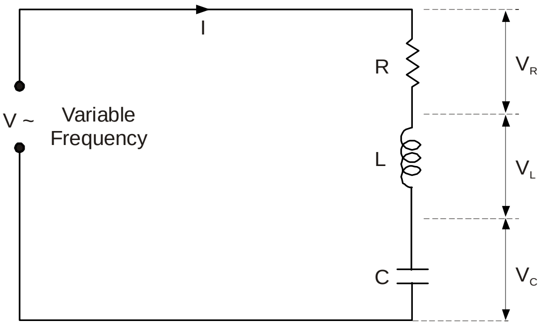
14-8 Fig 15 Power Factor of a Capacitor



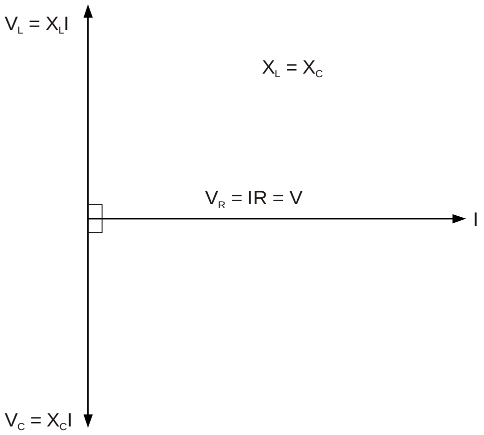
Series Circuit Resonance

31. Consider the circuit shown in Fig 16 where R is the combined loss resistance of L and C. When the frequency of the supply is such that the capacitive reactance XC is greater than the inductive reactance XL, the impedance is capacitive and the phase angle negative indicating that the applied voltage is lagging the current. When the frequency of the supply is such that XL is greater than XC, the impedance is inductive and the phase angle positive, indicating that the applied voltage is leading the current. In between these conditions, at some particular frequency, XL will equal XC, and VL will equal VC. This is shown vectorially in Fig 17. The circuit is then said to be at resonance. Where the frequency of the supply is fixed, either L or C can be varied to give the condition XL = XC. The circuit has then been tuned to resonance with the supply frequency.

14-8 Fig 16 Power Factor of a Capacitor



14-8 Fig 17 Series Resonance



32. **Conditions at Resonance**. At resonance:

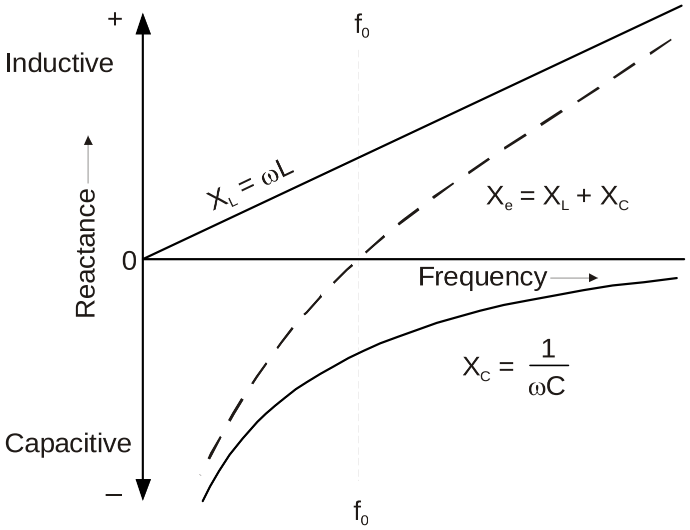
a. $\omega L = \frac{1}{\omega C}$, and the impedance Z is given by: Z $= \sqrt{R^2 + \left(\omega L - \frac{1}{\omega C}\right)^2}$ = R ohms. Thus, at resonance in a series tuned circuit, Z is a minimum and equal to R ohms. In a well-designed circuit, R will be small, being only the loss resistance of L and C.

b. The current $I = \frac{V}{Z} = \frac{V}{R}$ amperes has a maximum value.

c. The phase angle θ, given by tan θ = $\frac{(\omega L - 1/\omega C)}{R}$, is zero. Thus, at resonance in a series circuit, the current and the applied voltage are in phase.

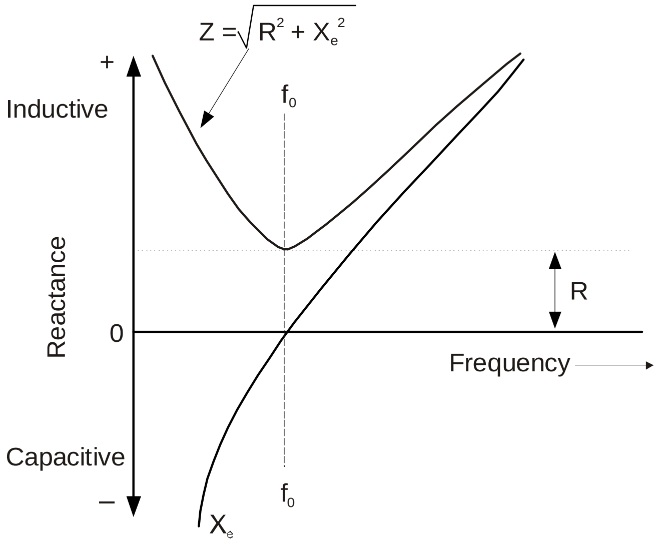
33. **Reactance Sketches**. A reactance sketch is a graph relating reactance and frequency. The reactance sketches for an inductance and for a capacitance are shown in Volume 14, Chapter 7. In a series circuit where both capacitance and inductance are present, the reactance sketches can be combined to give the total variation of reactance with frequency. This is shown in Fig 18. From this graph it is seen that at the resonant frequency (f_o), $X_L = X_C$ and the total reactance is zero. The circuit is then purely resistive with the current and the applied voltage in phase.

14-8 Fig 18 Reactance Sketch for a Series Tuned Circuit



34. **Variation of Impedance with Frequency**. If the graph for resistance R (which for practical purposes remains constant with frequency) is combined with the graph for the total effective reactance X_e, a curve for impedance Z against frequency can be obtained. To plot this graph the expression Z = $\sqrt{R^2 + X_e^2}$ ohms is used and as shown in Fig 19, Z is all above the horizontal axis. At the resonant frequency f_0 the impedance Z is minimum and equal to R ohms. Either side of resonance the impedance rises, and, as can be seen from the reactance curve, below resonance the circuit has capacitive reactance and above resonance the circuit has inductive reactance.

14-8 Fig 19 Variation of Z with Frequency



35. **Resonant Frequency**. At resonance, $X_L = X_C$, i.e.

$$2\pi f_0 L = \frac{1}{2\pi f_0 C} \quad \therefore\ f_0^2 = \frac{1}{4\pi^2 LC} \qquad \therefore\ f_0 = \frac{1}{2\pi\sqrt{LC}}$$

where f_0 = resonant frequency in hertz, L = inductance in henrys, C = capacitance in farads

This gives the resonant frequency of a series tuned circuit. By altering either L or C the frequency at which resonance occurs is changed. It should be noted that R has no effect on this.

Voltage Magnification

36. **Selectivity**. Selectivity is the property of a tuned circuit which makes it responsive to a particular frequency. Selectivity is dealt with more fully in the Annex, but it is sufficient here to note that it is represented by the symbol Q_0, and defined as the ratio of reactance to circuit resistance at the resonant frequency, i.e. $Q_0 = 2\pi f_0 L/R$ (or $\omega L/R$).

37. At resonance, in a series RLC circuit, the impedance is a minimum and equal to the resistance R. The current I is a maximum and equal to V/R. Voltages across the components shown in Fig 20a are as follows:

a. **Resistor**. The voltage across the resistance at resonance is:

$$V_R = IR = \frac{V}{R} \times R = V \text{ volts}$$

$$\therefore V_R = \text{applied voltage}$$

b. **Inductor**. The voltage across the inductance at resonance is:

$$V_L = X_L I = \omega L I = \frac{V}{R}$$

$$\therefore V_L = -\left(\frac{\omega L}{R}\right) \times V = QV \text{ volts, where}$$

$$Q_o = \frac{\omega L}{R}$$

The voltage across the inductance is Q_0 times the applied voltage V and voltage magnification has taken place.

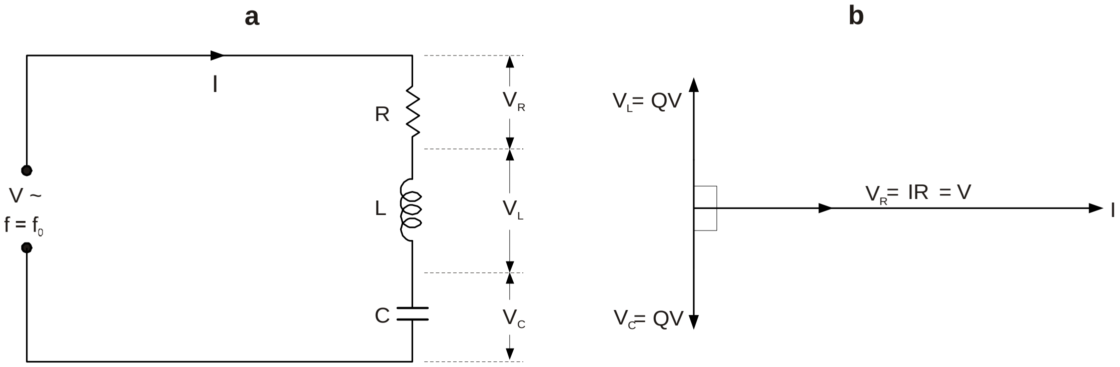
c. **Capacitor**. Q_0, as already defined, is equal to $\omega L/R$. However, at resonance, $\omega L = 1/\omega C$. Hence Q_0 is also equal to $1/\omega CR$ and the voltage across the capacitance at resonance is:

$$V_C = X_C I = \frac{I}{\omega C} = \frac{1}{\omega C} \times \frac{V}{R}$$

$$\therefore V_C = \left(\frac{1}{\omega CR}\right) \times V = QV \text{ volts}$$

Thus the voltage across the capacitance at resonance is equal in magnitude but opposite in polarity to the voltage across the inductance (Fig 20b).

14-8 Fig 20 Voltage Magnification



38. When a series circuit is resonant to a given input, the voltage across either L or C at the input frequency can be many times greater than the input voltage. If V = 0.1V and Q = 100, the voltage across either L or C is 10V and this voltage can be applied to another circuit. A resonant series circuit is, therefore, a voltage magnifier.

PARALLEL CIRCUITS

Introduction

39. In this part, circuits consisting of different types of components connected in parallel are considered. The phase and magnitude relationships for currents and voltages are discussed, as they were for the series circuits. There are important differences in the construction of vector diagrams for series and for parallel AC circuits, and these are summarized as follows:

a. Series Circuit.

(1) All components carry the same current in the same phase.

(2) The resultant applied voltage is the vector sum of the individual voltages across the separate components.

(3) Vector diagrams are constructed by drawing the voltage vectors relative to the current, which is the reference vector.

b. Parallel Circuit.

(1) All branches have the same voltage across them in the same phase.

(2) The resultant supply current is the vector resultant of the individual currents in the separate branches.

(3) Vector diagrams are constructed by drawing the current vectors relative to the voltage, which is the reference vector.

Resistance and Inductance in Parallel

40. The vector diagram for the circuit of Fig 21a is given in Fig 21b. It shows that:

a. The applied voltage V is common to both R and L and is the reference vector.

b. The current through the resistor is I_R = V/R amperes and is in phase with V. The vector I_R is drawn to any convenient scale in line with V.

c. The current through the inductor is $I_L = V/X_L$ amperes and is drawn to the same scale as I_R, lagging V by 90º.

d. The resultant supply current I is the vector sum of I_R and I_L and is found by completing the parallelogram and applying Pythagoras' theorem to Fig 21b. Thus:

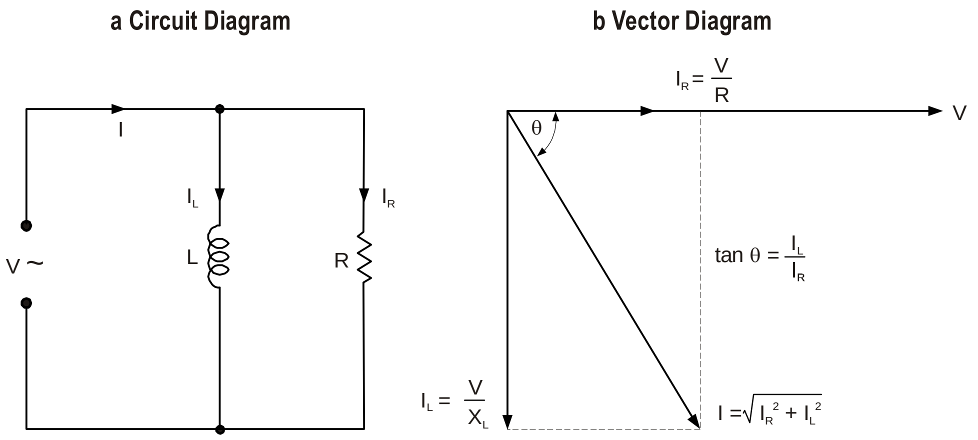
$$I = \sqrt{I_R^2 + I_L^2} = V\sqrt{\frac{1}{R^2} + \frac{1}{X_L^2}} \text{ amps}$$

e. I lags V by an angle θ, where:

$$\tan\theta = \frac{I_L}{I_R} = \frac{V/X_L}{V/R} = \frac{R}{X_L} = \frac{\text{resistance}}{\text{reactance}}$$

f. From the equation in d, $\frac{V}{I} = \frac{1}{\sqrt{\frac{1}{R^2} + \frac{1}{X_L^2}}}$ = impedance (Z), ohms.

14-8 Fig 21 R and L in Parallel



a Circuit Diagram

b Vector Diagram

Resistance and Capacitance in Parallel

41. The vector diagram for the circuit of Fig 22a is given in Fig 22b. This vector diagram is constructed as follows:

a. The applied voltage V is the reference vector.

b. The current through the resistor is $I_R = \frac{V}{R}$ amperes and is in phase with V.

c. The current through the capacitor is $I_C = \frac{V}{X_C}$ amperes and leads 90º on V.

d. The resultant supply current I is the vector sum of I_R and I_C and is found by completing the parallelogram and applying Pythagoras' theorem to Fig 22b. Thus:

$$I = \sqrt{I_R^2 + I_C^2} = V\sqrt{\frac{1}{R^2} + \frac{1}{X_C^2}} \text{ amps.}$$

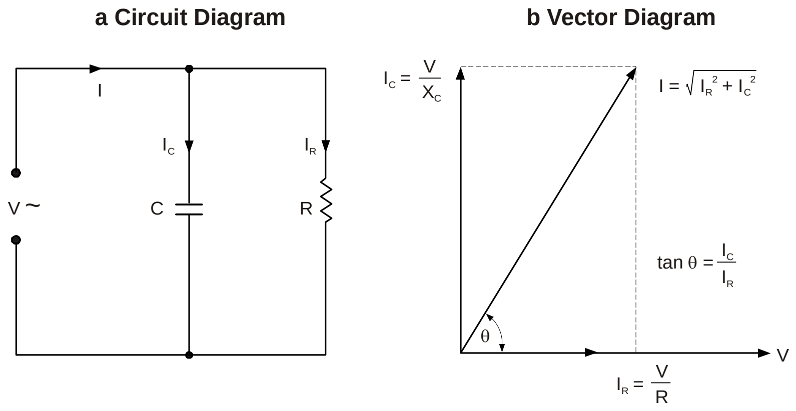
e. I leads V by an angle θ, where:

$$\tan\theta = \frac{I_C}{I_R} = \frac{V/X_C}{V/R} = \frac{R}{X_C} = \frac{\text{resistance}}{\text{reactance}}.$$

f. From the equation in d,

$$\frac{V}{I} = \frac{1}{\sqrt{\frac{1}{R^2} + \frac{1}{X_C^2}}}$$ = impedance (Z), ohms.

14-8 Fig 22 R and C in Parallel



Inductance and Capacitance in Parallel

42. Fig 23b is the vector diagram for the circuit of Fig 23a, and is constructed as follows:

a. The applied voltage V is the reference vector.

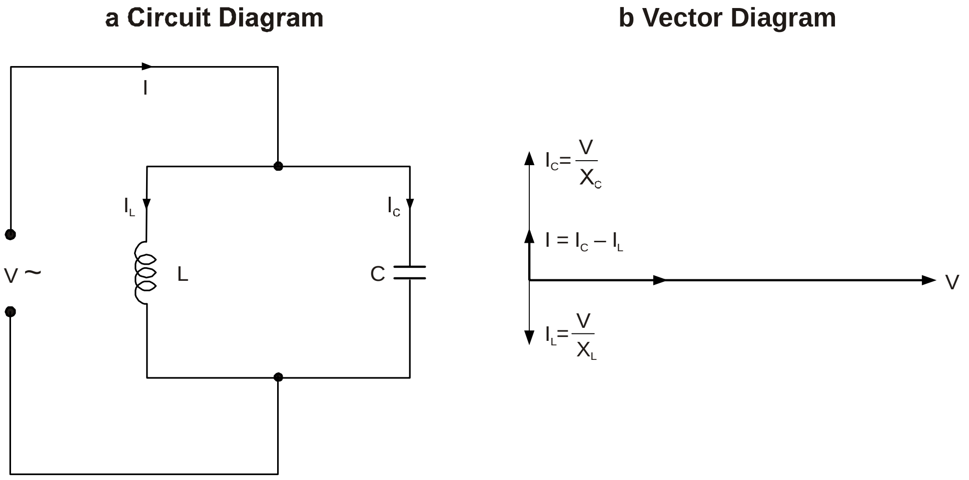
b. The current through the inductance is $I_L = V/X_L$ and lags V by 90º.

c. The current through the capacitance is $I_C = V/X_C$, and leads V by 90º.

d. The resultant supply current I is the vector sum of I_L and I_C and as I_L and I_C are 180º out of phase, I is therefore $I_C \sim I_L$.

e. I is 90º out of phase with V, leading or lagging V depending on the relative sizes of IL and IC. In Fig 23b I is shown leading E by 90º because IC > IL.

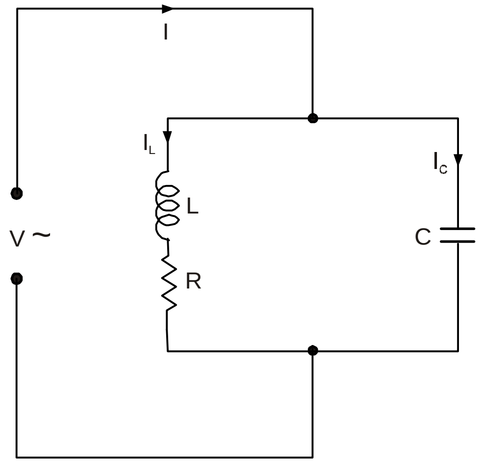
14-8 Fig 23 L and C in Parallel



Tuned Circuit with Resistance

43. Consider the circuit of Fig 24. This shows a practical arrangement of a coil and a capacitor in a parallel tuned circuit. The coil has certain power losses and these are represented by a resistance in series with the inductance of the coil. The capacitor losses are small in comparison and are ignored. This approximation is satisfactory for most purposes and is normally assumed when considering low power RF circuits.

14-8 Fig 24 Practical Parallel Tuned Circuit



44. The vector diagram for the circuit is constructed as shown in Fig 25, where:

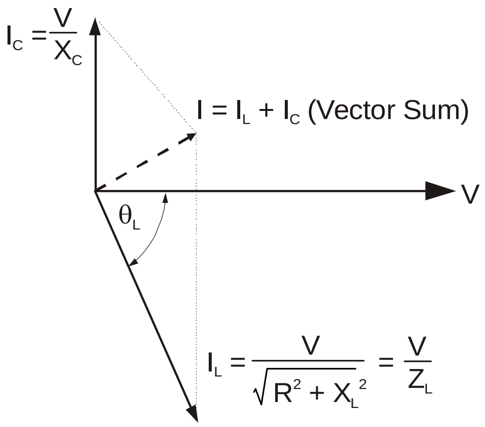
a. The applied voltage V is the reference vector.

b. The capacitive current is $I_C = V/X_C$ amperes and leads V by 90º. I_C increases with frequency.

c. The inductive current is $I_L = V/\sqrt{R^2 + X_L^2}$, and lags V by an angle θ_L which is less than 90º by virtue of the series resistance. I_L decreases as the frequency increases.

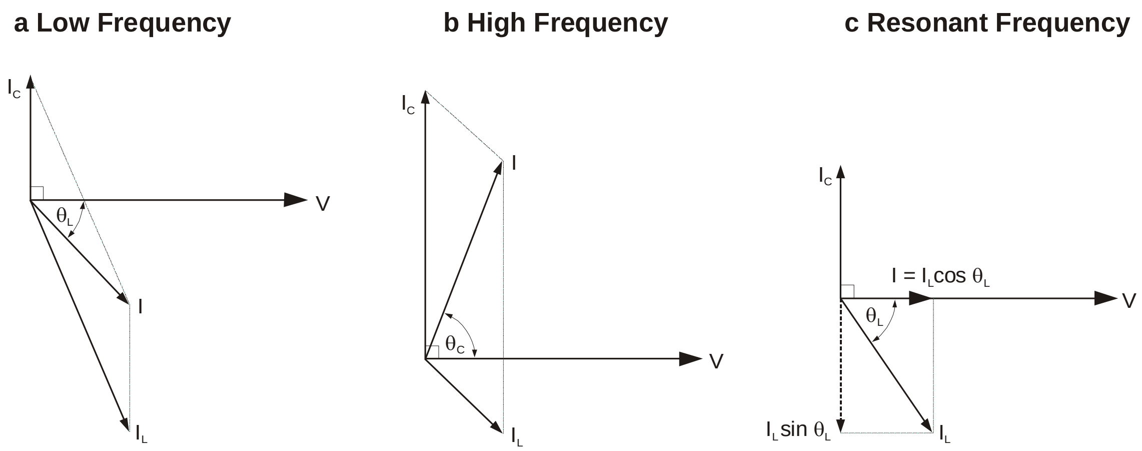
d. The resultant supply current I is the vector sum of I_C and I_L.

14-8 Fig 25 Vector Relationships in a Parallel Tuned Circuit


$I_C = \frac{V}{X_C}$
$I = I_L + I_C$ (Vector Sum)
V
θ_L
$I_L = \frac{V}{\sqrt{R^2 + X_L^2}} = \frac{V}{Z_L}$

45. The phase of the supply current I relative to the applied voltage V depends on the supply frequency, as well as on the values of L and C. At a low frequency, $I_C = V/X_C = V\omega C$ will be small and $I_L = \dfrac{V}{\sqrt{R^2 + \omega^2 L^2}}$ will be large so that I lags V by an angle θ_L (Fig 26a). At a high frequency, I_C will be large and I_L small, so that I leads V by an angle θ_C (Fig 26b). At a particular frequency (the resonant frequency) I will be in phase with V (Fig 26c).

14-8 Fig 26 Effect of Variation of Frequency



a Low Frequency

b High Frequency

c Resonant Frequency

46. Resonance occurs in a parallel tuned circuit when the supply current I is in phase with the applied voltage V. For this to be so, the out-of-phase (or reactive) component of I_L must equal I_C. There is then no reactance offered by the circuit and its impedance is a pure resistance. Thus the condition for resonance in a parallel tuned circuit is that of zero reactance.

47. **Resonant Frequency**. The out-of-phase component of a current I_L, which is at an angle I_C to the reference, is $I_L \sin\theta_L$ as shown in Fig 26c, and for resonance to occur $I_L \sin\theta_L = I_C$. It can be shown that at resonance, the resonant frequency is given by: $f_o = \dfrac{1}{2\pi}\sqrt{\dfrac{1}{LC} - \dfrac{R^2}{L^2}}$. Unlike in the series tuned circuit case, R has an effect on the resonant frequency of a parallel tuned circuit. In most cases, the resistance is very small and R^2/L^2 can be ignored, so that for practical purposes the resonant frequency can be taken as $f_o = \dfrac{1}{2\pi\sqrt{LC}}$ (Hz).

48. **Dynamic Impedance**. At resonance in a parallel tuned circuit there is zero reactance and the supply current I is in phase with the applied voltage V. From Fig 26c it is seen that at resonance, I is equal to the in-phase (or resistive) component of I_L. Thus at resonance $I = I_L \cos\theta_L$ and it can be shown that Z = L/CR. Note that the smaller R is, the greater is Z; if R is zero Z is infinite. The expression $Z_D = L/CR$ is known as the dynamic impedance and is the purely resistive impedance of a parallel tuned circuit at resonance.

49. **Frequency at or Near Resonance**. The frequency at which resonance occurs is given by:

$$f_o = \frac{1}{2\pi}\sqrt{\frac{1}{LC} - \frac{R^2}{L^2}}$$

If the applied frequency is greater than f_0 the current ($I_C = V\omega C$) through the capacitor will increase whilst that through the inductor $\left(I_L = V/\sqrt{R^2 + \omega^2 L^2}\right)$ will decrease. If the frequency is lower than f_0, the reverse occurs.

50. Supply Current at or Near Resonance. At resonance the supply current is in phase with the supply voltage and is at a minimum equalling VCR/L. When the frequency is above f_0, the supply current leads the supply voltage and the circuit is capacitive (see Fig 26b). When the frequency is below f_0, the circuit is inductive (see Fig 26a).

51. Impedance at or Near Resonance. At resonance the impedance equals L/CR and is at maximum. Either side of f_0 the impedance falls by an amount dependent on:

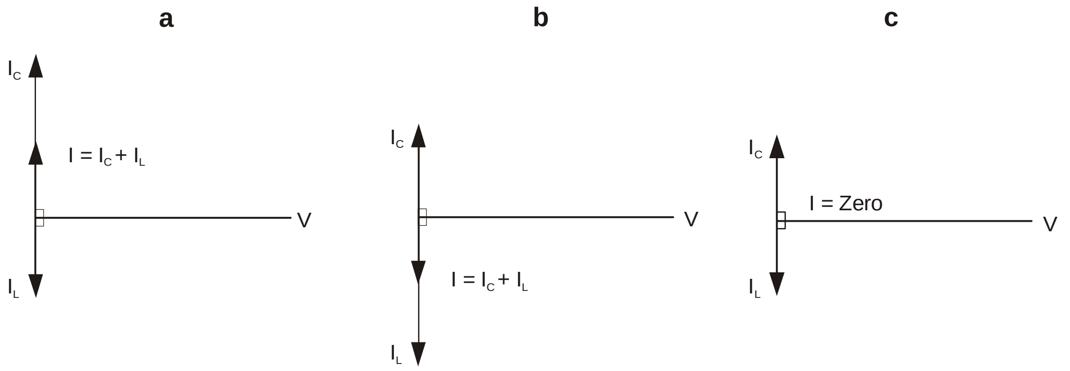
a. The departure from resonance.

b. The ratio of C to L.

c. The value of R.

52. **Resistance Ignored**. If the coil losses are so small that R can be ignored, the parallel tuned circuit has the 'ideal' form of Fig 23. The vector diagrams for this circuit will be as shown in Fig 27. If the applied frequency, or the values of L and C, are such that X_C is greater than X_L, I leads V by 90º and the circuit behaves as a pure capacitance (Fig 27a). If X_L is greater than X_C, I lags V by 90º, and the circuit behaves as a pure inductance (Fig 27b). If X_L equals X_C, the supply current into the circuit is zero and the circuit behaves as an infinite resistance (ie impedance is infinite). This is the resonant condition (Fig 27c), which occurs at a frequency $f_o = \frac{1}{2\pi\sqrt{LC}}$.

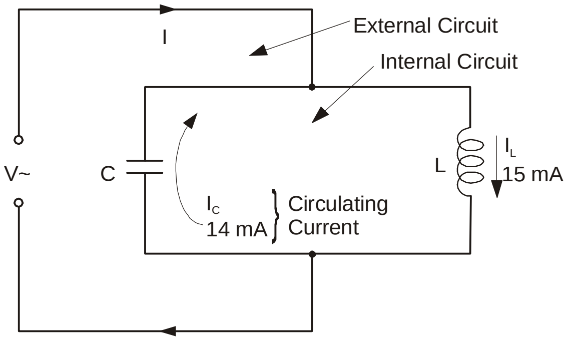
14-8 Fig 27 Effect of Varying X_C and X_L



Current Magnification

53. Although at resonance the current I into a parallel tuned circuit from the supply has a low value, this does not apply to the current within the closed LC loop. Thus the parallel tuned circuit may be thought of as having both an internal and an external circuit (Fig 28).

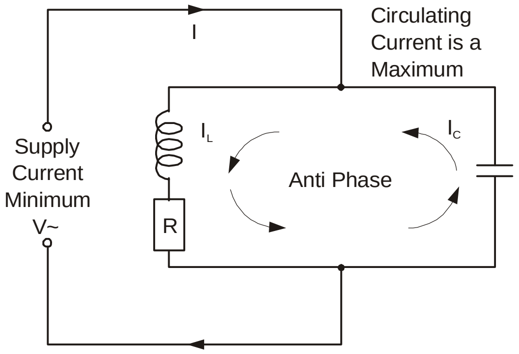
14-8 Fig 28 Supply and Circulating Current



54. In the ideal LC parallel circuit without resistance, the supply current I is the difference between I_L and I_C, since these currents are 180º out of phase. Thus I_L and I_C act in the same direction round the internal circuit, forming a circulating current. For example, if the ideal LC parallel circuit is near resonance, I_L may have a value of 15 mA and I_C a value of 14 mA. The circulating current is equal to the smaller of the two currents, and the supply current of 1 mA makes up the difference between I_L and I_C.

55. Fig 29 shows a practical parallel tuned circuit with resistance. At resonance, I has its minimum value of V/Z_D but the circulating current within the tuned circuit has a high maximum value equal to Q_0 times the supply current. It can be shown that at resonance $I_C = I_L = V\sqrt{\frac{C}{L}} = Q_0 I$. Thus a parallel tuned circuit is a current magnifier, as distinct from the series tuned circuit which is a voltage magnifier.

14-8 Fig 29 Current Magnification



56. **Selectivity**. The selectivity of a parallel tuned circuit is discussed at greater length in Volume 14, Chapter 9. It may be noted here that:

a. The higher the impedance at resonance in relation to the impedance at frequencies off resonance, the greater the selectivity of the circuit.

b. A circuit with a high value of Q has high selectivity.

CHAPTER 9 - SELECTIVITY OF TUNED CIRCUITS

Introduction

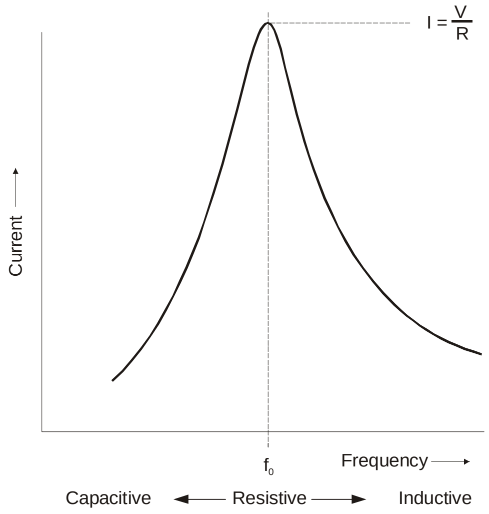
1. Selectivity of a tuned circuit is defined as the circuit's ability to pick out (select) a desired frequency, or band of frequencies, and reject the unwanted ones. The sharpness of response over a range of frequencies near resonance gives an indication of the selectivity of a circuit. In this Annex, considerations affecting the selectivity of series and parallel tuned circuits are examined.

SERIES TUNED CIRCUITS

General

2. For the series tuned circuit, selectivity is a measure of the ease with which the circuit can accept an input at the resonant frequency as compared with inputs off resonance. Fig 19 of Volume 14, Chapter 8 shows that the impedance Z of a series tuned circuit falls as resonance is approached. At the resonant frequency, such a circuit allows the maximum current to flow through it, and is often termed an acceptor circuit. A graph showing the variation of current with frequency is at Fig 1.

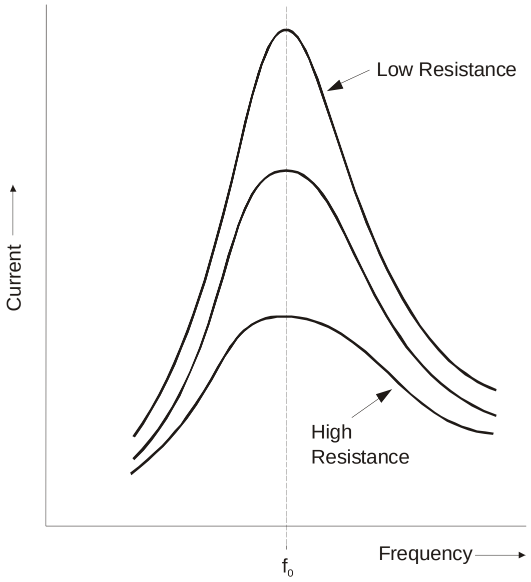
14-9 Fig 1 Variation of Current with Frequency - Series Tuned Circuit



The Effect of Resistance

3. At resonance I = V/R, thus, if R is doubled, the current at resonance is halved. In a properly designed acceptor circuit, above resonance X_L is much greater than R, and below resonance, X_C is greater than R. Thus, the effect of R is progressively less important as the frequency is moved away from resonance. The effect of resistance is to reduce the current at frequencies near resonance to a far greater extent than at other frequencies, i.e. to 'flatten' the response curve as shown in Fig 2. Selectivity is therefore reduced by an increase in resistance.

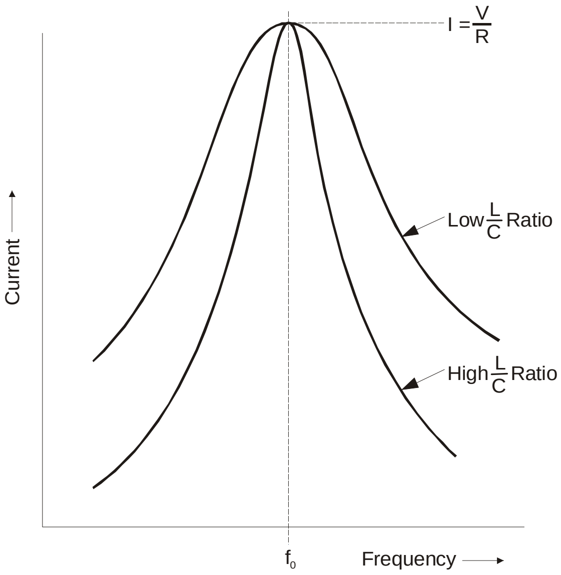
14-9 Fig 2 Effect of Resistance on Selectivity



The Effect of the L/C Ratio

4. The resonant frequency of an acceptor circuit is given by $f_0 = \frac{1}{2\pi\sqrt{LC}}$. It is dependent on the product LC*,* thus, if the inductance of a circuit is doubled and the capacitance halved the resonant frequency remains unaltered. However, the ratio of L to C has been increased four times and this has an effect on the selectivity of the circuit. Fig 3 shows the response curves for two circuits having the same resonant frequency and equal values of resistance but different L/C ratios. From this it is seen that the greater the L/C ratio the more selective is the circuit.

14-9 Fig 3 Effect of L/C Ratio on Selectivity



Q Factor

5. The quantity used to represent the selectivity of a circuit at resonance is denoted by Q_0 which is defined as the ratio of reactance to circuit resistance at the resonant frequency. It is usually considered for the coil, ie $Q_0 = \frac{\omega_0 L}{R}$

Thus, $Q_0 = \frac{2\pi\pi_o L}{R}$, but $f_0 = \frac{1}{2\pi\sqrt{LC}}$

$$\therefore \quad Q_0 = \frac{2\pi L}{R} \times \frac{1}{2\pi\sqrt{LC}} = \frac{L}{R} \times \frac{1}{\sqrt{LC}}$$

$$\therefore \quad Q_0 = \frac{1}{R}\sqrt{\frac{L}{C}}$$

This expression supports paras 3 and 4 as it shows that the Q_0 or selectivity of a tuned circuit is inversely proportional to R, and proportional to the ratio of L to C. Thus, a circuit with a high value of Q_0 has high selectivity.

Q_0 and Bandwidth

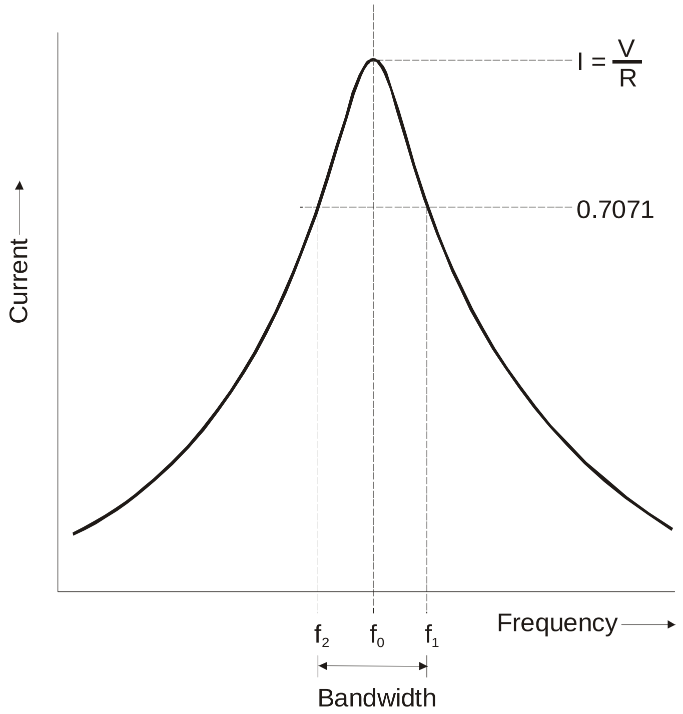
6. Series tuned circuits are used in radio to accept inputs at the resonant frequency and in the immediate neighbourhood of resonance. In order to present a high impedance to inputs considerably removed from resonance, a circuit must have a high Q_0 value. In practice, values of Q_0 vary from about 10 at audio frequencies to several hundreds at radio frequencies. The graph of current against frequency for a circuit having a high Q_0 shows a sharp response curve (Fig 4). An alternative way of describing this is to say that the circuit has a narrow bandwidth. Bandwidth is defined as the separation between two frequencies either side of the resonant frequency at which the power has fallen to 50% of the maximum power. This is known as the half-power bandwidth. Since power is proportional to the square of the current, reducing the power to one-half means reducing the current by a factor of $\sqrt{\frac{1}{2}}$ = 0.707. Thus in terms of current the bandwidth of a series tuned circuit is the difference between two frequencies f_1 and f_2 at which the current is 70% of the current at the resonant frequency f_0 (Fig 4). The selectivity of a series tuned circuit can therefore be defined either in terms of Q_0 or in bandwidth. For purposes of calculation, Q_0 and the bandwidth as defined above are related by the expressions:

$$Q_o = \frac{\text{resonant frequency}}{\text{bandwidth}} = \frac{f_o}{f_1 - f_2}$$

Thus if the resonant frequency of a circuit is 200 kHz and the bandwidth required is 12 kHz,

$$Q_0 = \frac{200}{12} = 16.7$$

14-9 Fig 4 The Half-power Bandwidth



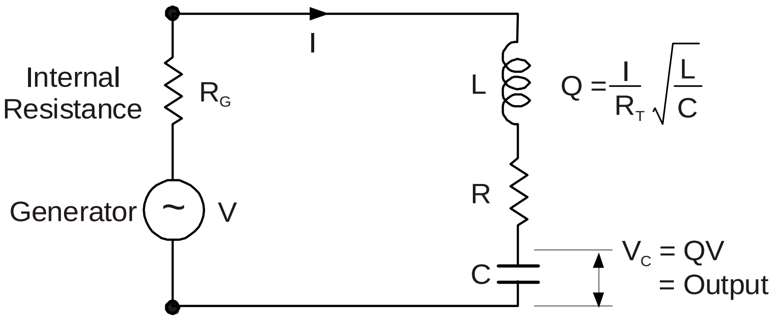
Effect of Supply Impedance

7. When a series resonant circuit is used for tuning purposes, it is often intended that the voltage across the capacitor be applied to another stage. The voltage applied to the circuit itself can be considered as the supply from a generator with an internal resistance R_G (Fig 5). If R_G is considered non-reactive, it will modify the tuned circuit characteristics as follows: total circuit resistance is $R_T = R + R_G$, ie:

$$\text{effective } Q = \frac{1}{R + R_G}\sqrt{\frac{L}{C}}$$

If R_G is large compared with R, then Q is reduced and the voltage across the capacitor at the resonant frequency ($V_C = QV$) and the selectivity of the circuit are both relatively small. Thus, the series circuit is more selective when fed by a generator of low internal impedance.

14-9 Fig 5 Effect of Supply Impedance on Selectivity



The Q Factor of Components

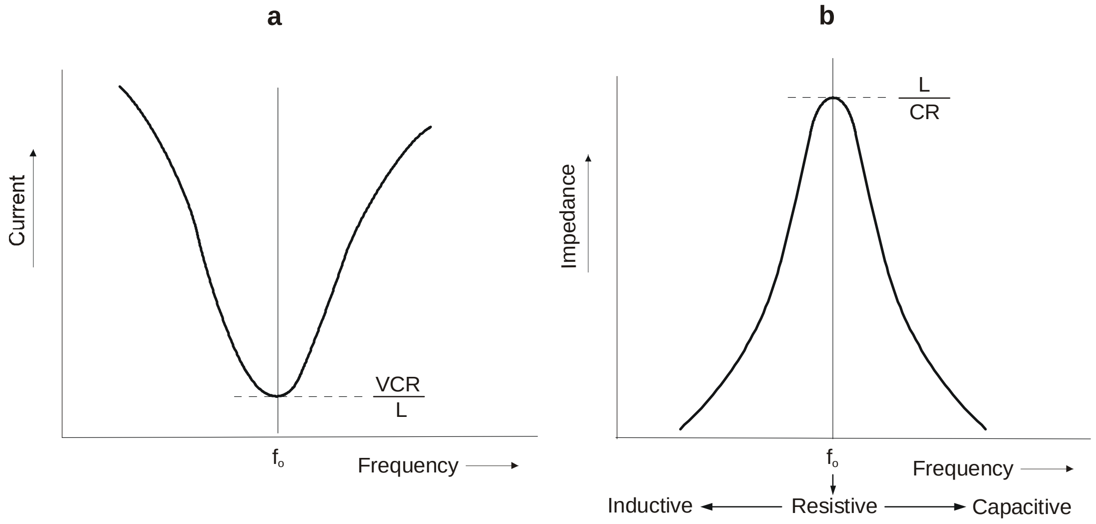
8. It was shown in Volume 14, Chapter 8 that the power losses associated with inductors and capacitors can be expressed in terms of the power factor, cos θ = R/Z = $\frac{R}{\sqrt{R^2 + X^2}}$ where R is the equivalent loss resistance of the component. In well-designed components, R is small compared with X, and the power factor approximates to R/X (i.e., R/ωL for inductors and ωCR for capacitors). It is, however, more convenient to express the quality of inductors and capacitors in terms of the reciprocal of the power factor, namely ωL/R = 1/ωCR = Q. Thus, a coil having a high value of Q indicates a component with low losses. Q remains relatively constant with frequency since the value of loss resistance (R) varies with frequency in much the same way as the reactance (X). At audio frequencies, Q values for coils rarely exceed 10, whilst coils normally used for radio frequencies have Q values around 50 to 300. Q values from 100 to 300 are common with paper dielectric capacitors and from 1,000 to 3,000 for mica capacitors.

PARALLEL TUNED CIRCUITS

General

9. The variation of supply current with frequency for a parallel tuned circuit is shown in Fig 6a. The sharpness of the response curve denotes the selectivity of the circuit. For parallel circuits, selectivity is often defined in terms of impedance rather than of current, and the higher the impedance at resonance in relation to the impedance at frequencies off resonance, the greater the selectivity of the circuit. Fig 6b shows the variation of impedance with frequency.

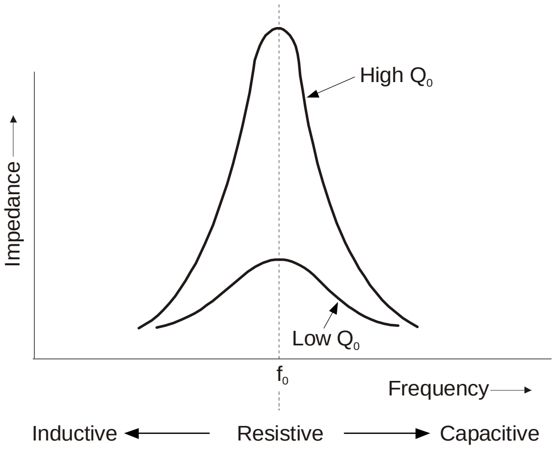
14-9 Fig 6 Variation of I and Z with Frequency about Resonance - Parallel Tuned Circuit



Q_0 and Selectivity

10. An increase in resistance R will reduce the impedance at resonance (Z = L/CR) to a greater extent than at frequencies removed from resonance. Since $Q_o = \frac{1}{R}\sqrt{\frac{L}{C}}$ is reduced, the response curve is 'flattened' and the circuit is made less selective. A variation in the ratio of L to C will also affect the impedance at resonance and, hence, the selectivity of the circuit. Thus, a circuit with a high value of Q_0 has a high selectivity, and vice versa (Fig 7).

14-9 Fig 7 Q_0 and Selectivity

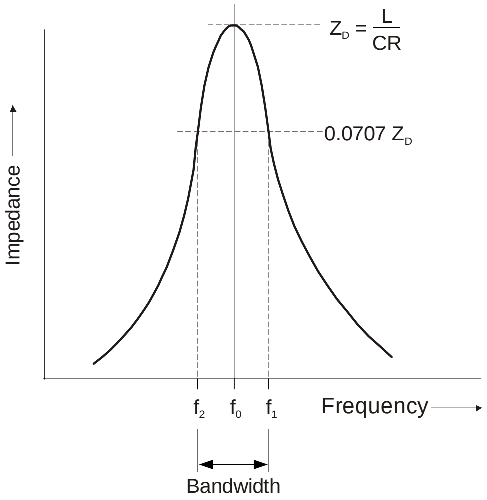


Q_0 and Bandwidth

11. Parallel tuned circuits are used in radio equipment to 'reject' inputs at, and near, the resonant frequency, by offering a high impedance at those frequencies. To inputs at frequencies considerably removed from the resonant frequency, the parallel tuned circuit should offer a low impedance. To do this satisfactorily, the circuit must have a high Q_0, ie a narrow bandwidth. The bandwidth of a parallel tuned circuit is defined as the difference between two frequencies f_1 and f_2 at which the impedance has fallen to 70% of the resonant value (Fig 8). For purposes of calculations, Q_0 and the bandwidth are related by the expression:

$$Q_o = \frac{\text{resonant frequency}}{\text{bandwidth}} = \frac{f_o}{f_1 - f_2}$$

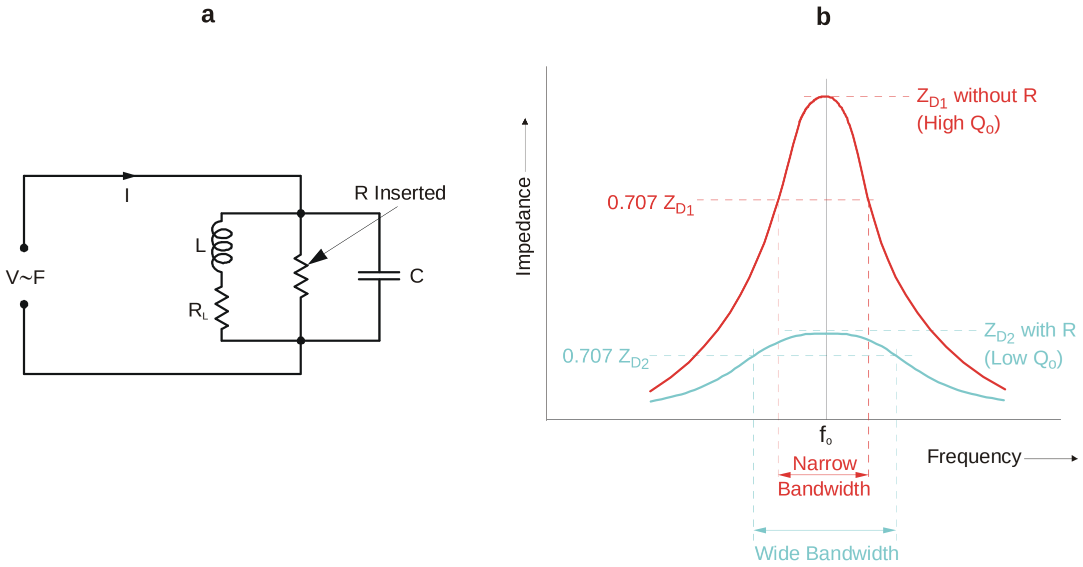
14-9 Fig 8 The Half - power Bandwidth



Damping of Parallel Circuits

12. In some circuits in radio equipments, it is desired to 'pass' a wide band of frequencies in the neighbourhood of resonance, and for this a circuit with a wide bandwidth is required. From para 11, it is seen that the bandwidth of a parallel tuned circuit is inversely proportional to Q_0 (ie bandwidth = f_0/Q_0). Thus, a circuit with a low value of Q_0 will give a wide bandwidth. To reduce Q_0 an actual resistor is inserted in parallel with the rejecter circuit to 'damp' the response (Fig 9a). The impedance at resonance is then the dynamic impedance Z_D(= L/CR_L) in parallel with R. Consequently, the reduction of Q results in reduced impedance at resonance and a flattened response curve with a corresponding increase in bandwidth (Fig 9b).

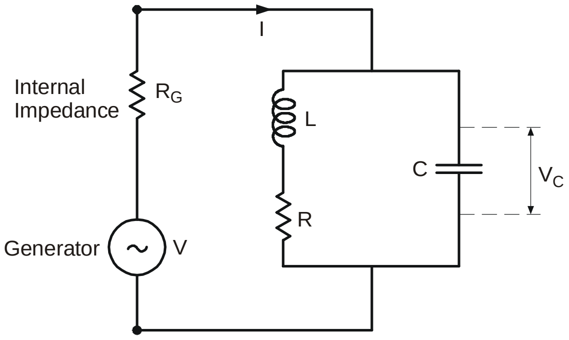
14-9 Fig 9 Effect of Damping on Selectivity



Effect of Supply Impedance

13. When a parallel circuit is used for tuning purposes, the idea is that the voltage across the capacitor shall be applied to a further stage. The voltage applied to the circuit itself can be considered to be applied from a generator, and the behaviour of the circuit then depends on the impedance of the generator as well as all the characteristics of the circuit itself. If the internal impedance of the generator (R_G) is small compared with the impedance of the parallel tuned circuit, then V_C will be approximately equal to V at all frequencies (Fig 10). If, however, R_G is large compared with the impedance of the circuit, then the current will be approximately equal to V/R_G. Thus, V_C will be proportional to the impedance of the parallel circuit and will vary with frequency, being a maximum at resonance. Hence, the parallel circuit is selective only if supplied from a generator having a high internal impedance.

14-9 Fig 10 Effect of Supply Impedance on Selectivity



CHAPTER 10 - AC GENERATORS AND MOTORS

POLYPHASE AC - GENERAL

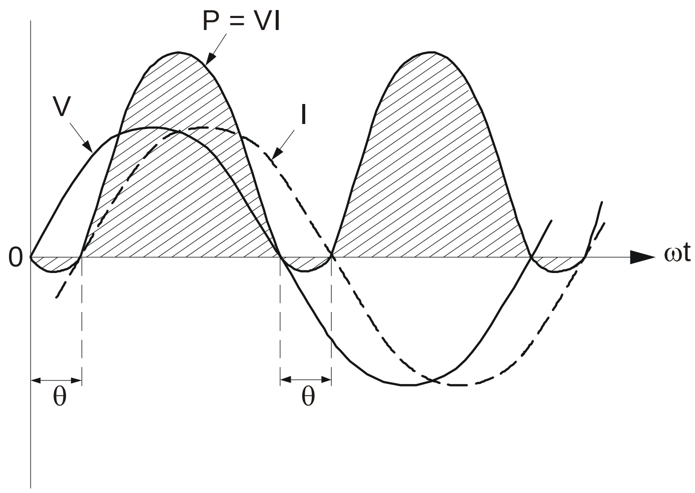
Introduction

1. The simple AC generator (sometimes called an alternator), discussed in Volume 14, Chapter 3, produces a single-phase AC output at its slip rings. In other AC generators, the armature winding consists of two or more groups of series-connected coils, with their outer ends connected to separate slip rings. Two or more alternating voltages are then produced at the slip rings, these voltages being, in general, out of phase with each other. Such machines are known as polyphase or multiphase AC generators, or alternators.

Effect of Load on Single-phase and Polyphase Supply

2. In polyphase systems, loads can be arranged to draw power from the generator at a uniform rate, so that the machine runs steadily under a uniform torque. Fig 1 shows the relationship between current and voltage when a single-phase AC generator is supplying an inductive load. The instantaneous power is the product VI at each instant and is indicated by the shaded portion of the graph. When the power graph is above the time axis, it indicates that energy is being drawn from the generator; below this axis, energy is being returned from the inductive load. Since the load on the generator not only fluctuates during each half cycle, but also changes sign, the torque changes in a similar manner. Though this may happen in the individual cycles of each phase in a multiphase system, the power peak and zero values of various phases will not normally coincide because of the phase difference. Power is therefore drawn from the generator at a much more uniform rate, and the torque remains steadier.

14-10 Fig 1 Power Variation with Inductive Load in a Single-phase Supply



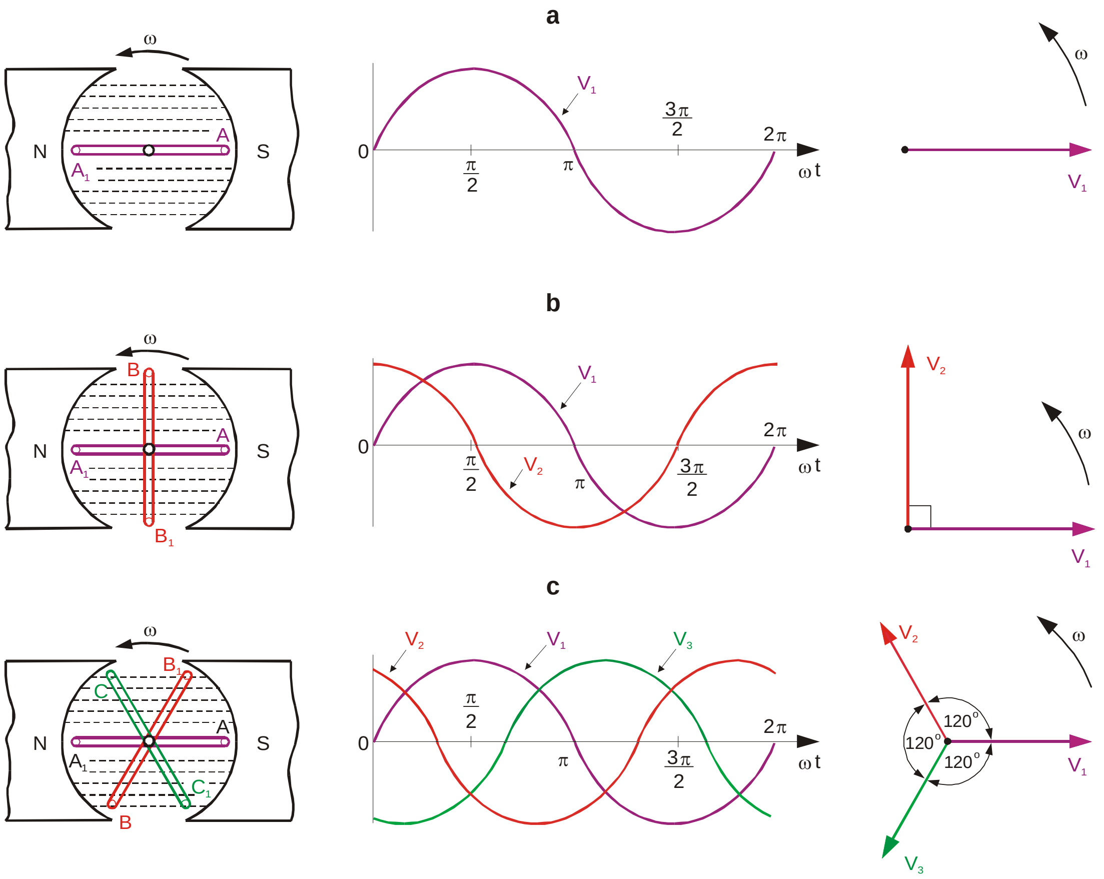
Generation of Polyphase Voltages

3. The output of a simple AC generator, consisting of a single group of coils rotating uniformly in a uniform magnetic field, is a sine wave (V_1), as shown in Fig 2a, which represents the voltage in the coils (AA_1) available at the slip rings.

4. If, on the same armature core, a second group of coils (BB_1) is mounted at right angles to the first and connected to a second pair of slip rings, two independent voltages are available, differing in phase by 90º. This arrangement, shown in Fig 2b, represents a two-phase generator.

5. With three groups of coils (AA_1, BB_1 and CC_1) wound independently on the same armature, and connected to three separate pairs of slip rings, three independent voltages are generated, as indicated in Fig 2c. The start of one coil is at an angle of 120º to the start of the next (i.e. A is 120º removed from B) and the three voltages differ in phase by 120º, the arrangement representing a three-phase generator. It should be noted that, when the voltage is in one phase at peak value, the voltage in each of the other phases is at half peak value in the opposite direction. The sum of the voltages in the three phases is thus zero; this holds good for all points in the cycle.

14-10 Fig 2 Generation of Polyphase Voltages



6. The armature can be wound with any number of symmetrically spaced groups of coils and independent pairs of slip rings; six-phase and twelve-phase AC supplies are sometimes used. However, since many of the advantages of polyphase voltages are available with three phases, the three-phase system is that in most general use for the generation and transmission of power.

Advantages of Polyphase Systems

7. The advantages of polyphase systems can be summarized as follows:

a. Other parameters being unchanged, the power rating of a polyphase machine increases with the number of phases.

b. Both the heating loss for a given power transmitted, and the line voltage drop, are less than they would be if the whole power was transmitted by a single phase only.

c. Loads can be arranged to draw power from the generator at a uniform rate (see para 2).

d. Polyphase alternators can be made to work in parallel without much difficulty.

e. Polyphase motors may be self-starting and provide uniform torque.

Symmetrical and Balanced Systems

8. A polyphase system is said to be:

a. Symmetrical, when the phase voltages have the same amplitude and are displaced from one another by equal angles.

b. Balanced, when voltage, current and phase angle are the same for each phase. In a balanced three-phase system, the sum of the instantaneous values of voltage (or current) is zero.

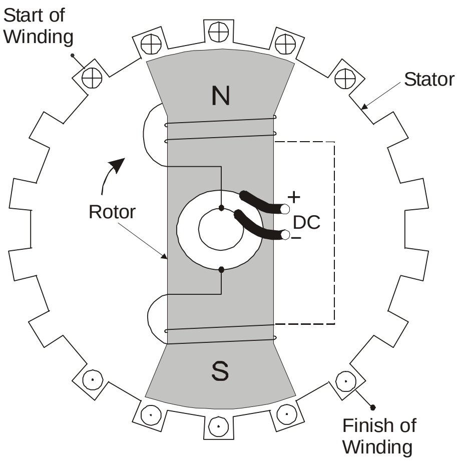
AC GENERATORS

Single-phase Generators

9. A simple form of AC generator (the rotating armature type), consisting of a coil caused to rotate in a magnetic field by a prime mover, is described in Volume 14, Chapter 3. Most actual machines are of the rotating field type, designed the other way around. That is, the rotating part (rotor) consists of electromagnets energized from a DC supply, while the coils in which the generated emf is induced are wound on a fixed frame (stator). The two arrangements are electrically equivalent, since the relative motion of flux and coils that gives rise to an induced emf is the same in both cases. However, fixed connections to stationary windings are easier to insulate at high voltages than slip rings would be, and whether the output is single-phase or polyphase, only one pair of slip rings is needed for the relatively low voltage, DC energizing supply.

10. Fig 3 shows the arrangement of a two pole, single-phase AC generator of the rotating field type. The stator winding consists of a number of coils connected in series and inserted in slots cut in the inner surface of the laminated frame. The rotor is driven by a prime mover and carries the field windings, which are energized from a DC source via slip rings, as shown.

14-10 Fig 3 Two-pole Single-phase AC Generator



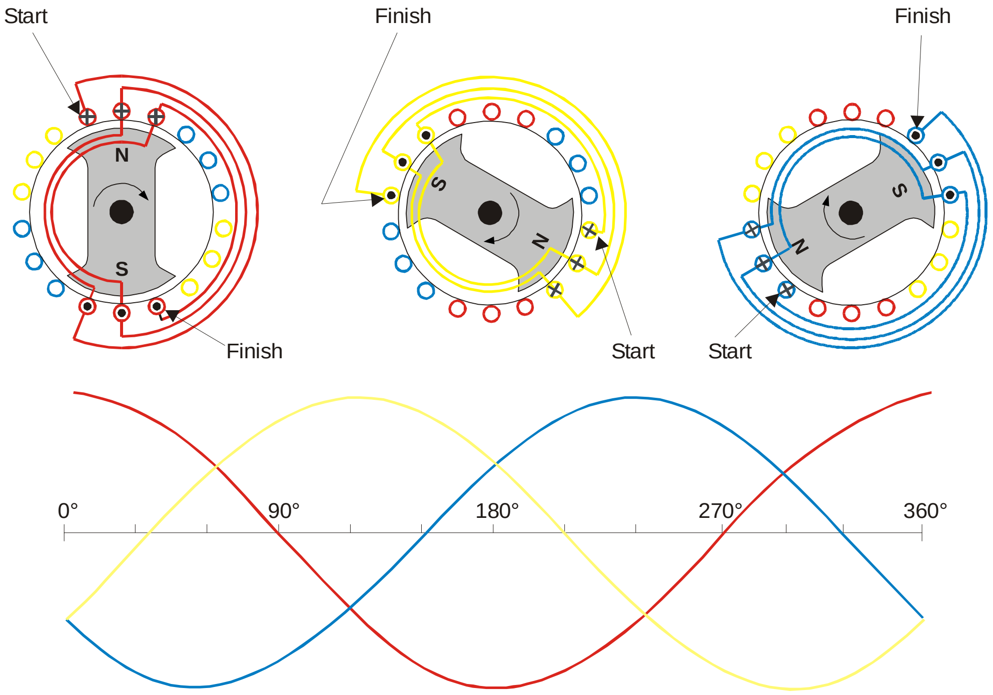
Two-phase Generators

11. By mounting two separate coils on the rotor at right angles to each other, a two-phase output is produced. The voltages induced in each coil will be of the same magnitude and frequency, but there will be a 90º phase difference between the voltages. Two-phase supplies have limited application, mostly confined to direction control, where phase relationship determines the direction of rotation.

Three-phase Generators

12. The stator of a three-phase AC generator has three separate windings equally spaced round the interior of the frame, and, between them, occupying all the slots in the laminations. The poles of the rotor sweep past each winding in turn, so that three alternating voltages are produced, differing in phase by 120º. This is shown in Fig 4. Standard practice identifies the phases by the colours red, yellow (or white), and blue; each phase then being referred to by its colour.

14-10 Fig 4 Two-pole Three-phase AC Generator



13. The emf generated in each conductor completes one cycle as it is passed by a pair of rotor poles. The machine shown in Fig 4 has two poles (one pair) on the rotor. However, some machines have many pairs of poles on the rotor, with the stator winding spaced accordingly, in a manner similar to that for the multi-pole DC generator discussed in Volume 14, Chapter 6. Thus, in one revolution of the rotor, the emf will complete a number of cycles corresponding to the number of pairs of poles in the generator. The frequency of the supply given by the machine will be:

$$\text{Frequency (F)} = \frac{p\eta_s}{60} \text{ Hz,}$$

where p = number of pairs of poles and ηS = speed in rpm.

The factor η_s is called the synchronous speed; it is the speed at which the machine must run in order to generate the required frequency. Thus:

$$\eta_s = \frac{f}{p} \times 60 \,(\text{rpm})$$

INTERCONNECTION OF PHASES

General

14. Each phase of a three-phase generator may be brought out to separate terminals and used to supply separate loads independently of one another. This method requires a pair of lines for each phase. The number of wires may be reduced, with a consequent saving in cable, if the phases are interconnected. There are two main methods of connecting the generator windings and the loads in three-phase systems:

a. The Star, or Y, Connection. A development of this is the four-wire star connection.

b. The Delta, or Mesh, Connection.

Star (Y) Connection

15. A star, or Y, connection is achieved by joining similar ends (the starting or the finishing ends) of the three windings to a common point (the neutral or star point). The other ends are joined to the line wires. In Fig 5, let E_a, E_b and E_c be the phase voltages, and I_a, I_b and I_c the phase currents, lagging behind their respective phase voltages by a constant angle θ (ie a balanced load). The arrows indicate the assumed positive directions of the phase voltages. Voltage between lines is equal to vector difference between the two-phase voltages. Hence, the voltage between lines 1 and 2 (${}_1V_2$) is obtained by the vector addition of vector E_a and vector E_b reversed. Thus:

$${}_1V_2 = E_a - E_b = 2E \cos 30^o$$

$$= 2E\,\frac{\sqrt{3}}{2}$$

$$= E\sqrt{3}$$

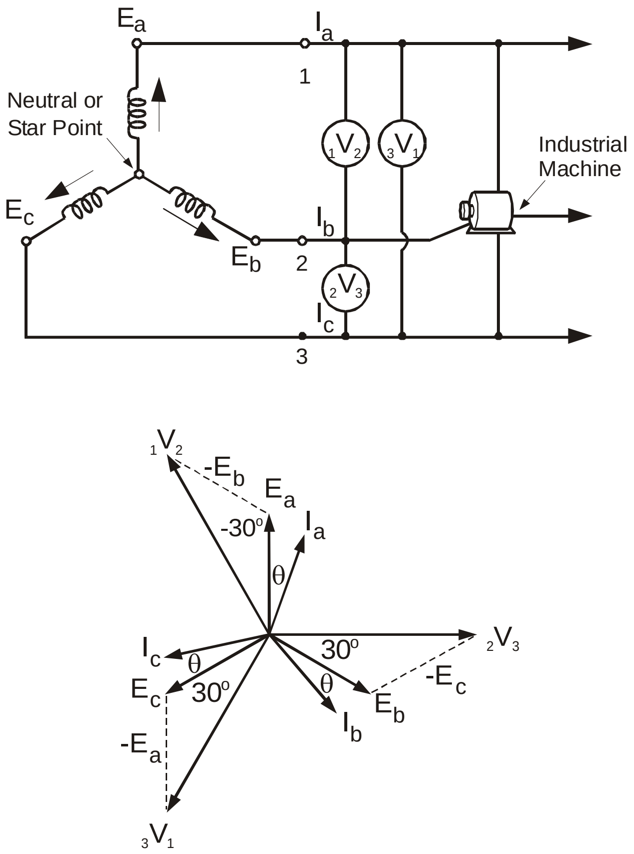
where the magnitudes of $E_a = E_b = E_c = E$, the phase voltage.

Similar results are obtained for ${}_2V_3$, and ${}_3V_1$, thus line voltages are each equal to $\sqrt{3}$ × phase voltage, or, generally:

$$V_L = E\sqrt{3}$$

Since each line is in series with its individual phase winding, the line current must equal the phase current, i.e. $I_L = I$.

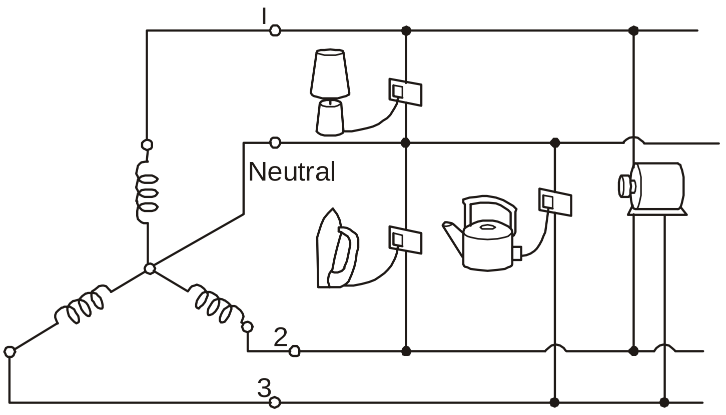
14-10 Fig 5 Three-phase Star Connection



Three-phase Four-wire Star Connection

16. In the three-phase four-wire star connection system, used in the National Grid, the phases of the alternator or transformer are connected in star, and three-line wires taken from the terminals. In addition, a fourth wire is taken from the star point (Fig 6). This enables lighting loads and domestic services to be taken at phase voltage E (240 V) by connecting between line and neutral, and at the same time allows industrial machines to be supplied at line voltage $\sqrt{3}$ E (415 V).

14-10 Fig 6 Three-phase Four-wire System



17. Three-phase motors constitute a balanced load, but domestic loads across the phases may not be balanced; in this case, the live conductors carry unequal currents, and the out of balance current is carried by the neutral.

Mesh or Delta (Δ) Connection

18. In the Mesh, or Delta, system the phase windings are connected to form a closed mesh (Fig 7). Assuming, as before, a balanced load:

voltage between lines = phase voltage, i.e. $V_L = E$.

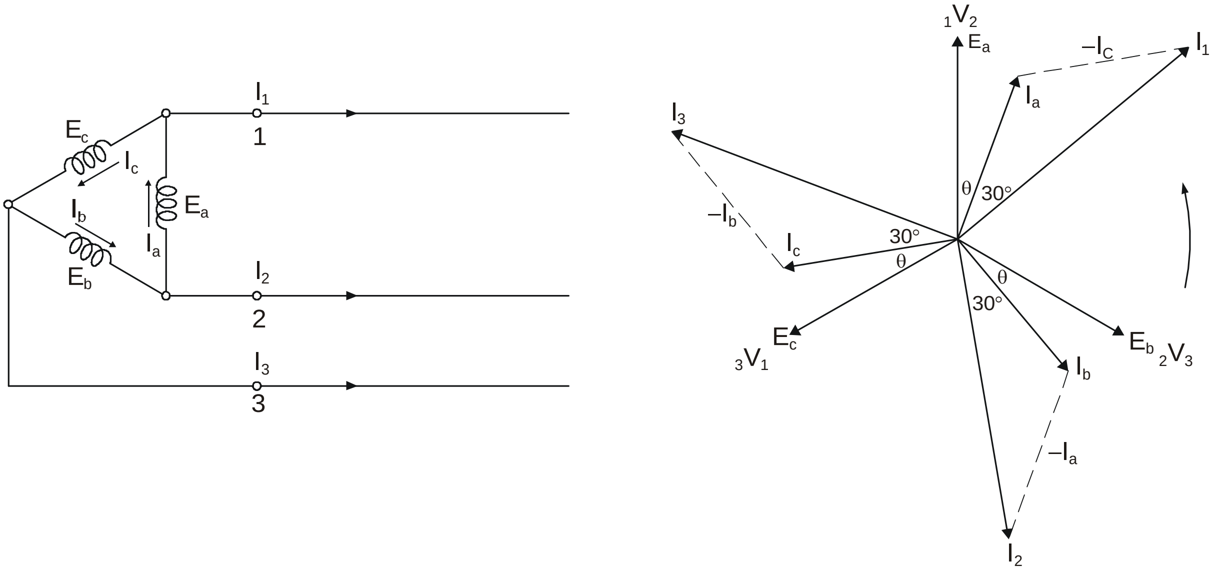
In this case, the line current is equal to the vector difference between two phase currents. Hence:

$I_1 = I_a - I_c = 2I \cos 30^o$

$= I\sqrt{3}$

where the magnitudes of $I_a = I_b = I_c = I$, the phase current. Similar results are obtained for I_2 and I_3, thus line currents are each equal to $\sqrt{3}$ × phase current, or generally, $I_L = I\sqrt{3}$

14-10 Fig 7 Three-phase Delta Connection



Power in Three-phase Circuits

19. In a balanced three-phase circuit, the total power equals three times the power per phase, ie

total power = 3 × phase voltage × phase current × power factor,

or, $P = 3VI \cos \theta$ watts,

where θ is the angle between phase voltage and phase current.

Expressing this in line quantities:

a. **Star Connection**.

$$P = 3\frac{V_L}{\sqrt{3}} I_L \cos\theta = \sqrt{3} V_L I_L \cos\theta$$

b. **Delta Connection**.

$P = 3V_L \frac{I_L}{\sqrt{3}} \cos\theta = \sqrt{3} V_L I_L \cos\theta$, as for star connection.

AC MOTORS

Introduction

20. There are three main types of AC motor - the synchronous motor, the induction motor, and the series or commutator motor. Of these, the induction motor is the most common in industry; the commutator motor is the most common in the home.

21. As most large AC motors do not have commutators (some do not even have slip rings), they are more trouble-free in operation than DC motors. Some are also useful as constant speed motors; their speed being determined by the AC supply frequency.

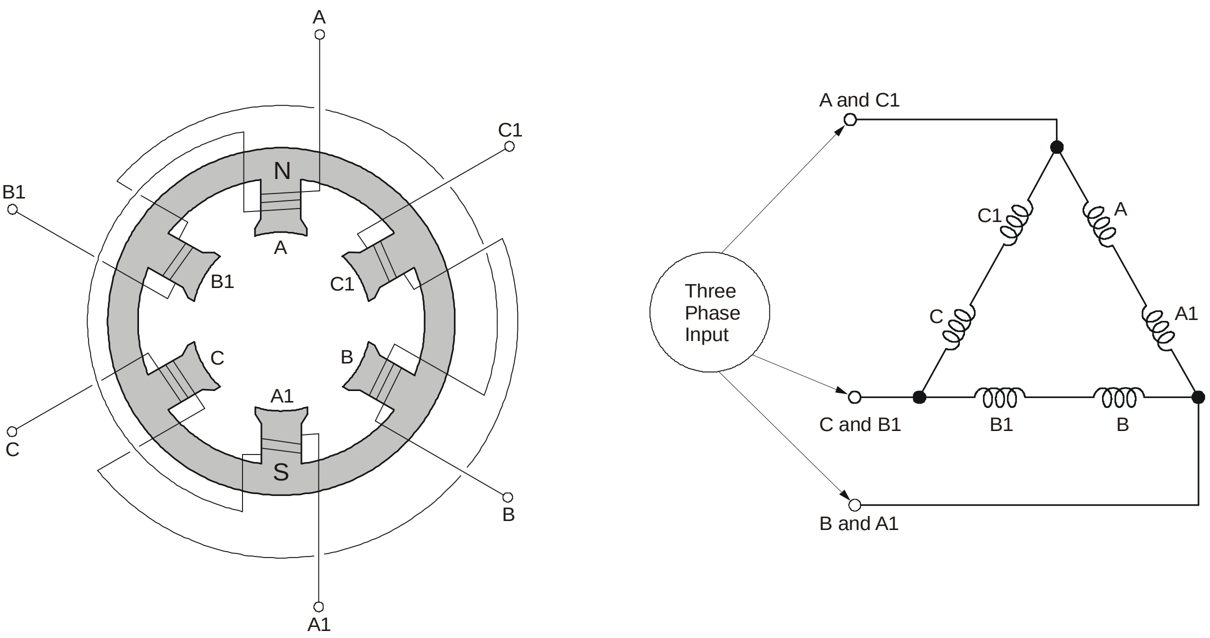
22. AC motors may be operated from single-phase or polyphase supplies; the basic principles of operation are similar in both cases. Synchronous and induction AC motors work on the principle that AC applied to the stator produces a rotating magnetic field, which causes the rotor of the machine to turn with the field.

Production of a Rotating Magnetic Field

23. A rotating magnetic field can be produced by applying a three-phase supply to a stationary group of coils, provided the latter are suitably wound, spaced, and connected. The field produced is of unvarying strength, and its speed of rotation is directly related to the frequency of the supply.

24. Fig 8a shows a typical three-phase stator. The two windings in each phase (eg A and A_1) are connected in series and are so wound that current flowing through the two windings produces a North pole at one of them and a South pole at the other. Thus, if current is flowing in the A phase in the direction from the A to the A_1 terminal, pole-piece A becomes a North pole and A_1 a South pole. In operation, the three-phase stator is connected in delta, as shown in Fig 8b, so that only three terminals, each common to two of the windings, are provided for the three-phase input.

14-10 Fig 8 Three-phase Stator and Connections



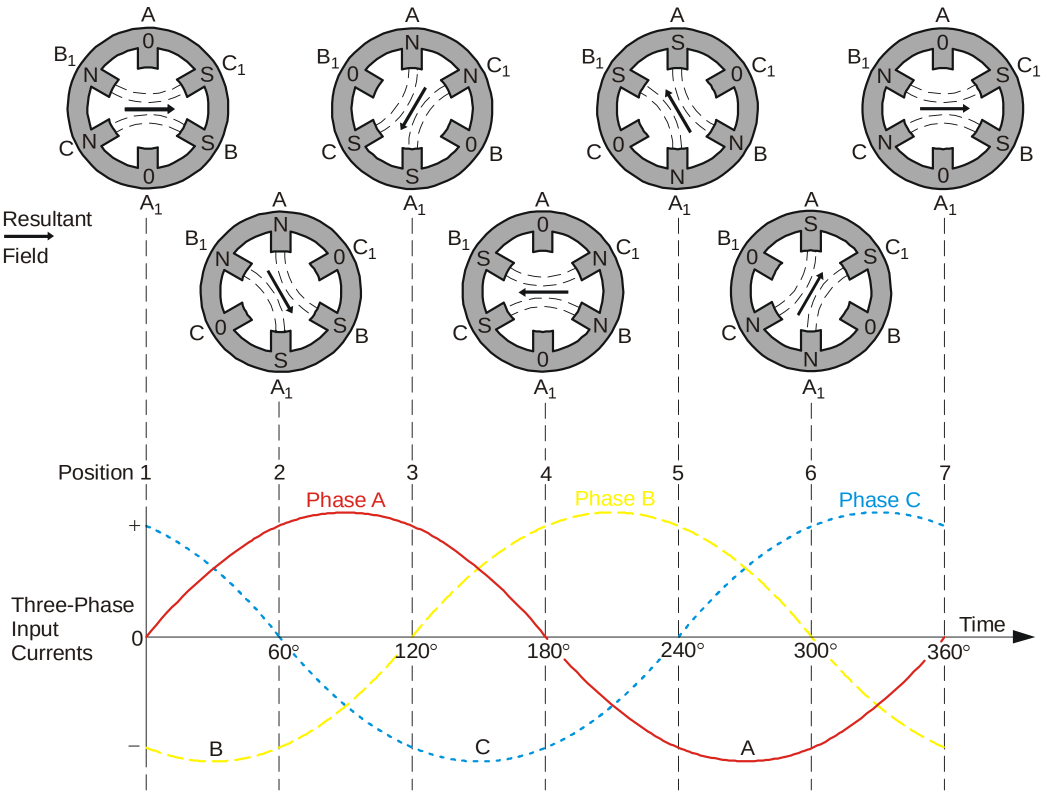
a Stator

b Delta Connection

25. At any instant, the magnetic field generated by one particular phase is proportional to the current in that phase; as the current alternates, so does the magnetic field. As the currents in the three phases are 120º out of phase, the three magnetic fields will also alternate 120º out of phase with each other, and the resultant magnetic field is the vector sum of these three.

26. Fig 9 shows how the magnetic fields add up to give a resultant magnetic field which continuously shifts in direction. After one complete cycle of AC input, the resultant magnetic field has shifted through 360º, or one revolution. Thus, although the coils are stationary, the application of three-phase AC produces a magnetic field that rotates at the frequency of the supply.

14-10 Fig 9 Production of a Rotating Magnetic Field



27. At position 1 in Fig 9, the current in phase A is zero, as shown by the graph; the current in phase C is positive and flows in the direction C to C_1, and that in phase B is negative and flows in the direction B_1 to B. Equal currents therefore flow in opposite directions through the B and C phase windings, and magnetic poles are established as shown. Since the magnetic fields of the B and C phases are equal in amplitude (equal currents), the resultant field lies in the direction shown by the arrow.

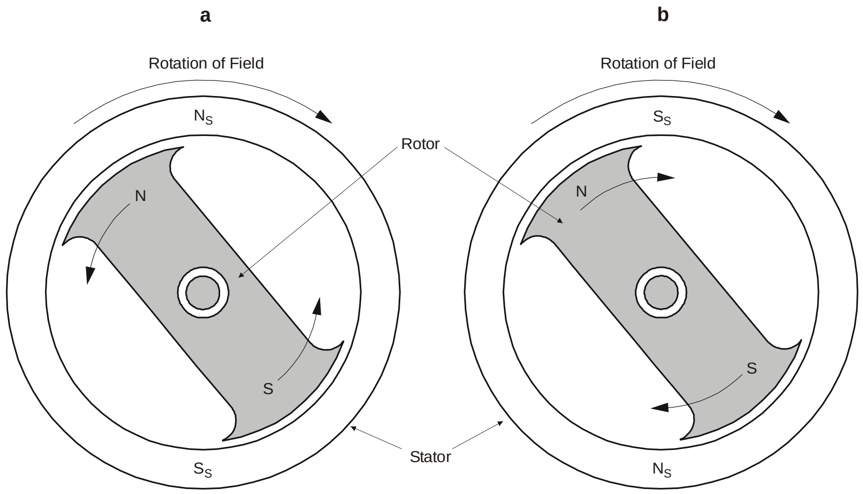
28. Position 2 shows the condition when the supply cycle has advanced 60º. The current C is now zero; A is positive and B negative. The resultant field is as shown. The other positions shown are at intervals of 60º. Thus, the field rotates one complete revolution during one complete cycle of the AC supply. An input frequency of 50 Hz produces a field rotating at 50 revolutions per second, i.e. at 3,000 rpm.

Synchronous Motors

29. The AC generator, like the DC generator, is a reversible machine; if supplied with electrical energy, it will run as a motor. An AC generator used in this manner operates as a synchronous motor (so called, because the speed of the machine is dependent upon the synchronous speed of the rotating field produced by the input to the stator windings).

30. When a three-phase stator winding is supplied with three-phase AC, a magnetic field, of constant magnitude and rotating at synchronous speed, is produced within the stator gap. In a two-pole stator supplied at 50 Hz, this is equivalent to two poles (N_S and S_S) rotating at 3,000 rpm. The rotor carries the field windings, which are supplied with DC to produce two poles (N and S). With the rotor stationary in the position shown in Fig 10a, there will be repulsion between N and N_S, and between S and S_S. A torque will therefore be exerted on the rotor in an anti-clockwise direction.

14-10 Fig 10 Principle of Operation of a Synchronous Motor



31. Half a cycle (i.e. 0.01 sec) later, the poles of the rotating stator field will have changed position, as shown in Fig 10b. There is now attraction between N and S_S, and between S and N_S, so that a clockwise torque is applied to the rotor. Because of its inertia, the rotor cannot respond to this rapidly alternating torque, and so remains at rest. The synchronous motor cannot, therefore, be started simply by applying AC to the stator winding, even if the rotor is already supplied with DC.

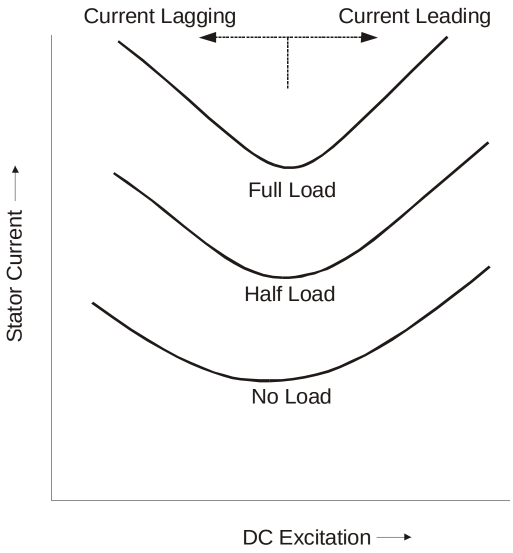
32. Suppose now that the rotor, driven by an external force, is already turning in a clockwise direction at just below synchronous speed. The relative speed between the rotating field and the rotor will be low, and eventually the N pole of the rotor will become adjacent to the S_S pole of the rotating field, as the latter overtakes the rotor. At this point, the two magnetic fields will lock together, and the rotor will maintain its position relative to the rotating field; that is, it will rotate at synchronous speed. The synchronous motor is usually started and run up towards the synchronous speed with the help of a small induction motor.

33. **Characteristics**.

a. **Effect of Load**. When a mechanical load is applied to a synchronous motor, the electrical input to the motor increases, the speed remaining constant. If too great a load is applied, the machine is pulled out of synchronism and comes to rest. The torque at which this occurs is called the 'pull-out torque'.

b. **Effect of DC Excitation**. If the mechanical load is kept constant, and the DC excitation varied, the back emf varies, and hence the supply current to the stator varies. A graph showing the variation of stator current as the excitation is varied is shown in Fig 11. The greater the load on the machine, the greater is the current and the higher is the curve on the graph. At a low value of DC excitation, the stator current is large and lags the applied voltage. At normal excitation, the current is a minimum and the phase angle zero. At a high value of excitation, the current again increases, but it now leads the applied voltage. This is an important property of the synchronous motor since, by over-exciting the field, the machine is made to take a leading current which compensates for any lagging current taken by other apparatus connected to the same supply. The power factor of the supply is thereby improved.

14-10 Fig 11 Effect of Varying the DC Excitation to a Synchronous Motor



34. **Summary**. The synchronous motor:

a. Requires to be started by an external prime mover.

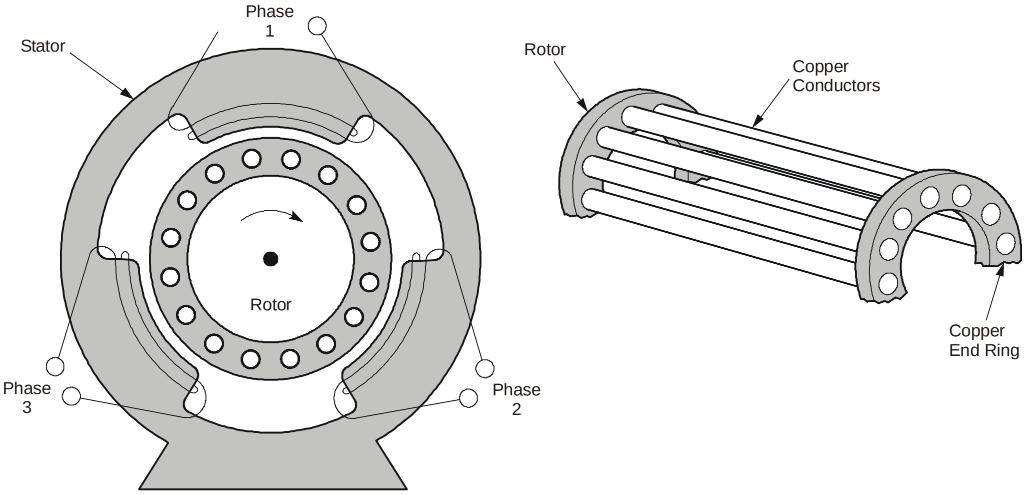
b. Runs only at the synchronous speed.

c. Can be used to adjust the power factor of a system at the same time as it is driving a mechanical load.

Three-phase Induction Motors

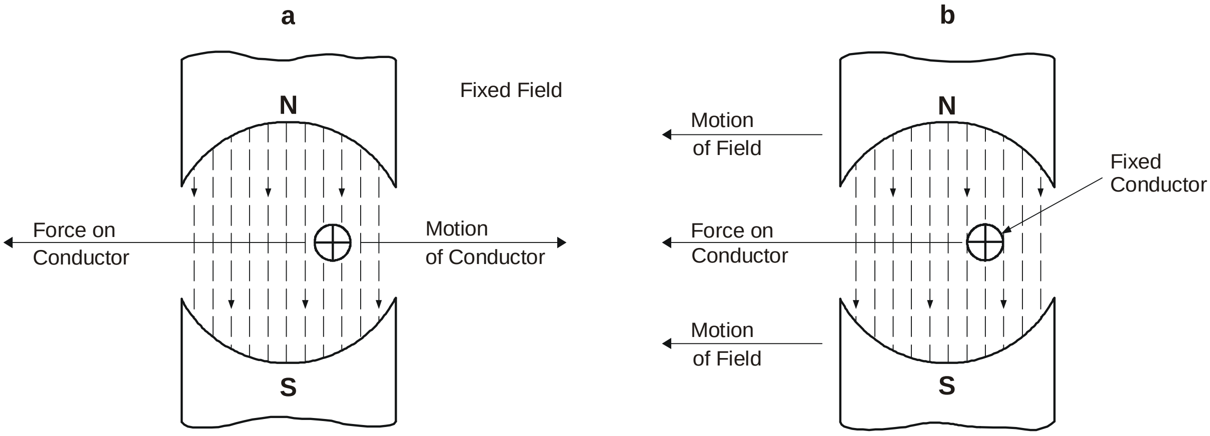
35. **Construction**. Fig 12 shows the basic construction of a simple type of three-phase induction motor (note that the stator windings are usually in pairs of poles as in Fig 8a). It consists of a three-phase stator winding supplied with three-phase AC to produce a rotating magnetic field. The rotor consists of a set of stout copper conductors, laid in slots in a soft-iron armature, and welded to copper end rings, thus forming a closed circuit. This is termed a 'squirrel-cage' rotor and it will be seen that there is no electrical connection to the rotor.

14-10 Fig 12 Three-phase Induction Motor



36. **Principle**. If a conductor is set at right angles to a magnetic field, as in Fig 13a, and moved across the flux from left to right, the direction of the induced emf will be into the paper (Fleming's Right-hand Rule – Volume 14, Chapter 6). If this conductor is part of a complete circuit, a current will be established in the direction of this voltage, and there will be a force on the conductor tending to urge it from right to left (Fleming's Left-hand Rule – Volume 14, Chapter 6). The same relative motion of field and conductor applies if the conductor is stationary and the field moves from right to left, as in Fig 13b; if the circuit is completed so that current can be established in the conductor, the conductor tends to move from right to left. The conductor experiences a force moving it in the same direction as the field's motion. As the conductor follows the field, the relative motion is reduced, thereby reducing the conductor current and the force on the conductor. Thus, the conductor speed is limited to something less than that of the field, otherwise there will be no relative motion, no current, and no torque.

14-10 Fig 13 Principle of Induction Motor



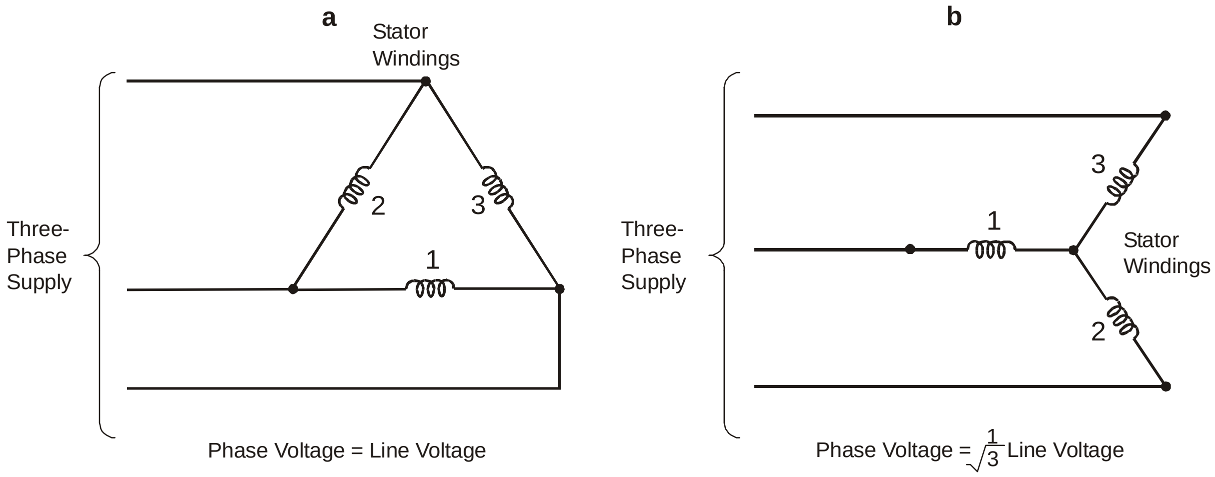
37. The squirrel-cage rotor of the induction motor, set in the rotating field of the stator, should accelerate until it is running steadily at a speed which is slightly less than the synchronous speed at which the magnetic field rotates. Thus, the rotor runs slightly slower than the rotating field, the amount depending on the load: the larger the load the greater the speed difference. In practice, very little speed change occurs between a light and a heavy load, and the main use of an induction motor is to drive a load at relatively constant speed.

38. **Slip**. The difference between the synchronous speed and the rotor speed, measured in rpm, is termed the 'slip speed'. The ratio of slip speed to synchronous speed, expressed as a percentage, is

termed 'slip'. For example, a six-pole motor supplied at 50 Hz will have a synchronous speed of $\eta_S = 60f/p = 60 \times 50/3 = 1{,}000$ rpm. With a rotor speed of 960 rpm, the slip speed will be 1,000 - 960 = 40 rpm and the slip will be 40 × 100/1,000 = 4%. In practice, the value of slip varies from 2%, for large machines, to 6% for small machines.

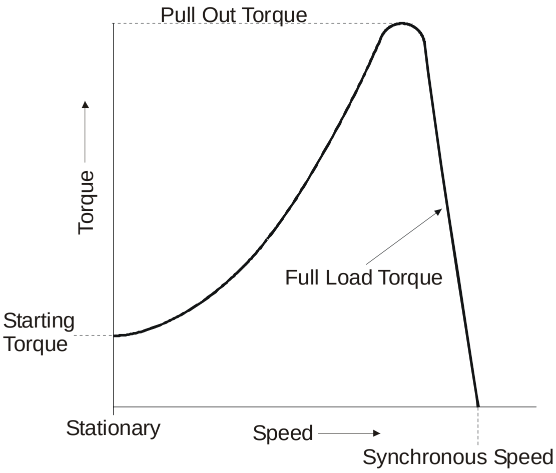
39. **Starting Current**. When the stator winding is energized, and the rotor stopped, the slip is 100% and maximum emf is induced in the rotor. A heavy current is thus established in the rotor, and this produces a flux which opposes and weakens the stator flux. The self-induced emf in the stator is therefore reduced, and a heavy current is taken by the stator winding on starting. To reduce this heavy starting current, the voltage applied to the stator windings should be at a reduced level on starting, until the rotor is turning at such a speed that its effect on the stator current is negligible. The normal way of doing this is to use a 'star-delta' starting switch. For normal running, the motor is designed to operate with the stator phases mesh or delta connected to the supply via the switch (Fig 14a), so that the phase voltage is equal to the line voltage. For starting, the stator windings are connected up in star (Fig 14b) via the switch to the supply, so that the phase voltage is $\frac{1}{\sqrt{3}}$ of the normal voltage. This reduced voltage limits the starting current to a safe level.

14-10 Fig 14 Star-delta Starter


a
Stator Windings
2
3
1
Three-Phase Supply
Phase Voltage = Line Voltage
b
3
1
Stator Windings
2
Three-Phase Supply
Phase Voltage = $\frac{1}{\sqrt{3}}$ Line Voltage

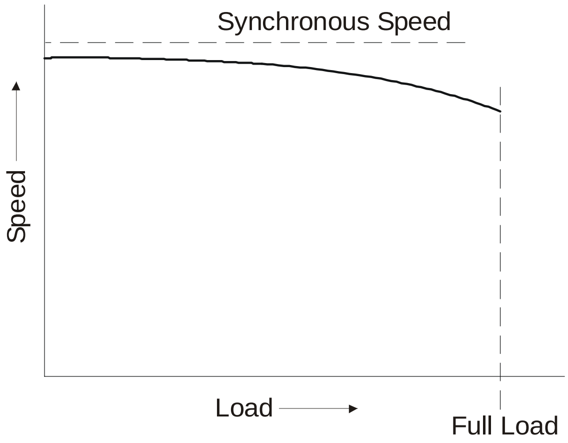
40. **Torque**. The frequency of the current induced in the rotor is the frequency with which the stator field rotates relative to each conductor. When the rotor is at rest, this frequency equals the supply frequency. When the motor is running lightly loaded, the slip is small, and the frequency of the induced rotor current may be only a few cycles per second. However, the resistance of a squirrel-cage rotor is small and its inductance high. Its impedance will, therefore, be large at the frequency of the supply when the rotor is stationary, and much less when it is running. Thus, on starting, the rotor current and the rotor emf are nearly 90º out of phase. The flux produced by this lagging rotor current is such that there is little interaction between it and the stator flux, and the starting torque is poor. As the rotor current comes round into phase with the rotor emf with increased rotor speed (decreased slip and inductive reactance), the rotor and stator fluxes come more into phase, and the torque increases. A typical torque-speed characteristic curve is shown in Fig 15.

14-10 Fig 15 Torque-Speed Characteristics of an Induction Motor



41. **Load-Speed Characteristics**. Off load, the torque developed by the rotor is only that required to overcome friction, and, in this condition, the rotor speed is almost synchronous. When a load is applied, the rotor slows down to the point at which the resultant increased driving torque balances the load torque. The fall in speed from no load to full load is small, as shown in Fig 16. As the load is increased, slip increases, until eventually a load is reached at which any further increase in slip, instead of increasing torque, reduces torque, and the motor will shut down. Usually, therefore, motors are designed so that the pull-out torque is at least twice the normal full load torque, to give an ample operating margin.

14-10 Fig 16 Load-Speed Characteristics of an Induction Motor



42. The speed of a squirrel-cage motor is not easily controlled, since it is related to synchronous speed, and its main use is on devices where a fairly constant speed is required. One typical use is as the prime mover for generators used in control systems, where ratings of 2 to 30 hp at speeds of up to 3,000 rpm may be required. The squirrel-cage motor can be readily adapted for frequent reversing. To do this, it is necessary to reverse the direction of rotation of the stator field, and this is achieved by changing over any two of the three connections at the stator terminals.

43. **Summary**. The three-phase squirrel-cage motor:

a. Has a high starting current (reduced by a star-delta starter).

b. Has a poor starting torque.

c. Runs at almost synchronous speed, and the speed cannot easily be varied.

d. Can be adapted for frequent reversing.

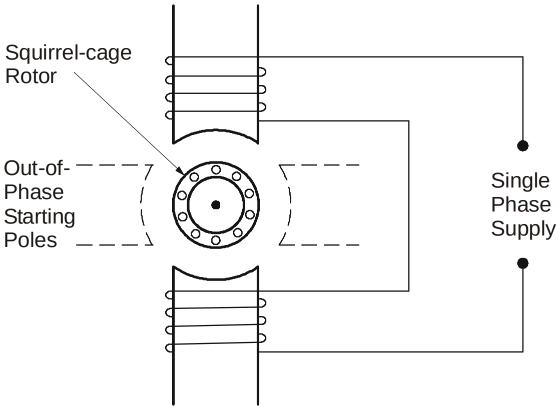
Two-phase Induction Motors

44. A rotating magnetic field is also produced if two phases, 90º out of phase with each other, are used instead of a three-phase supply. The production of the rotating magnetic field in a two-phase induction motor employs similar methods to those described for the three-phase motor, and the motor principles are also similar to those of the three-phase motor. Two-phase motors are less efficient than three-phase motors, and therefore less widely used.

Single-phase Induction Motors

45. Fig 17 represents a single-phase induction motor with one pair of stator poles and a squirrel-cage rotor. Such a motor is not capable of producing a rotating field in the manner previously described, and it is not self-starting.

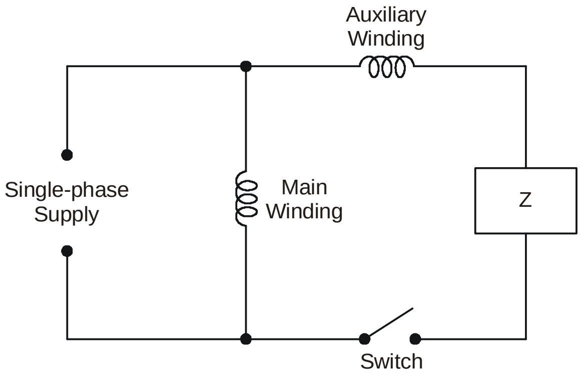
14-10 Fig 17 Single-phase Induction Motor



46. The field produced by the single-phase winding alternates according to the frequency of the supply. As the field changes polarity every half-cycle, it induces currents in the rotor which try to turn it through 180º but, as the force is exerted on the rotor axis, there is no turning force on the rotor at rest. If the rotor is turned by external means to overcome its inertia, an impulse every half-cycle will keep it rotating. As the field is pulsating rather than rotating, single-phase motors are not as smooth running as two or three-phase motors.

47. The starting device often takes the form of an auxiliary stator winding spaced 90º from the main winding (Fig 17) and connected in series with an impedance to the main supply. This impedance is chosen to produce as great a phase displacement as possible between the currents in the main and auxiliary windings, so that the machine starts up virtually as a two-phase motor (Fig 18). A switch, usually operated by centrifugal action, cuts out the auxiliary winding when a fair speed has been attained, and the machine continues to run on the main stator winding. Single-phase induction motors (operating on 230 V 50 Hz supply and rated up to 5 hp) are used to provide a relatively constant-speed drive to DC generators.

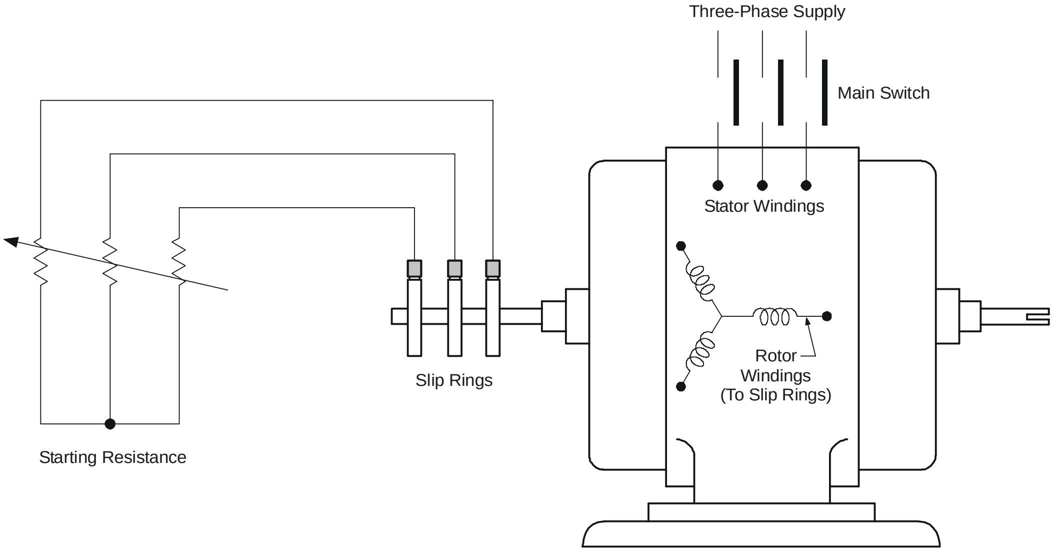
14-10 Fig 18 Starting Device for a Single-phase Induction Motor



Wound Induction Motors

48. The squirrel-cage motor takes a large starting current, and has a poor starting torque, due to the low resistance rotor. These features are improved in the wound or slip-ring type of induction motor. The stators of wound induction motors are identical to those of squirrel-cage motors, but the rotor conductors are insulated and form a three-phase, star-connected winding, the three ends of which are connected to three insulated slip rings mounted on the motor shaft (Fig 19). When the motor is running normally, these slip rings are short-circuited to give a low resistance rotor equivalent to the squirrel cage. For starting, the slip rings are connected to a three-phase, star-connected resistance, as shown in Fig 19, and maximum resistance is inserted in the rotor circuit, improving the starting torque and giving a lower starting current. The resistance is gradually cut out as the machine speeds up until, finally, the three slip rings are short-circuited, and the motor runs as for a squirrel-cage machine.

14-10 Fig 19 Slip Ring Induction Motor

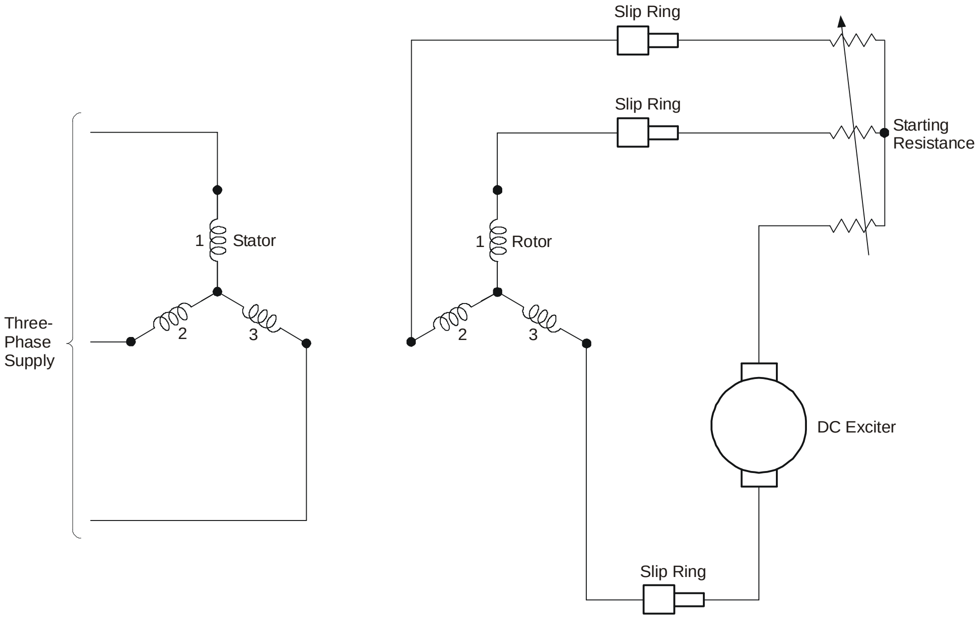


Synchronous Induction Motors

49. The features of constant speed and capacitor action of the synchronous motor are often combined with the self-starting feature of the induction motor by using a machine which is virtually a hybrid of the two types. One type of synchronous induction motor in wide use consists of a slip-ring induction motor coupled to a DC exciter; the rotor windings are connected, via the slip rings, to the exciter and a three-phase starting resistance, as shown in Fig 20. The exciter is driven by the motor, which is started up as a slip-ring induction motor. There is no appreciable DC from the exciter until the motor has attained some 90% of full speed. As synchronous speed is approached, the machine

automatically pulls itself into synchronism, and continues to run as a synchronous motor, with DC applied to the rotor windings from the exciter. If the machine is overloaded, it pulls out of synchronism like an ordinary synchronous motor but continues to run at reduced speed as an induction motor.

14-10 Fig 20 Synchronous Induction Motor



Single-phase Commutator Motors

50. The induction and synchronous types of AC motor, although possessing many excellent features, are not variable-speed machines. Furthermore, the starting torque of a squirrel-cage motor is poor, it takes a large starting current, and the power factor is low. For these reasons, AC commutator motors have been developed, in an attempt to obtain improved speed control, starting torque, and power factor.

51. An ordinary DC series motor, when connected to an AC supply, will rotate and exert a unidirectional torque. The ordinary DC series motor would not, however, be capable of giving an efficient and satisfactory performance for the following reasons:

a. Large eddy currents in the field magnets would cause undue heating.

b. Each coil of the armature winding, when short-circuited by the brushes, would give rise to destructive sparking at the brushes.

c. The power factor would be very low because of the highly inductive nature of the field and armature windings.

52. To overcome these drawbacks, a series motor designed for use on AC supply (or for use on either AC or DC, in which case it is known as a universal motor), is modified in the following manner:

a. The entire magnetic circuit is laminated.

b. The field winding is distributed in core slots, like the stator winding of an induction motor.

c. The armature winding is sub-divided to a greater extent than in DC machines. This necessitates a relatively large number of commutator segments, and gives a large commutator for the size of the motor.

d. The brushes are of high resistance carbon, and each brush is restricted in width so that it bridges only two commutator segments.

e. Connection between armature coils and commutator segments is often made through resistors, instead of by direct soldering. This reduces the circulating current when a coil is short-circuited by a brush.

53. The main characteristics of the AC series motor are similar to those of DC series motors (Volume 14, Chapter 6), and the variable series resistor method for starting and speed control of the DC motor is used for the AC motor. AC commutator motors are not very efficient, and their use is confined to fractional horsepower motors. They are widely used in household appliances.

ROTARY INVERTERS AND CONVERTERS

Rotary Inverters

54. A rotary inverter combines the functions of a DC motor and an AC generator. It is similar in construction to a rotary transformer (Volume 14, Chapter 6), except that the generator windings are connected to slip rings. The function of the machine is to invert the DC input to an AC output.

Rotary Converters

55. A rotary converter is a combination of an AC motor and a DC generator and converts an AC input to a DC output.

CHAPTER 11 - TRANSFORMERS

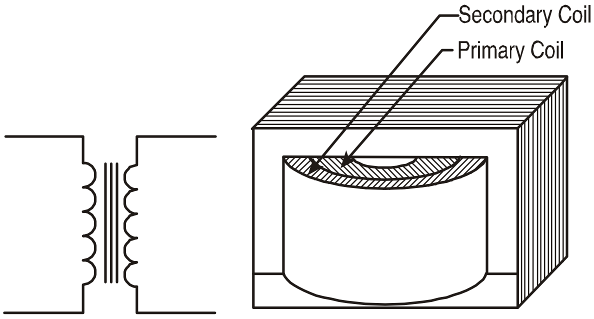
Introduction

1. In the main, transformers are used as a means of increasing or decreasing AC voltages. They are devices with no moving parts and depend on mutual inductance between two electric circuits for their action. Transformers are not amplifiers or generators, as no external power supplies are used. The power out of the device is slightly less than the power put in to it.

The Basic Voltage Transformer

2. The voltage transformer is a mutual inductor in which the coupling is arranged to be a maximum by concentrating the flux in an iron core, so that the flux through the input (primary) and output (secondary) windings is identical. Under ideal conditions, with the same flux in both primary and secondary windings, the transformer is said to have a coupling coefficient (k) equal to unity. In reality, device losses have to be taken into account and unity is never achieved. Fig 1 shows the construction of a typical transformer along with its symbolic representation.

14-11 Fig 1 A Typical Voltage Transformer

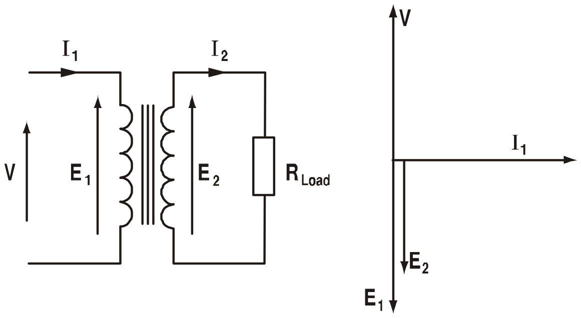


3. In the circuit shown in Fig 2, the applied AC voltage V causes a current I to flow in the primary which induces the same number of volts per turn in both primary and secondary windings of the transformer. The voltage induced in the primary, E_1, is a back emf and as such is in opposition to the applied voltage. If the primary has n_1 turns in its windings and the secondary has n_2, then the voltage induced in the secondary, E_2, is given by:

$$E_2 = \frac{n_2}{n_1} \times E_1 \text{(Back emf)}$$

The phasor diagram representation of this equation is also shown in Fig 2.

14-11 Fig 2 Voltages and Currents in a Transformer



Transformer Turns Ratio

4. The number of turns in the secondary winding (n_2) divided by the number of turns in the primary (n_1) is known as the turns ratio of the transformer. The secondary voltage can be made many times larger or smaller depending on the turns ratio. Transformers are referred to as either step-up or step-down devices.

5. Some transformers are designed with tapings on their primary and secondary windings in order to accommodate a wide range of voltages. In these cases, the transformer ratio is determined by the number of active windings in circuit for a given configuration.

Transformer Losses

6. During the process of transformer action two phenomena occur within the metal core, namely hysteresis and eddy current effects. Both of these give rise to power losses within the device and are present whether a load is connected across the secondary or not. These losses are grouped under the heading of magnetization or iron losses, and, because they are associated with the input, they may be likened to the internal resistance losses of a battery or generator. These losses may be minimized as follows:

a. Using magnetic materials with a high permeability and a small hysteresis loop area (see Volume 14, Chapter 2).

b. Constructing a core made up of laminated sheets rather than a solid block. The laminations are insulated from each other, giving high resistance to the eddy currents tending to flow.

7. When current flows in the secondary circuit of a transformer, ie load connected, additional losses are introduced which are known as copper losses. Copper losses are caused by the resistance of the windings, and, because they are associated with the output, they may be likened to the load across a battery or generator.

8. For any transformer, maximum efficiency occurs, i.e. maximum transference of power, when the load losses (copper losses) equal the internal fixed losses (iron losses). This is very much in line with the maximum power transfer theorem stated in Volume 14, Chapter 1.

Transformer Regulation

9. The rated voltage of a transformer is the voltage that appears across the secondary terminals on no-load, when the rated primary voltage is applied to the primary. However, the secondary voltage on-load will generally be less than this value; the drop depending on the load, the load power factor and the losses within the transformers. The value of the on-load voltage is particularly important because it is the value which applies in practice when the transformer is in operation with a load connected. The numerical value of the reduction is known as the inherent voltage regulation and is expressed as:

Secondary voltage, off-load (E_S) – Secondary voltage, on-load (V_S).

This is normally expressed as a percentage, or as a per-unit (pu) value.

$$\text{As a percentage} = \frac{E_S - V_S}{E_S} \times 100$$

$$\text{As a pu} = \frac{E_S - V_S}{E_S}$$

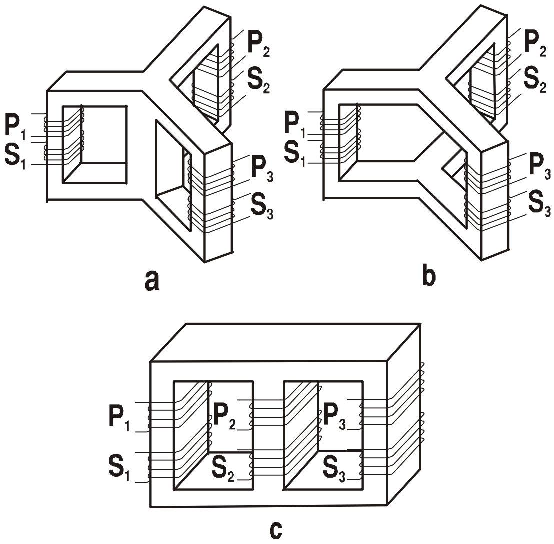
Three-phase Transformers

10. In a three-phase AC system, three separate transformers could be used in order to provide the necessary voltage transformation; however, it is more economical to use a single unit. It is usually more efficient, requires less space and weighs less than three transformers of the same capacity. Fig 3 shows the development of the three-phase core.

11. Fig 3a shows three single-phase transformers brought together to share a common central leg. However, the sum of the fluxes in the common leg will always be zero, and it is therefore not necessary. Fig 3b shows a combined core without the central leg, but this shape is difficult to make. The practical shape is shown at Fig 3c and is a distorted version of the previous form.

12. The primary and secondary coils of a three-phase transformer may be connected together in various star and delta configurations, depending on load and circuit requirements.

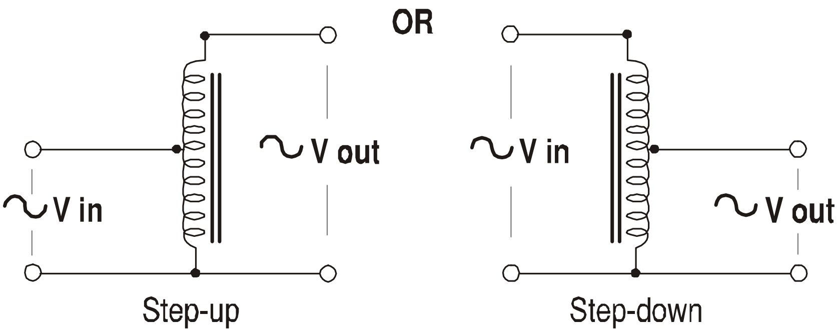
14-11 Fig 3 The Development of a Three-phase Core



Auto Transformers

13. Any tapped inductance can be used as a transformer, and this is illustrated in Fig 4. This device is called an auto transformer and the action is similar to that of the conventional type. The main difference is that part of the inductance is common to both primary and secondary circuits and as such provides direct coupling between them. Primary and secondary currents are essentially in antiphase and equal in value when the turns ratio is zero. The inductance then carries zero current. The auto transformer is particularly efficient for small transformer ratios.

14-11 Fig 4 The Auto Transformer (Principle of Operation)



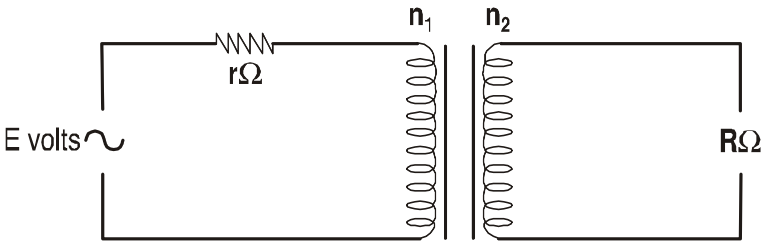
Transformer Matching

14. The maximum power transfer theorem established that maximum transference of power takes place when the resistance of the load is equal to the internal resistance of the supply source. Under these conditions the power source and load are said to be matched.

15. Unfortunately, in practice, circuit loads very rarely equal the internal resistance of the supply and transformers are used with suitable turns ratios to provide the required matching. This is shown in Fig 5. A turns ratio $\left(t = \frac{n_2}{n_1}\right)$ is chosen so that:

$$r = \frac{R}{t^2} \text{ or } t = \sqrt{\frac{R}{r}}$$

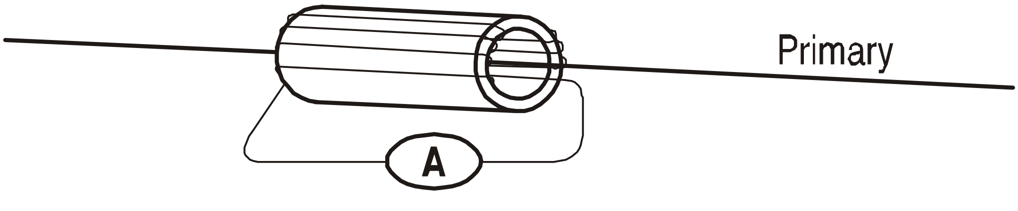
14-11 Fig 5 Transformer Matching



Current Transformers

16. Current transformers, sometimes referred to as instrument transformers, utilize the current or voltage transformation properties of the normal transformer. The main difference in operation is that the primary winding forms part of the circuit under test (see Fig 6).

14-11 Fig 6 A Current Transformer



17. Current transformers are used for:

a. Measuring current or power in high voltage systems.

b. Measuring high current with low range instruments.

c. Operating relays and other control devices according to the current in a system.

CHAPTER 12 - FUNDAMENTAL ELECTRONIC COMPONENTS

Introduction

1. Electronics is the science and technology of controlling suitably modified electron flow so as to convey information. A good deal of electronic equipment looks complicated and complex on first examination, but it has to be realized that all electronic circuits are made up of relatively simple basic units each performing a specific function. The way in which these units, or 'building blocks', are assembled governs the function of the complete system.

2. The whole of modern electronic circuitry is built on a foundation of resistors, capacitors, inductors, and semiconductor devices. From these simple components, intermediate units are assembled, such as:

a. Amplifiers, to increase the size of signals.

b. Oscillators, to generate desired waveforms and signal frequencies.

c. Mixers, to combine signals of different frequencies.

d. Power supplies, to energize the whole system.

e. Modulators, to superimpose information on radio waves.

f. Demodulators (detection), to remove information from radio waves.

g. Transducers, to convert a physical quantity (force, light, sound, etc) to an electrical signal, and vice versa.

h. Logic gates, to perform logic functions.

i. Shift registers, for use in microprocessors.

3. Electronic components fall into two main categories, namely active and passive devices. Resistors, capacitors, and inductors are regarded as being 'passive', while transistors and allied semiconductor devices are regarded as 'active', because they modify the power supplied to them. This chapter is devoted to the subject of active devices; passive devices are covered in detail in Volume 14, Chapter 1.

SEMICONDUCTORS

Introduction

4. Most materials used in electricity and electronics are either conductors or insulators. However, a few materials do not fall into either of these categories because conduction through them is too small for them to be classed as conductors, and too large for them to be insulators. Such materials are known as semiconductors. Germanium and silicon are the two most common semiconductors.

Doped Semiconductors

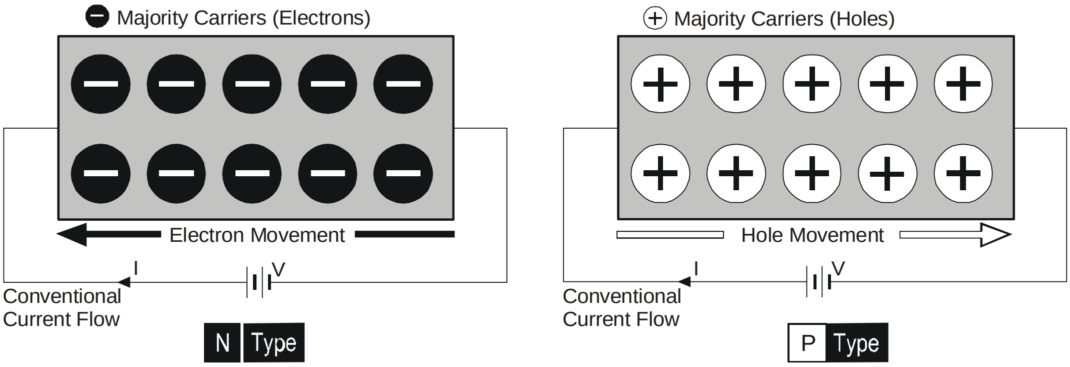
5. At ordinary room temperatures, a pure semiconductor has few free electrons. However, if very small quantities of another selected element, such as indium or arsenic, are combined with the pure semiconductor, conductivity is increased. This process of adding 'impurities' to the pure germanium or silicon is known as 'doping'. The resulting conductivity of the semiconductor depends upon the amount of doping, and can be strictly controlled.

N-Type and P-Type Semiconductors

6. Doped semiconductors can be one of two types (n-type or p-type) depending upon the element that is added during doping. For an n-type semiconductor, the result of doping is to produce free

electrons which, as negative charges, become available as n-type 'charge carriers' (n for negative). Conversely, for a p-type semiconductor, doping produces positive (or p-type) charge carriers known as 'holes'. A hole has a positive charge, equal and opposite to that of an electron. If a voltage is applied to an n-type semiconductor, the resulting current is due mainly to a movement of electrons; in a p-type, it is due mainly to a movement of holes (Fig 1). The moving charges in each case are known as 'majority carriers'.

14-12 Fig 1 Flow of Electrons and Holes


Majority Carriers (Electrons)
Majority Carriers (Holes)
Electron Movement
Hole Movement
I
V
I
V
Conventional Current Flow
Conventional Current Flow
N Type
P Type

Note: While hole movement is in the **same** direction as conventional current flow, electron movement is in the **opposite** direction.

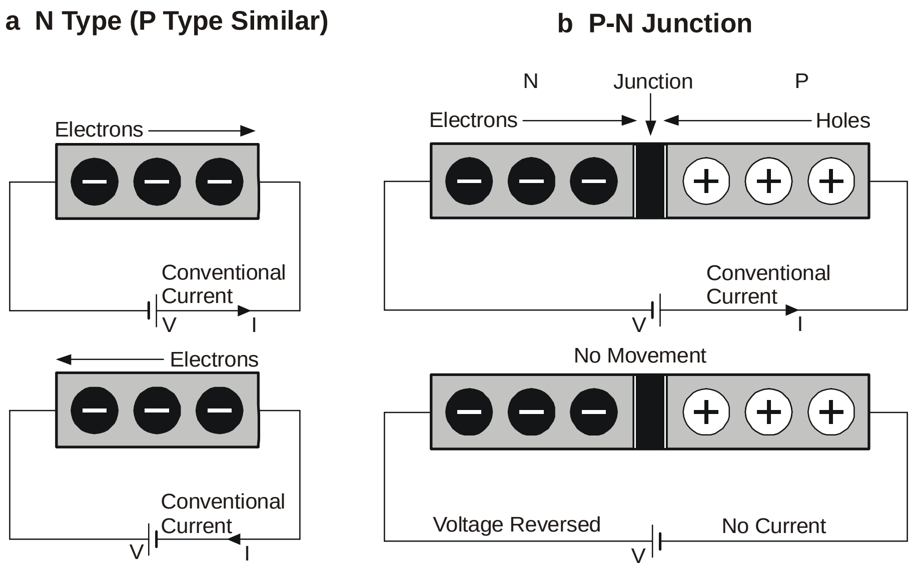
7. Although the majority of charge in an n-type is carried by electrons, there is a small flow of charge due to hole movement, known as 'minority carriers'. Similarly, minority carriers are electrons in p-type material where the majority carriers are holes. Minority carriers can be ignored when considering the basic operation of semiconductor devices, but they are very important when determining breakdown voltages, heating, and the efficiency of devices.

P-N Junctions

8. A single piece of n-type or p-type semiconductor conducts readily in either direction; reversing the applied voltage merely reverses the direction of current (Fig 2a).

9. In a continuous piece of pure semiconductor that has been doped in such a way that one end of the semiconductor has n-type properties and the other end p-type, conduction takes place easily in only one direction (Fig 2b). This is the basis of the p-n junction diode. Note that it is not sufficient merely to join a piece of n-type semiconductor to a piece of p-type, the junction must occur in a continuous piece of single-crystal semiconductor.

14-12 Fig 2 Direction of Flow



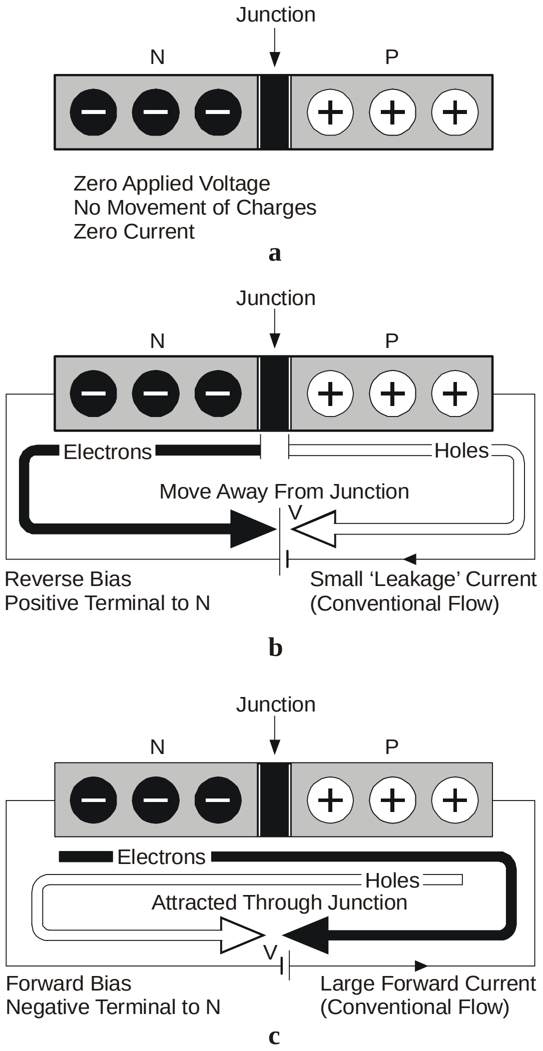
Properties of P-N Junctions

10. When a p-n junction is formed, the distribution of charges in the n and p-type materials is such that there is no net flow of current across the junction. Thus, under the condition shown in Fig 3a, with no external voltage applied, there is no current flow.

11. If a battery is connected across the semiconductor slice, as shown in Fig 3b, the positive terminal attracts electrons in the n-type region away from the junction, and the negative terminal of the supply has the same effect on the holes. There is now even less possibility of current flow through the p-n junction. A junction diode operating under these conditions is 'reverse biased'. In practice, in a reverse biased diode, a very small reverse current flows (measured in micro amps) and this is called the leakage current; due to minority charge carriers.

12. If the external battery is reversed, electrons are attracted from the n-type region through the junction to the positive terminal of the battery, and holes are attracted the other way. The two movements combine to give a large forward current through the junction in the direction shown. A junction diode operating under these conditions is 'forward biased', as shown in Fig 3c. A large current (normally measured in milliamps) flows for a small forward bias voltage; this indicates that the diode has a low resistance in the forward direction.

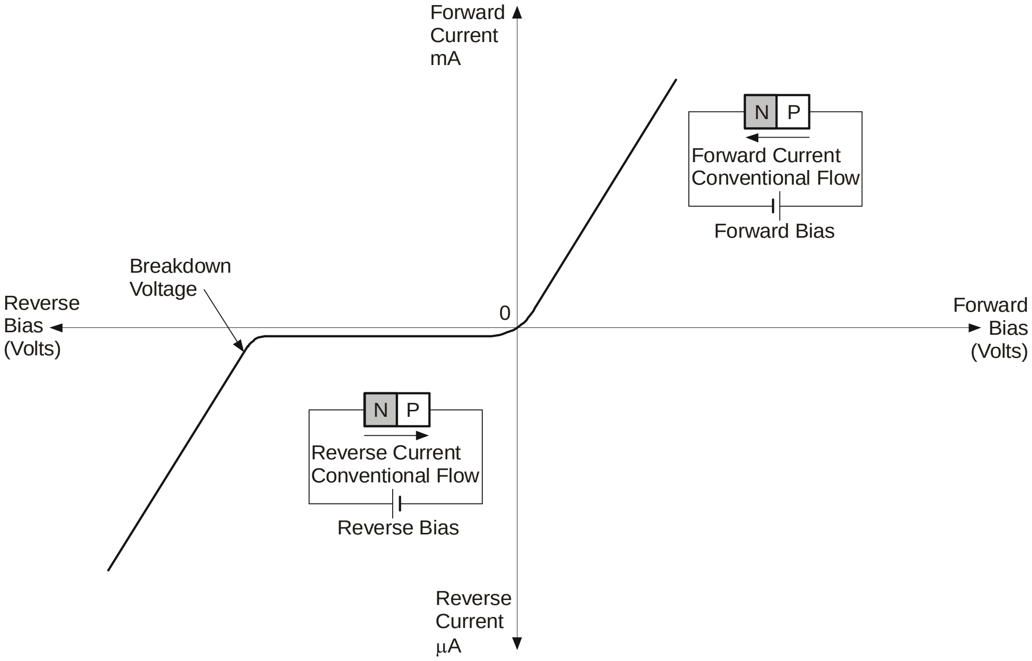
14-12 Fig 3 Effect of Bias in P-N Junction Diode



P-N Junction Diode Characteristics

13. By measuring the current flowing in the external circuit for various values of reverse and forward bias voltages applied to a diode, we can obtain a series of readings relating the current and the voltage. From these readings, a 'characteristic' of the diode may be plotted showing how the diode current varies with the voltage applied to it (Fig 4).

14-12 Fig 4 Typical Diode Characteristics



Note: Because forward current scale is in mA and the reverse current scale is in μA, reverse current is scaled up 1,000 times. Without this scale magnification, the reverse current would appear to be zero up to the breakdown point.

14. Fig 4 shows that current flows more easily in the forward direction than in the reverse direction. After the initial 'bend' around zero volts, the current in the forward direction is proportional to the forward bias voltage and the graph becomes approximately a straight line. When the diode is reverse biased, only a very small reverse (leakage) current flows until a certain 'breakdown' point is reached. At this point, the reverse current rises to a high value, often sufficient to destroy the diode. The reverse voltage at which the diode breaks down depends upon the construction of the device, but a diode must not be used in circuits where the reverse voltage that can be applied to it exceeds its breakdown voltage.

Junction Diode as a Switch

15. When a p-n junction is forward biased, its resistance is very low, often a fraction of an Ohm. When reverse biased, the very small leakage current, even at high reverse voltages, means that the resistance in this direction is very high, of the order of megohms. Thus, by applying suitable forward and reverse voltages to a diode, it may be used in the same way as a switch: ON when forward biased; OFF when reverse biased. Diode switches have many uses in electronics. The switching is faster than that of a mechanical switch, the diode may be very small, and there are no moving parts to wear out.

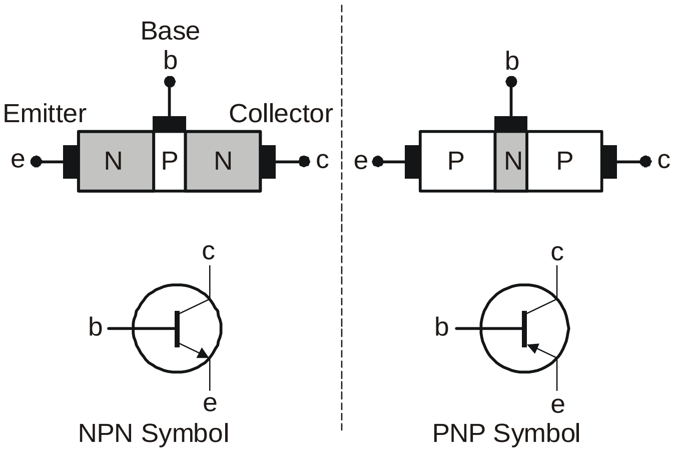
16. For rapid switching (of the order of picoseconds) at high powers (up to l0 kW peak), a PIN diode is often used. This has an undoped (intrinsic) layer sandwiched between the p and n layers. This layer alters the diode characteristics to give the high speed, high power requirements.

TRANSISTORS

Introduction

17. A semiconductor diode is a two-electrode device, one end having p-type properties and the other end n-type. A transistor is a three-electrode device and has three layers of semiconductor, alternately p-type and n-type. The three layers merge into one another to form a sandwich (Fig 5). One of the outer layers is known as the 'emitter', the other is the 'collector', and the thin centre layer is known as the 'base'.

14-12 Fig 5 Two Types of Transistor



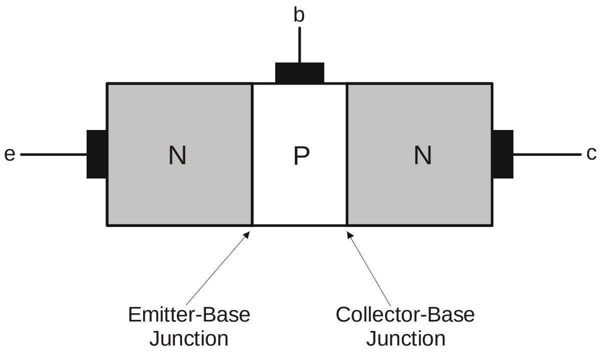
18. There are p-n-p or n-p-n transistors. In an n-p-n transistor, the emitter and collector are n-type semiconductors and the base is p-type. In a p-n-p transistor, the semi-conductor materials are interchanged. The circuit symbols are also shown in Fig 5. The arrow on the emitter points in the direction of conventional current flow; for an n-p-n transistor, this is into the emitter, whilst in a p-n-p type it is away from the emitter.

19. A transistor has many functions: it may be used as a switch, as a control device, or as the active device in amplifier and oscillator circuits. It has the advantages of being very small and light; but, even more important, it is reliable, it requires very little power for its operation and, if operated correctly, has a long life.

Transistor Junctions and Bias Voltages

20. In a semiconductor junction diode, there is one p-n junction. In a transistor, we have two: one between the emitter and the base, the other between the collector and the base (Fig 6).

14-12 Fig 6 Transistor Junction



21. For a transistor to operate, it must have suitable voltages applied across each junction, and these voltages must be of the correct polarity; in other words, the transistor has to be properly biased. A p-n junction is forward biased when the p-type semiconductor is connected to the positive terminal of the supply; under this condition, the resistance of the junction is low and current flows easily. A junction is reverse biased when the p-type semiconductor is connected to the negative terminal of the supply; the resistance of this junction is high and the only current flowing is that due to leakage.

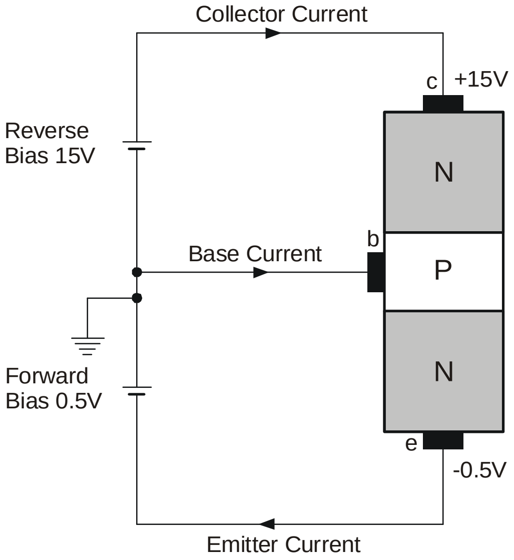
22. Under operating conditions, the emitter-base junction in a transistor is forward biased and the collector-base junction is reverse-biased. This applies whether the transistor is n-p-n or p-n-p.

Because of the forward and reverse bias, the emitter-base junction has a low resistance and the collector-base junction a high resistance.

Current Flow in Transistors

23. Transistor action relies upon a controllable current flowing between collector and emitter. Consider the n-p-n configuration shown in Fig 7. Electrons are flowing from the emitter to the base across the forward-bias junction. These electrons would normally flow out of the base, and could not go on to the collector because the base-collector junction is reverse biased. However, if the base junction is very thin, there are more electrons arriving from the emitter than there are positive holes in the p-type material. The base-collector junction then effectively becomes forward biased, so most of the free electrons in the base can continue their journey to the collector.

14-12 Fig 7 Current in NPN Transistor



24. There is a large current between collector and emitter and a small current into the base. Controlling the current which can enter the base, controls the junction biasing and the current which flows between the collector and the emitter, ie a small base current controls a large collector-emitter current.

25. Similar remarks apply to a p-n-p transistor, except that the batteries must be reversed to provide the correct bias. The forward-biased p-type emitter injects holes into the n-type base, with most of the holes moving through the base region to be collected by the collector.

Advantages of Transistors

26. Transistors and their allied semiconductor devices have now replaced thermionic valves in almost every type of electronic equipment. Their success has been due to the great advantages they offer over valves, that is:

a. Much smaller size.

b. They do not require heating.

c. The ability to withstand greater levels of vibration and shock.

d. No warming up period.

e. Supply voltages are much lower.

f. Longer life.

g. They are better suited for use in remote locations, such as space satellites and booster/repeater amplifiers.

27. The early problems associated with limited power output, noise, frequency response, and operating temperature are all being overcome as research intensifies. Nowadays there are very few applications in which a thermionic device is superior to a transistor.

FIELD EFFECT TRANSISTORS

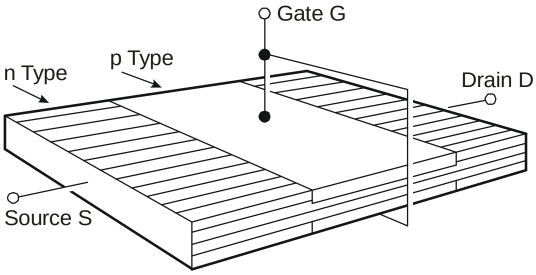
Introduction

28. The type of transistor so far discussed is often called the 'bipolar' transistor because there is a simultaneous flow of both main and leakage currents. Another type of transistor which has been introduced has certain advantages and disadvantages when compared with the bipolar type. It is known as the field effect transistor (FET). There are basically two types of field effect transistor, the junction-gate (or 'jugfet') and the insulated gate FET (or 'igfet').

The Junction Gate FET

29. The schematic outline of a basic FET is shown in Fig 8. It consists of a slice of n-type silicon into which two p-type regions are inserted opposite each other. The p-type regions are joined together externally and form a control electrode known as the gate (G). One end of the slice is referred to as the source (S) and the other end is the drain (D). The source is comparable to the emitter in a bipolar transistor, the gate to the base, and the drain to the collector.

14-12 Fig 8 Basic Construction of the FET



30. Fig 9 indicates how the FET works. It normally has a conducting n-type channel between S and D. Therefore, if D is made positive relative to S, charge carriers (electrons in this case) flow from S to D, channelling between the p-type gate regions. If G is made negative to S, it effectively repels the moving electrons and 'squashes' them into a narrower funnel. In other words, the conducting channel becomes narrower and the drain current decreases. By making G more negative, a point is reached where the channel is 'pinched off' altogether and the drain current falls to zero.

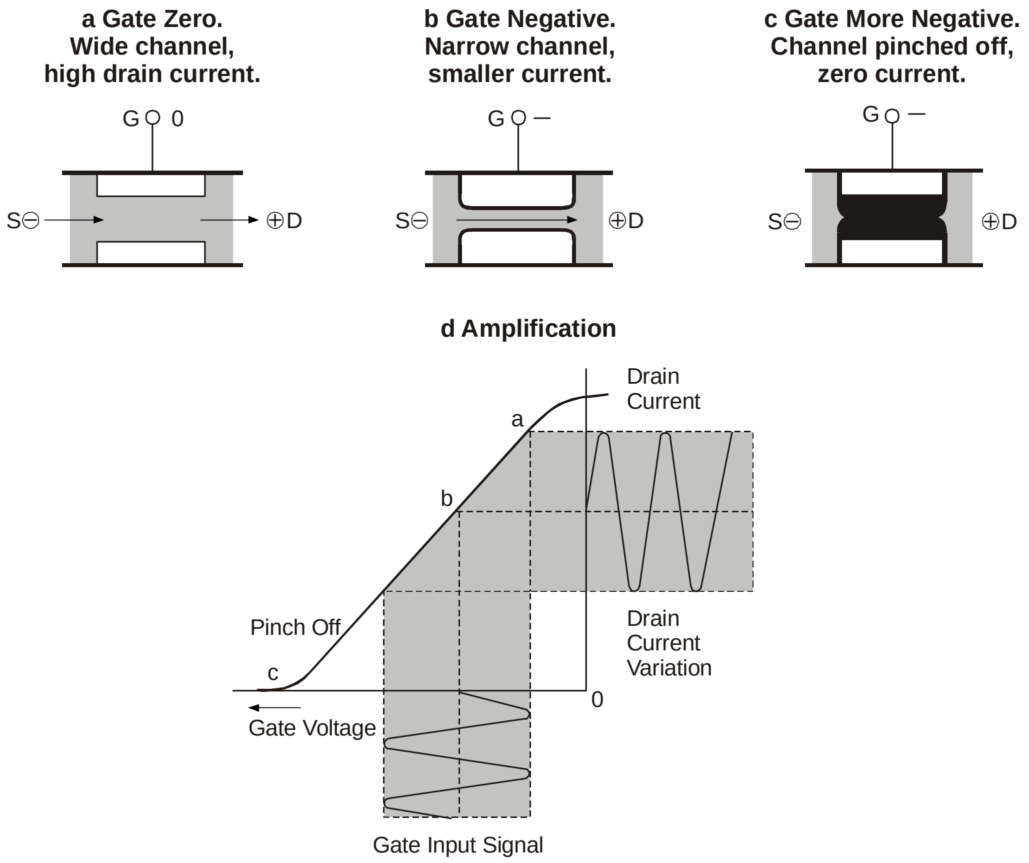
31. If a signal voltage is applied to G, the drain current is caused to vary in sympathy. By suitable choice of load resistor in the drain circuit, the output voltage variations produced by the drain current can be much larger than the applied signal voltage. Amplification may therefore be obtained. The characteristics of a FET differ from bipolar transistors in three ways:

a. The bipolar transistor is a current-operated device, ie it depends for its operation on the current applied as input to the emitter-based circuit. For a current to flow for a small signal voltage, the input resistance of the device must be low. In a FET, there is no current in the gate electrode. Thus, the FET is a voltage operated device, and has a very high input impedance, which is a requirement in many circuits.

b. The FET has practically no leakage current.

c. Transit time effects are negligible, so that the FET has a high cut-off frequency.

14-12 Fig 9 How the FET Works



a Gate Zero. Wide channel, high drain current.

b Gate Negative. Narrow channel, smaller current.

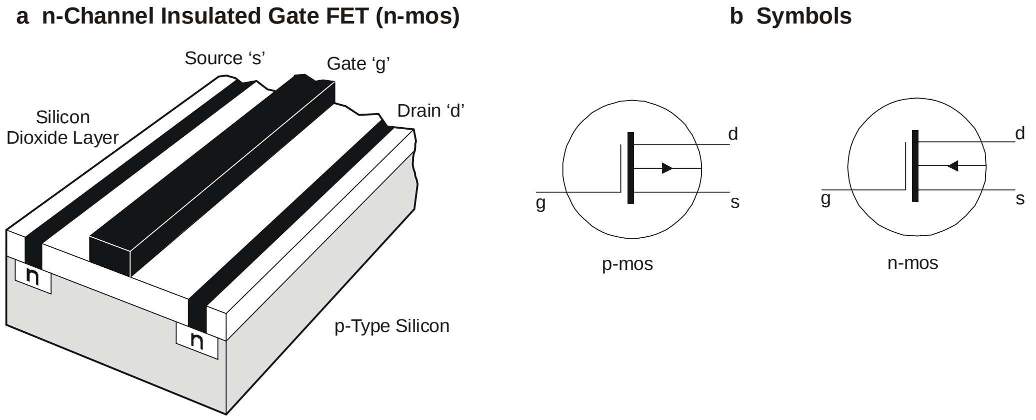
c Gate More Negative. Channel pinched off, zero current.

d Amplification

The Insulated Gate FET

32. Junction FETs are now only manufactured for replacement parts; all new FETs are of the Insulated Gate family. The basic construction of an insulated gate FET is shown in Fig 10. In this type of FET, the gate is insulated from the channel and, in the example shown, the source and drain connections are to two n-regions contained within a p-region. The gate is a metallic layer separated from the p-region by an insulating layer of silicon dioxide. The sequence of materials - metal, oxide, semiconductor - has given the device its alternative name of 'mosfet'. Although this is an n-channel device, there is, in fact, no n-layer between the source and the drain until the gate is biased positively with respect to the p-region (known as the 'base' or 'substrate'). Such a potential attracts free electrons from the n and p-regions to the upper surface of the p-region, where they form an n-channel bridging the two n-regions and so providing a conducting path between source and drain. It is significant that in the absence of a gate bias, no drain current can flow; a positive gate bias is necessary to give a drain current.

14-12 Fig 10 MOSFET



INTEGRATED CIRCUITS

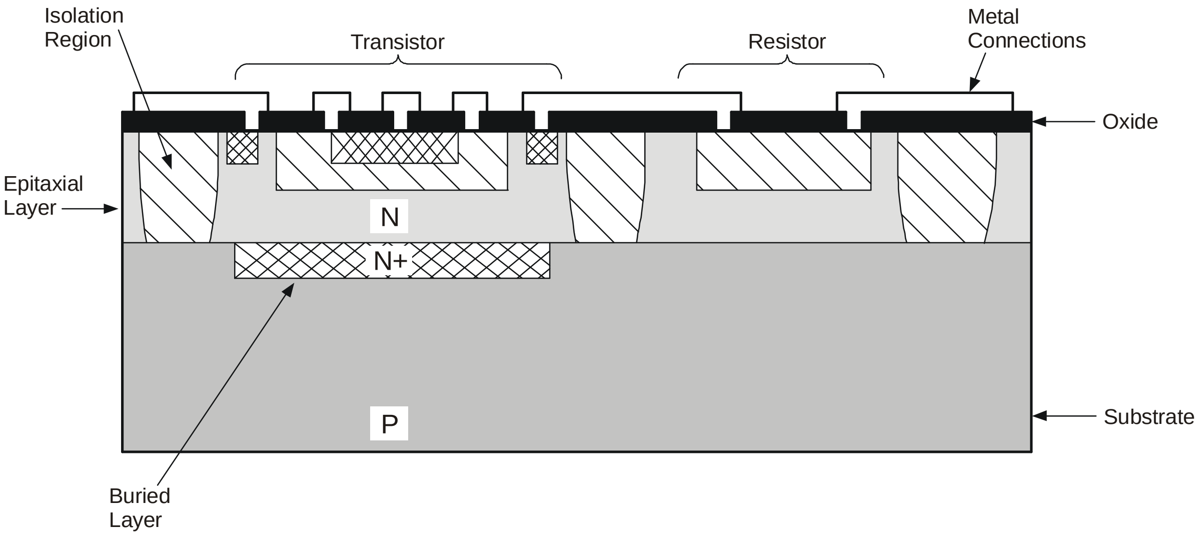
Introduction

33. The twin requirements for small size and reliability, particularly in military systems, have led to many different approaches to the miniaturization of electronic circuits. Perhaps the most important of these is the monolithic silicon integrated circuit.

Composition

34. Silicon integrated circuits are manufactured by a diffusion process developed from the planar technique, used for making transistors. In this process, hundreds, or even thousands, of transistors and other components can be diffused into a single silicon slice, which is then separated into individual circuit wafers with dimensions typically 1.5 mm × 1.5 mm × 0.3 mm. These components are connected together in the desired circuit configuration by a network of metal interconnections on the surface of a wafer. Fig 11 represents a simplified cross-section of a piece of such an integrated circuit containing a resistor and a transistor. Diodes can also be incorporated into integrated circuits, and small-value capacitors can be provided by using reverse-biased diodes which exhibit capacitance of a few pF. Larger value capacitors and inductors must be added externally to the diffused integrated circuits. A far greater degree of miniaturization can be achieved using MOS techniques; because of their high input resistance and simple construction, thousands of these devices may be diffused on a single silicon slice.

14-12 Fig 11 Piece of an Integrated Circuit



Applications

35. The range of applications for integrated circuits is almost limitless; they can be used to provide all the basic computer circuits, such as NAND and NOR gates, and binary counters. Indeed, it is the widespread use of integrated circuits which has made possible high capacity, small size computers. There are also integrated circuits for single and multiple linear amplifier stages, pre-amplifying circuits and operational amplifiers.

Development

36. A development in the field of integrated circuits is the concept of forming the system wiring between integrated circuits while they are still on the whole silicon slice, so that a single slice will form a complete electronic sub-system. This technique is generally called large-scale integration (LSI). It is important because one of the major causes of failure in electronic equipments is failure in the interconnecting wiring. The integrated circuit, with its system of interconnections within each circuit (formed as part of the slice fabrication process), gave a significant improvement in reliability. It is a logical step, in order to obtain still higher reliability, to produce the interconnections between circuits while they are still on the slice. This concept has been extended to cover very-large-scale integration (VLSI) and extra-large-scale integration (ELSI).

CATHODE RAY TUBES

Introduction

37. The cathode ray tube (CRT) is an electron tube in which electrons emitted from a cathode are formed into a narrow beam, accelerated to high speed, and directed to a screen coated with a fluorescent material. This material glows at the point of impact, producing a visible dot. A changing field, either electrostatic or electromagnetic, between the source of electrons and the screen causes the dot of light to move in accordance with the field variations.

38. There are two basic types of CRT; the electrostatic CRT and the electromagnetic CRT. In the electrostatic CRT, focusing and deflection are done with electronic fields, while in the electromagnetic CRT, magnetic fields do these jobs.

The Electrostatic CRT (ESCRT)

39. Fig 12 shows the main parts of an electrostatic CRT. These parts can be divided into three groups:

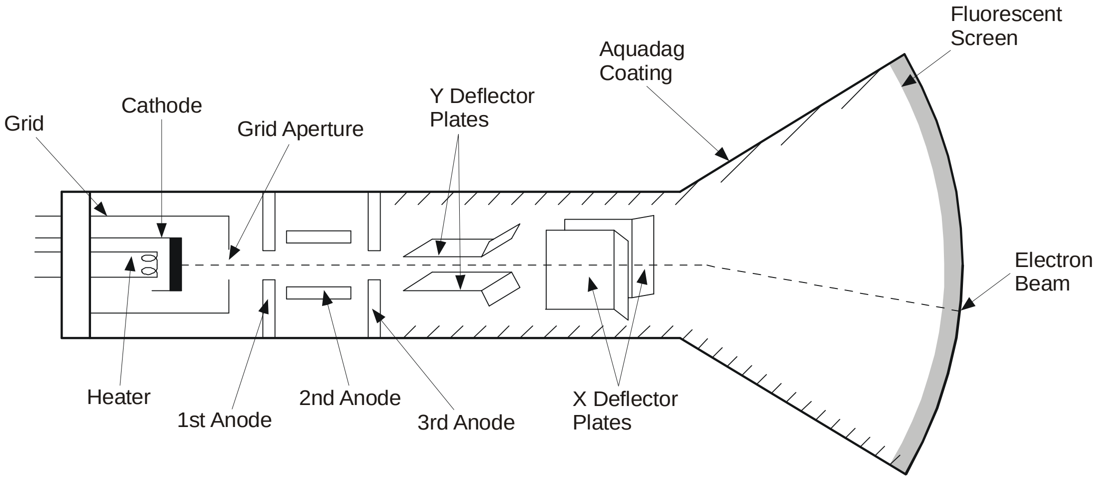
a. The electron gun, which produces a stream of fast-moving electrons and focuses them into a narrow beam. Sometimes the electron gun refers only to the cathode and grid; the anodes are then called the focusing system.

b. The deflecting plates, which enable the beam of electrons to be moved up and down and from side to side.

c. The fluorescent screen, which shows the movement of the beam by producing a spot of light.

All of the electrodes are enclosed in an evacuated glass envelope.

14-12 Fig 12 The Electrostatic CRT (ESCRT)



40. The cathode is a small tube, coated at the end with an oxide which emits electrons when heated. The grid is a hollow cylinder surrounding the cathode, with a central hole through which the electrons pass. The grid is made negative with respect to the cathode and, by varying this voltage, the number of electrons in the beam is varied, thus controlling the brilliance of the spot of light on the CRT screen. The brilliance control alters the voltage on the grid of the CRT. If the grid is made sufficiently negative, the electron beam will be completely cut off, and the spot on the screen will be blanked out.

41. The first and third anodes are circular plates, with holes through their centres. They are held at a high positive voltage relative to the cathode (several hundred, or even thousand, volts), and so they accelerate the electrons to a high speed. The third anode voltage is higher than the first anode voltage.

42. The second anode is a hollow cylinder mounted between the first and third anodes. Its purpose is to focus the electrons into a narrow beam and, for this reason, it is made negative with respect to the other anodes; this voltage can be varied to adjust the focusing of the beam. Hence, the focus control varies the voltage on the second anode.

43. In practice, the third anode is earthed and the other electrodes are made negative with respect to it. Typical EHT (extra high tension) voltages used in CRTs are: cathode –4 kV; grid –4.02 kV (variable); first anode –2 kV; second anode –3 kV (variable). Third anode and screen OV EHT voltages are needed in order to give the electrons enough speed to produce light on the fluorescent screen.

44. Two sets of deflecting plates are mounted after the third anode, as shown in Fig 12. By applying varying voltages to these plates, the focused beam can be swung in any direction. The plates nearer the third anode (the Y plates) are used to move the beam vertically, and those nearer the screen (the X plates) deflect the beam horizontally. The plates are often flared at the ends to provide the required amount of deflection without fouling the beam. When a voltage is applied to the plates, the electron beam will be attracted or repelled, depending on the polarity.

45. For horizontal deflection, a sawtooth voltage is applied to the X plates. This voltage increases from minimum to maximum values at a uniform rate (the sweep) and then returns rapidly to minimum (the fly-back). The waveform is repetitive and causes the beam to move from the left-hand to the right-hand side of the screen and then to return quickly to the left-hand side to start another sweep. The fly-back may be 'blanked out' by applying either a negative pulse to the grid or a positive pulse to the cathode during this period.

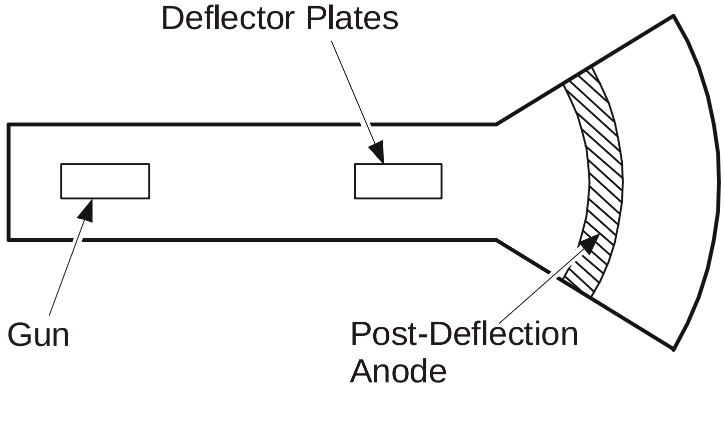
46. Thus, with this waveform applied to the X plates, a horizontal line is produced on the screen. When a voltage pulse is applied to the Y plates, the beam is vertically deflected for the duration of the pulse and a 'blip' appears on the screen.

47. The fluorescent screen is a chemical coating on the inside end of the glass tube. When fast-moving electrons hit the screen, they cause it to glow with a colour which depends upon the type of chemical used. The spot of light remains for a time after the electron beam has moved away; this effect is after-glow and enables a complete steady picture to be seen. In sonic types of radar display, the time base trace rotates slowly round the centre of the CRT screen and a chemical with a long after-glow is used to retain the picture.

48. When electrons hit the screen, secondary electrons are emitted. These are conducted via a powdered graphite coating, called the 'aquadag', to the third anode (earth). The coating prevents the screen becoming negatively charged.

49. In some types of CRT, an extra anode is used. It is made of a conducting ring of powdered graphite held at about twice the voltage of the third anode with respect to the cathode. This enables small input signal voltages to produce large displays on the screen, ie the deflection sensitivity of the tube is increased (Fig 13).

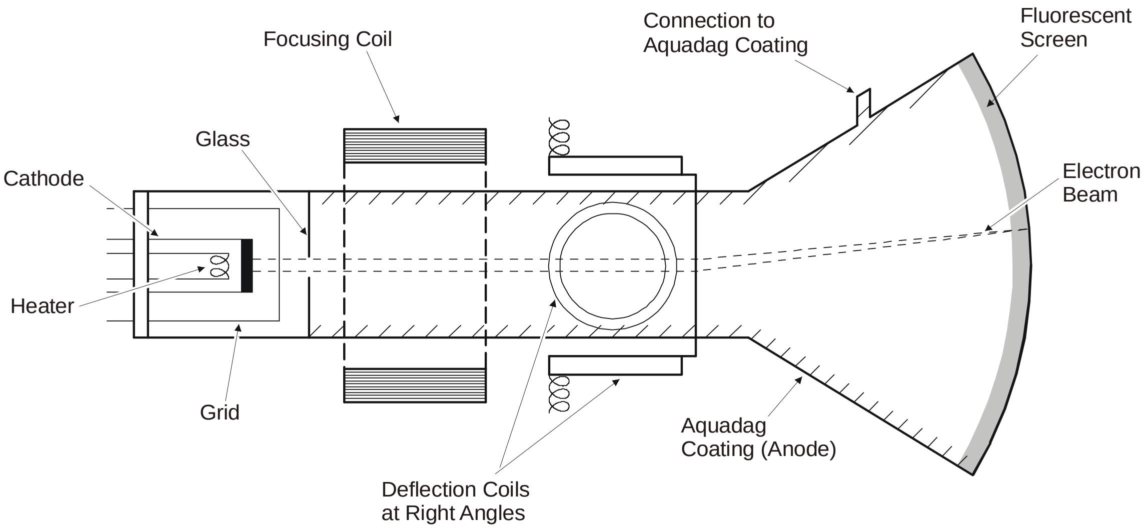
14-12 Fig 13 Post-deflection Acceleration



The Electromagnetic CRT (EMCRT)

50. The construction of an electromagnetic CRT, employing magnetic focusing and deflections, is shown in Fig 14. The heater, cathode, and grid are the same as for an electrostatic CRT, but there is usually only one anode. This may be the aquadag coating, connected to a suitable voltage. Deflecting and focusing currents are passed through coils mounted on the outside of the neck of the tube. The electron beam is acted upon by a sideways force as it passes through the magnetic field around the coil, in much the same way as a current-carrying conductor has a force exerted on it in an electric motor.

14-12 Fig 14 The Electromagnetic CRT (EMCRT)



51. Focusing is done by passing DC through a specially shaped coil. The position of the coil, and the amount of current, control the focusing of the beam. The focus control is a rheostat, which varies the current through the coil.

52. Deflection is produced by two pairs of coils, mounted at 90º to each other, round the neck of the tube. In Fig 14, only one pair is shown for clarity. The amount of deflection depends on the strength of current in the coils. If the current is reversed, the direction of deflection reverses. To produce a horizontal time base, a sawtooth waveform of current must be passed through the horizontal deflection coils. Vertical deflection can be obtained by passing a signal current through the vertical deflection coils, but it is more usual to show signals on an electromagnetic CRT by using intensity modulation. In this case, a positive-going signal voltage is applied to the grid of the CRT. The number of electrons in the beam increases for the duration of the signal, and a bright spot appears on the time base.

The ESCRT and EMCRT Combined

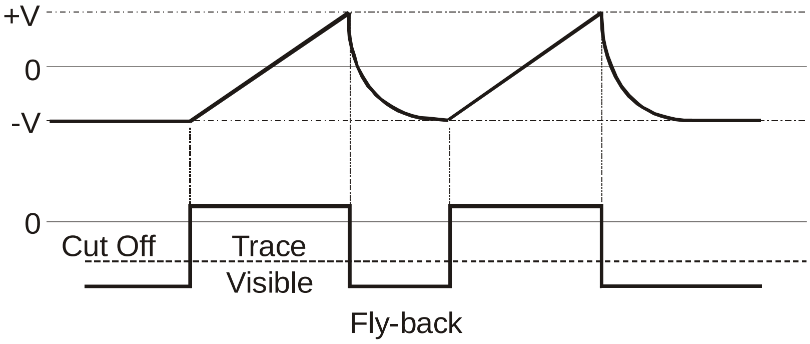
53. Combinations of the two CRT systems may be used, the most popular being that using electrostatic focusing and electromagnetic deflection. There are various advantages and disadvantages of the two types, and, until recently, the greatest advantage of the electromagnetic tube has been its higher resolution, ie ability to reduce the spot size to very small proportions. However, recent research has improved the resolution of the electrostatic CRT, although it has increased the complexity of the tube and necessitated the use of low-power magnetic correcting coils. Present-day electromagnetic tubes have spot sizes of about 0.25 mm.

CRT Displays

54. **Time Base Production**. All CRT displays need some form of time base; a large number of displays use the A-scope time base. With this type, the spot moves linearly across the face of the screen, usually from left to right, then returns rapidly to its starting point. During the rapid return, known as 'fly-back', the spot of light is normally extinguished by application of a suitable blackout waveform to the grid of the tube. The ESCRT uses a voltage waveform, whilst the EMCRT uses a current waveform, as follows:

a. **The ESCRT Waveform**. Fig 15 shows the voltage waveform required to produce a type A time base. The frequency of the waveform is equal to the pulse recurrence frequency of the radar. A circular time base, rotating with angular velocity ω, can be obtained by applying voltages proportional to sin ωt and cosine ωt to the X and Y plates respectively.

14-12 Fig 15 A-scope Waveform



b. **The EMCRT Waveform**. As the coils of the EMCRT are subject to inductance, current modulations must be used to produce the sawtooth waveform. A PPI display can be produced by physically rotating the coils.

55. **Calibration Markers**. For accurate measurement of range, the time base of a CRT must be set up against some accurate standard. This is done by a calibrator unit which is synchronized with the transmitter pulse. The calibration markers appear as pips or bright rings.

56. **Types of Display**. The CRT in a radar receiver can be used to present one, two or three-dimensional information concerning the target.

a. **One-dimensional Displays**. If only one-dimensional information is presented, the trace is deflection-modulated (ie the spot is deflected from its normal path to indicate the presence of an echo signal from a target). Examples of this type of display are the A-scope, I-scope, and J-scope.

b. **Two-dimensional Displays**. For two-dimensional displays, the trace is intensity-modulated (ie the brilliance of the trace is varied by an echo signal from the target). Examples of two-dimensional displays are:

(1) The PPI, sector-PPI, and B-scope, each of which shows the situation in a bearing and range plan format.

(2) The C-scope, which shows elevation and azimuth error, and is therefore of special use in fighters, where the display corresponds to the pilot’s view through the windscreen. The C-scope information can also be readily transferred to the pilot’s Head-up Display.

c. **Three-dimensional Displays**. Three-dimensional information is displayed in a variety of ways, often by combining complementary two-dimensional displays. For example, a C-scope, showing elevation and azimuth error, can be superimposed on to a B-scope showing range and azimuth. The I-scope shows target range, displacement from the axis, and angle-off error.

In all types of display, the spot moves in some pre-determined manner on the screen, this movement being termed the time base. Some of the commonly-used radar displays are illustrated in Volume 11, Chapter 1.

57. **Strobes**. In some radar equipments, it is necessary to expand a section of the main time base on either side of the echo blip in order to measure the range more accurately. To do this, a strobe pulse is generated at some definite time after the start of the main time base, the delay being controlled by the operator. The strobe pulse is made to trigger a strobe time-base circuit, with its own calibration on a larger scale.

CHAPTER 13 - POWER SUPPLIES

Introduction

1. In order to function, all electronic equipment needs to be energized by means of a power supply. In the great majority of cases, this power needs to be delivered at a steady or fixed voltage. In the early days of radio and electronics, power was derived from batteries, but with the advent of thermionic valves, which called for high currents and voltages, this source of power became inconvenient. Equipment power supplies were therefore developed which relied on the main domestic supply as the energy source. With the invention of the transistor, battery packs came back into favour, and remain so due to modern digital technology allowing for equipment portability and modest power requirements. For larger power requirements or fixed installations mains or AC generator-driven power packs are usually more economical. For remote locations, such as space satellites, the solar cell has been developed to convert solar energy into electrical power.

BATTERIES

The Simple Cell

2. When a metal electrode is immersed in an electrolyte, either:

a. Atoms of the electrodes of the more reactive metals (like zinc) are dissolved in the solution as positive ions, leaving the electrode negative with respect to the electrolyte, or

b. Positive ions from the electrolyte are attracted to the electrode of less reactive metals (like copper) making it positive with respect to the electrolyte.

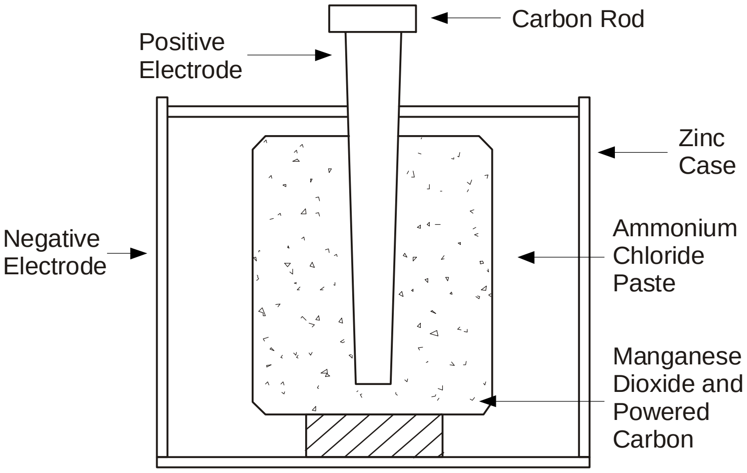
3. If a zinc rod and a copper rod are placed in dilute acid and are joined by a wire, then the positive charge on the copper will tend to flow along the wire and neutralize the negative charge on the zinc. This will cause more zinc to dissolve in the acid and more hydrogen to be liberated at the copper so as to maintain the potential difference between the electrodes. Current will flow until all the zinc is dissolved away or until the acid is exhausted.

4. One disadvantage of the simple cell is that hydrogen bubbles being liberated at the copper electrode tend to insulate it from the solution and stop the chemical action. This effect is called polarization. Another disadvantage is that of 'local action'. Impurities present in the zinc form small local cells causing the zinc to be dissolved away without supplying power to the main circuit.

The Dry Cell

5. Fig 1 depicts an early dry cell, also known as a 'Leclanché cell'. The positive electrode is the carbon rod and the negative electrode is the zinc case. The electrode is ammonium chloride, in paste form, and is separated from the carbon rod by a depolarizing agent. This is a mixture of powdered manganese dioxide and carbon, which oxidizes the hydrogen molecules to water.

14-13 Fig 1 An Early Dry Cell

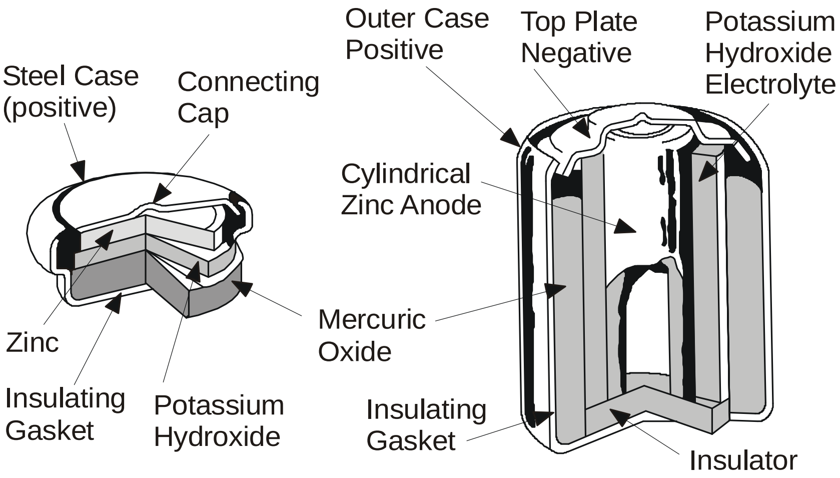


6. Similar materials were used in a later version of the Leclanché with the addition of mercuric chloride to reduce local action, and potassium dichromate to inhibit the corrosion of zinc. A disadvantage of this type of cell is that current quickly falls off, principally because hydrogen forms more quickly than it can be removed by the depolarizer. For this reason, they are best suited for intermittent work.

Mercury Cells

7. The mercury cell, the full name of which is zinc-mercuric oxide, was developed during the Second World War for use in portable equipment where maximum energy with minimum volume was the prime aim. The modern day mercury cell, shown at Fig 2, is capable of providing a steady voltage over nearly all of the cell's useful discharge period. Over long periods of operational use, or long storage, a voltage regulation within one per cent of the initial voltage is still maintained.

14-13 Fig 2 Mercury Cells (Flat and Cylindrical)



8. In the mercury cell, the negative electrode is zinc, as in the Leclanché cell, but the positive electrode and its depolarizer consist of graphite and mercuric oxide. The electrolyte consists of potassium hydroxide.

Solar Cells

9. The necessity for continuous electric power generation on space satellites led to the development of the solar cell. These devices, which are generally made from thin slices of highly pure single crystal silicon, produce electric power from radiant energy. Wavelengths in the range 400 to 1,100 nanometers are the most efficient, and about half the solar spectrum falls in this range.

10. Solar cells have no storage capacity, and for terrestrial applications they are used in conjunction with electrical storage batteries.

Non-rechargeable Batteries

11. Considerable effort and research has been put into rechargeable batteries over recent years but non-rechargeable batteries continue to fill an important niche market in applications such as wristwatches, remote controls and electronic keys. Non-rechargeable batteries are useful when charging is impractical or impossible, such as in some military applications. They possess high specific energy, can be stored for protracted periods and be ready for immediate use. Most non-rechargeable batteries are relatively inexpensive to produce and environmentally friendly.

12. Carbon-zinc general purpose batteries are used for applications with low power drain such as remote controls and torches. One of the most common non-rechargeable batteries is the *alkaline-manganese*, or alkaline battery. Alkaline batteries deliver more energy at higher load currents than carbon-zinc and do not leak when depleted, as carbon-zinc does.

13. Non-rechargeable batteries have one of the highest energy densities. Although rechargeable batteries have improved, a regular household alkaline battery provides 50 percent more energy than a lithium-ion one. The most energy-dense non-rechargeable type is the lithium battery made for film cameras and military combat and holds over three times the energy of lithium-ion.

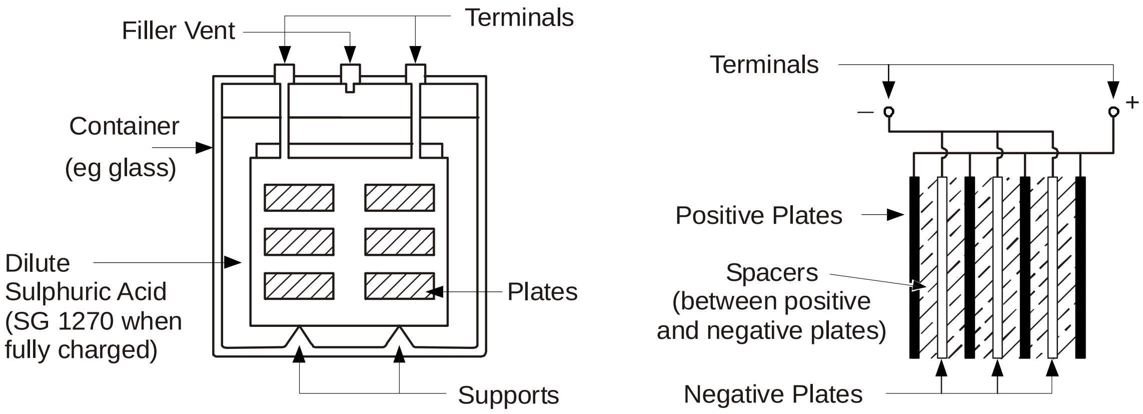
Rechargeable Batteries

Lead Acid Batteries

14. A lead-acid battery is a device for storing electricity. It produces power by means of chemical reaction. Unlike the batteries described so far, this reaction is reversible; so that when the battery has discharged power, it can be recharged by passing an electric current through it.

15. In its simplest form, shown in Fig 3, the lead-acid battery consists of two groups of interleaved plates, one group of lead and the other group of lead dioxide, immersed in a weak solution of sulphuric acid. Connecting the two groups of plates together electrically causes a chemical reaction to take place and current flows between the plates. This reaction continues until both plates change to lead sulphate, and the acid turns to water.

14-13 Fig 3 The Lead-acid Battery



16. The battery is recharged by passing an electric current (from a generator) through it in the other direction. This reverses all the electrical forces, and all the chemical reactions reverse themselves as a result. The lead sulphate changes back to lead oxide in the positive plates, and lead in the negative plates. The water changes back to sulphuric acid.

17. During the mid 1970s, researchers developed a maintenance-free lead acid battery that could operate in any position. The liquid electrolyte was transformed into moistened separators and the enclosure was sealed. Safety valves allow the venting of gas generated during charge and discharge.

18. Driven by different applications, two battery designations emerged. They are the small sealed lead acid (SLA), and the large valve regulated lead acid (VRLA). Technically, both batteries are the same. The term SLA is a misnomer as no lead acid battery can be totally sealed. As a result, the batteries are fitted with a valve to control the venting of gases during stressful charge and rapid discharge. Unlike the conventional lead acid battery where the plates are submerged in a liquid, the electrolyte is impregnated into a moistened separator. The design resembles nickel and lithium based systems and enables the battery to be operated in any physical orientation without leakage.

19. Both the SLA and VRLA are designed with a low over-voltage potential to prohibit the battery from reaching its gas-generating potential during charge. Excess charging would cause gassing and water depletion and consequently, these batteries should never be charged to their full potential.

20. Leaving the lead acid battery on float charge for a prolonged time does not cause damage. The battery's charge retention is the best among rechargeable batteries, for example, a Nickel-cadmium battery (NiCd) self-discharges approximately 40 percent of its stored energy in three months, the SLA self-discharges the same amount in one year. The SLA is relatively inexpensive but the operational costs can be higher than the NiCd if full cycles are required on a repetitive basis.

21. The SLA not suitable for fast charging with typical charge times of 8 to 16 hours. The SLA must always be stored in a charged state as leaving the battery in a discharged condition causes sulphation, a condition that makes the battery difficult, if not impossible, to recharge.

22. Unlike the NiCd, the SLA is not suitable for deep cycling. A full discharge causes extra strain and each cycle robs the battery of a small amount of capacity. This loss is small while the battery is in good operating condition, but the fading increases once the performance drops to half the nominal capacity. This wear-down characteristic also applies to other battery chemistries in varying degrees.

23. Depending on the depth of discharge and operating temperature, the SLA provides 200 to 300 discharge/ charge cycles. The primary reason for its relatively short cycle life is grid corrosion of the positive electrode, depletion of the active material and expansion of the positive plates. These changes are most prevalent at elevated operating temperatures and high-current discharges.

24. The optimum operating temperature for the SLA and VRLA battery is 25 °C. As a rule of thumb, every 8 °C rise in temperature will cut the battery life in half. VRLA that would last for 10 years at 25 °C will only be good for 5 years if operated at 33 °C. The same battery would endure a little more than one year at a temperature of 42°C.

25. Among modern rechargeable batteries, the lead acid battery family has the lowest energy density, making it unsuitable for handheld devices that demand compact size. They work well at cold temperatures and are superior to lithium-ion when operating in subzero conditions.

Table 1 The Advantages and Limitations of Lead Acid Batteries

Advantages	Limitations
Inexpensive and simple to manufacture	Should not be stored in a discharged condition
Mature, reliable and well-understood technology	Low energy density
Durable and provides dependable service	Limited number of full discharge cycles
Low self-discharge	Environmentally unfriendly
Low maintenance requirements	Transportation restrictions on flooded lead acid
Capable of high discharge rates	Thermal runaway possible with improper charging
Good low and high temperature performance	

26. **Absorbent Glass Mat (AGM) Batteries**. AGM technology became popular in the early 1980s as a sealed lead acid battery for military aircraft, vehicles and uninterrupted power supplies to reduce weight and improve reliability. The acid is absorbed by a very fine fibreglass mat making the battery spill-proof. The plates can be made flat to resemble a standard flooded lead acid pack in a rectangular case or wound into a cylindrical cell. AGM has very low internal resistance, is capable of delivering high currents on demand and offers a relatively long service life, even when deep-cycled. AGM is maintenance free, provides good electrical reliability and is lighter than the flooded lead acid type. It stands up well to low temperatures and has a low self-discharge. The leading advantages are a charge that is up to five times faster than the flooded version, and the ability to deep cycle. As with all gelled and sealed units, AGM batteries are sensitive to overcharging.

Table 2 The Advantages and Limitations of AGM Batteries

Advantages	Limitations
Spill-proof	Higher manufacturing cost
High specific power, low internal resistance, responsive to load	Sensitive to overcharging
Up to 5 times faster charge than with flooded technology	Capacity has gradual decline
Better cycle life than with flooded systems	Low specific energy
Vibration resistance	Must be stored in charged condition
Stands up well to cold temperature	Not environmentally friendly

Nickel Based Batteries

27. **Nickel-cadmium Cells**. Rechargeable nickel-cadmium cells are useful in electronic equipment since they can be sealed, thus avoiding the effect of corrosive fumes which are given off by lead-acid accumulators. The sealed type of cell has a long life and can be completely discharged without ill effects.

Table 3 The Advantages and Limitations of NiCd Batteries

Advantages	Limitations
Fast and simple charging after prolonged storage	Relatively low specific energy compared with newer systems
High number of charge/discharge cycles (over 1000 with proper care and maintenance)	Memory effect; needs periodic full discharges
Good load performance; rugged and forgiving if abused	Environmentally unfriendly; cadmium is toxic
Long shelf life; can be stored in a discharged state	High self-discharge; needs recharging after storage
Simple storage and transportation	
Good low-temperature performance	
Economical	
Wide range of sizes and performance options	

28. **Nickel-metal-hydride (NiMH)**. NiMH provides 40 percent higher specific energy than a standard NiCd and does not contain toxic metals. NiMH also has a high self-discharge rate and loses about 20 percent of its capacity within the first 24 hours, and 10 percent per month thereafter. The NiMH battery in a torch in storage will be ‘flat’ after only a few weeks.

9-12 Table 1 The Advantages and Limitations of NiMH Batteries

Advantages	Limitations
30 % to 40 % higher capacity than a standard NiCd	Limited service life; deep discharge reduces service life
Less prone to memory effect than NiCd	Requires complex charge algorithm
Simple storage and transportation	Does not absorb overcharge well; trickle charge must be kept low
Environmentally friendly	Generates heat during fast-charge and high-load discharge
Nickel content makes recycling profitable	High self-discharge
	Performance degrades if stored at elevated temperatures; should be stored in a cool place at about 40 % charge

Lithium Based Batteries

29. **Lithium (Li) Batteries**. There are two types of lithium battery. Lithium-ion batteries are rechargeable and used in equipment such as laptops, tablets and smart phones. Lithium-metal batteries are non-rechargeable and are found in items such as watches. Although lithium batteries are safe, their high energy levels mean they can pose a fire risk if damaged or charged incorrectly.

Rechargeable batteries with lithium metal on the anode could provide high energy densities; however, cycling produces unwanted growth particles (dendrites) on the anode which can penetrate the separator and cause an electrical short. When this occurs, the cell temperature rises quickly and approaches the melting point of lithium causing thermal runaway, also known as "venting with flame". As a result Li batteries must be treated with care and stowed appropriately during flight. Lithium batteries are known to have been the cause of a number of fires onboard aircraft and during ground handling. Additionally, poor quality or counterfeit batteries, which are known to be in wide circulation, pose an additional safety risk.

30. **Lithium Ion (Li-ion) Batteries**. The inherent instability of lithium metal, especially during charging has resulted in the production of a non-metallic solution using *lithium ions*. Although lower in specific energy than lithium-metal Li-ion is safe provided manufacturers and battery packers follow the correct safety measures in limiting voltages and currents to secure levels. The specific energy of a Li-ion battery is twice that of NiCd and it is also low-maintenance. The battery has no memory and does not need exercising (deliberate full discharge). Self-discharge is less than half that of nickel-based systems. The disadvantages of the LI-ion battery are the need for protection circuits to prevent abuse and high cost.

31. Similar to lead and nickel based architectures, lithium-ion uses a cathode (positive electrode) made from a metal oxide, an anode (negative electrode) made of porous carbon and electrolyte as the conductor. During discharge, the ions flow from the anode to the cathode through the electrolyte and separator. Charging reverses the direction and the ions flow from the cathode to the anode. Li-ion batteries come in many varieties which vary in performance where the choice of cathode material determines the unique characteristics of the battery. Most manufacturers use graphite as the anode to attain a flatter discharge curve. Graphite stores lithium-ions effectively when the battery is charged and has long-term cycle stability.

Table 4 The Advantages and Limitations of Li-ion Batteries

Advantages	Limitations
High energy density	Requires protection circuit to limit voltage and current
Relatively low self-discharge; less than half that of NiCd and NiMH	Subject to aging, even if not in use
Low maintenance	Transportation regulations when shipping in larger quantities
No periodic discharge is needed	
No memory	

Table 5 Six Common Lithium-ion Batteries

Li-ion Battery	Characteristics
Lithium Cobalt Oxide($LiCoO_2$)	High specific energy / Relatively short life span and limited load capabilities (specific power).
Lithium Manganese Oxide ($LiMn_2O_4$)	Low internal cell resistance - fast charging and high-current discharging / Capacity that is roughly one-third lower compared to Li-cobalt.
Lithium Iron Phosphate($LiFePO_4$)	Enhanced safety, good thermal stability, tolerant to abuse, high current rating and long cycle life / cold temperature reduces performance, and elevated storage temperature shortens the service life.
Lithium Nickel Manganese Cobalt Oxide ($LiNiMnCoO_2$)	Can also be tailored to high specific energy or high specific power, but not both.
Lithium Nickel Cobalt Aluminium Oxide ($LiNiCoAlO_2$)	High specific energy and power densities / long life span.
Lithium Titanate ($Li_4Ti_5O_{12}$)	Li-titanate replaces the graphite in the anode / can be fast-charged and delivers a high discharge current / High cycle rate / Safe / Excellent low-temperature discharge characteristics.

Lithium-polymer Battery

32. Lithium-polymer differs from other battery systems in the type of electrolyte used. The original polymer design used a solid (dry) polymer electrolyte that resembles a plastic-like film. This insulator allows the exchange of ions and replaces the traditional porous separator that is soaked with electrolyte. A solid polymer has a poor conductivity at room temperature and the battery must be heated to between 50 and 60 °C (122 to 140 °F) to enable the current to flow.

33. To make the Li-polymer battery conductive at room temperature, gelled electrolyte is added. All Li-ion polymer cells today incorporate a micro porous separator with moisture. The correct term is “Lithium-ion polymer” (Li-ion polymer or Li-polymer for short). Li-polymer can be built on many systems, such as Li-cobalt, Li-phosphate and Li-manganese. For this reason, Li-polymer is not considered a unique battery chemistry.

34. Li-polymer cells may also be shaped into a flexible foil-type case (polymer laminate or pouch cell) that resembles a food package. While a standard Li-ion needs a rigid case to press the electrodes together, Li-polymer uses laminated sheets that do not need compression. Although less durable that conventional packaging, a foil-type enclosure reduces the weight by more than 20 percent over the classic hard shell and thin film technology liberates the format design and the battery can be made into any shape.

35. Charge and discharge characteristics of Li-polymer are identical to other Li-ion systems and do not require a special charger. Safety issues are also similar in that protection circuits are needed. Gas build-up during charge can cause some Li-polymer in a foil package to swell. Li-polymer is not limited to a foil package shape and can also be made into a cylindrical design.

36. **Li-cobalt**. Most Li-polymer packs for the consumer market are based on Li-cobalt. With gelled electrolyte added the lithium polymer is essentially the same as the lithium-ion battery. Both use

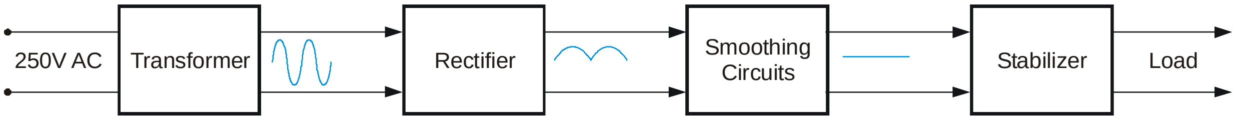
identical cathode and anode material and contain a similar amount of electrolyte. Although the characteristics and performance of the two systems are alike, the Li-polymer is unique in that a micro porous electrolyte replaces the traditional porous separator. The gelled electrolyte becomes the catalyst that enhances the electrical conductivity. Li-polymer offers slightly higher specific energy and can be made thinner than conventional Li-ion.

DC POWER FROM AC SOURCES

Overview of Process

37. When comparatively large amounts of power are needed, the sources of supply are nearly always alternating currents. However, before the power can be used inside a piece of electronic equipment it needs to be converted to a DC of the correct magnitude. The component parts of this process are shown in Fig 4.

14-13 Fig 4 Component Parts of a Power Supply



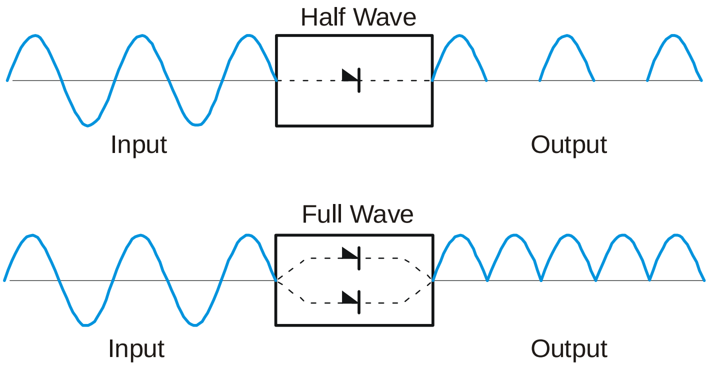
Transformation

38. In most cases, the amplitude of the AC voltage delivered to a piece of electronic equipment is of the incorrect value and needs to be transformed, up or down, to suit the equipment needs. This transformation is accomplished by means of a transformer. The subject of transformers is covered in some detail in Volume 14, Chapter 11.

Rectification

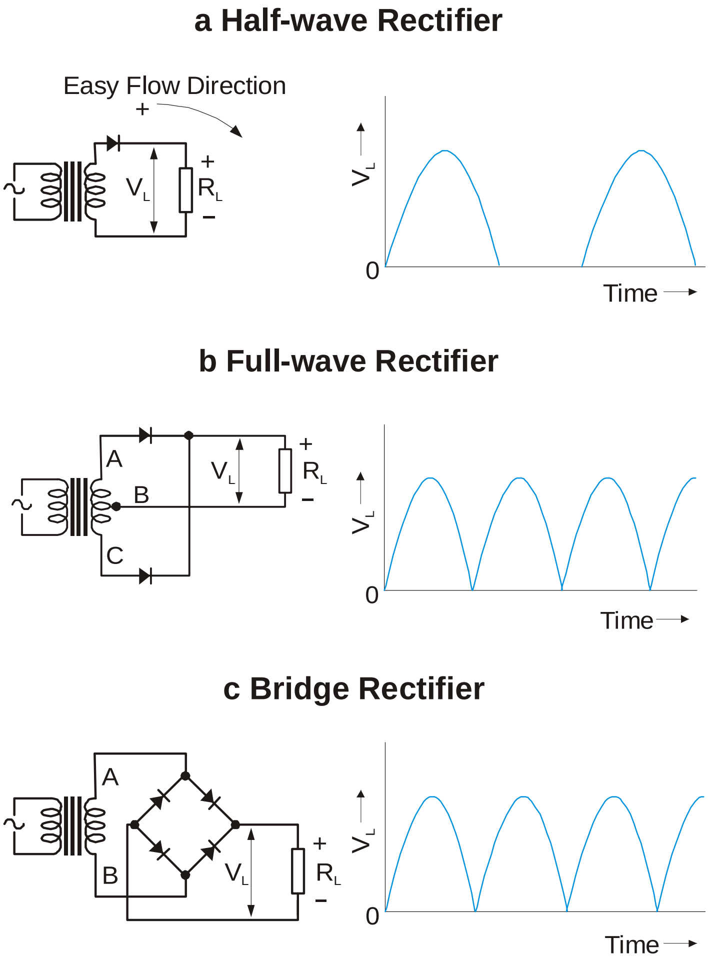
39. Once the AC voltage has been transformed to a level that is suitable for use in the equipment, the first stage of converting AC into a DC takes place. This stage is known as rectification, and involves the production of a pulsating unidirectional output from an alternating input. The process can be either half-wave rectification or full-wave rectification (see Fig 5).

14-13 Fig 5 Rectification



40. **Half-wave Rectification**. The diode (thermionic or semi-conductor) is the fundamental component in all rectification circuits. As the diode conducts only when the anode is positive with respect to the cathode, pulses of current flow through the load on alternate half-cycles of the voltage waveform. The output of a half-wave rectifier is shown in Fig 6a.

14-13 Fig 6 Types of Rectification



a Half-wave Rectifier

b Full-wave Rectifier

c Bridge Rectifier

41. **Full-wave Rectification**. By using two diodes, the full applied waveform can be used to produce an output, as shown in Fig 6b. The anodes of the two diodes are connected to opposite ends of a centre-tapped transformer winding. The voltages at the diode anodes are thus in anti-phase, so that the diodes conduct on alternate half-cycles of the applied waveform. The current through the load is always in the same direction, regardless of which diode is conducting, so producing a unidirectional flow of pulses through the load on every half-cycle. Note that a double diode could be used instead of two separate diodes.

421. **Bridge Rectification**. An alternative form of a full-wave rectifier is the bridge rectifier. This does away with the need for a centre-tapped transformer, but uses four diodes (thermionic or semi-conductor) arranged like a Wheatstone Bridge, with the supply connected across one diagonal of the bridge and the load across the other. The diodes act in pairs and, as before, the current through the load is always in the same direction (Fig 6c).

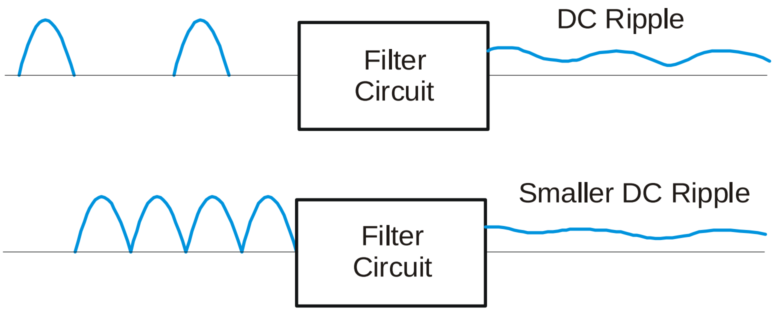
43. **P-N Junction Diode as a Rectifier**. Since a junction diode passes current only when the voltage across it is of one polarity and allows almost none to flow when the polarity is reversed, it may be used as a rectifier.

44. **Practical Junction Diode Rectifiers**. Junction diodes are manufactured in a wide range of sizes and cases. In general, the more power a diode has to handle, the bigger it will be. If the diode is passing a large current, heat is developed because of the resistance of the diode. This heat must be removed if the diode is not to be damaged, and the easiest way to do this is to provide large surface areas exposed to the air. Very often special 'heat sinks' are provided to give quick removal of heat from the diode.

Filters or Smoothing Circuits

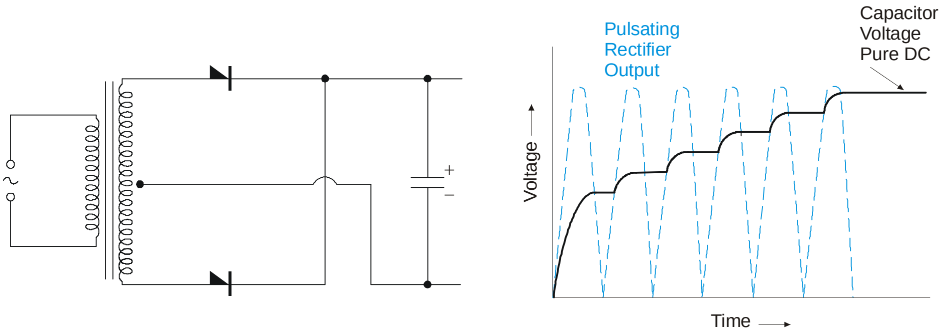
45. The fluctuating output from the rectification stage cannot be regarded as a unidirectional supply suitable for equipment use; the AC component needs to be minimized. This is achieved by a filter, or smoothing circuit and the action is illustrated in Fig 7.

14-13 Fig 7 Filter/Smoothing Action

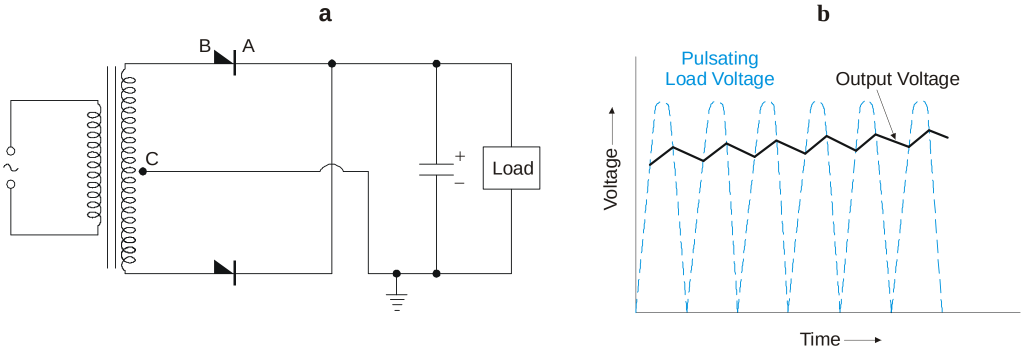


46. **Reservoir Capacitor**. If a large value capacitor is placed across the output terminals of a rectifier, and the load is disconnected, a steady DC voltage is developed across the capacitor. A rectifier has resistance, due mainly to the internal resistance of the rectifying device, therefore, when a capacitor is placed across the output of a rectifier, it charges up every time the rectifier conducts. If it does not reach the full output voltage at the end of the first half-cycle, it will do so during the next few half-cycles. It cannot discharge between the peaks of the pulsating DC because the rectifier will not pass current in the reverse direction. Therefore, after a few half-cycles, a steady DC voltage appears across the capacitor, as shown in Fig 8, and the AC has been filtered out. If a load is connected across the capacitor (Fig 9a), the capacitor begins to discharge and its voltage will fall. However, in a full-wave rectifier a new voltage peak occurs 100 times per second and this voltage peak recharges the capacitor. The resulting output waveform is shown in Fig 9b. The DC is not quite steady, but the amplitude of the pulses is much smaller than if there were no capacitor at all. The voltage waveform is now DC plus a small component of AC called the 'ripple'. The voltage is not smooth enough for the final HT output and more components are needed in the smoothing circuit.

14-13 Fig 8 Rectifier Output with Capacitor



14-13 Fig 9 Rectifier Output with Capacitor and Load



47. **Smoothing Circuit Resistor**. The ripple can be reduced by adding a resistor and capacitor, which will act as an AC voltage divider, to the smoothing circuit. It is common for two resistors to be connected across the capacitor. This would normally give constant voltage to within 1 to 2% of the required HT supply. If this is not satisfactory, another RC section could be added.

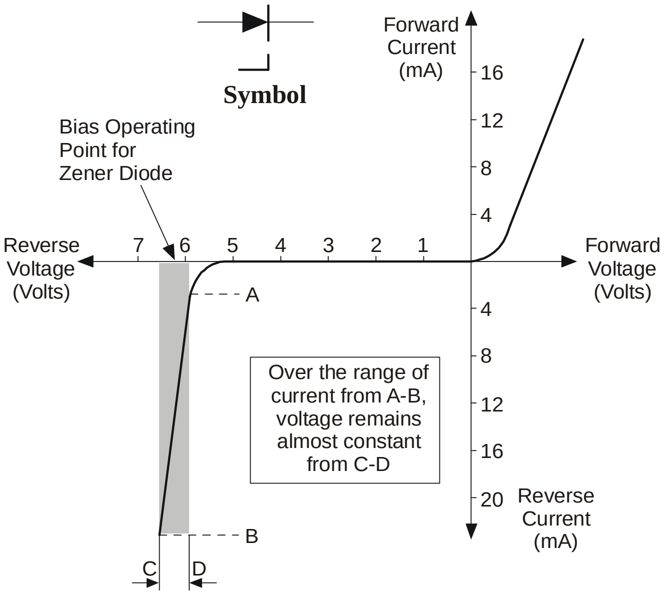
Stabilization

48. If it is required to ensure that the DC supply from the smoothing circuits remains constant, regardless of the applied load, compensation is needed. This compensation is generally known as stabilization.

49. Although there are a number of such circuits, they are usually based around the cold-cathode, gas-filled diode for HT circuits and the zener diode for low voltage circuits.

50. **Zener Diode**. Most junction diodes will be permanently damaged by the high reverse current flowing if the breakdown voltage is exceeded. The zener diode is specially designed to make use of this part of the diode characteristic, and is not damaged by the high reverse current. If the zener diode is reverse biased to a value just beyond the breakdown voltage, as shown in Fig 10, the voltage across the diode remains practically constant for large variations of current through it. Therefore, the zener diode may be used to provide a constant reference voltage point in a circuit, or as a means of stabilizing the voltage over a wide range of circuit currents. A whole range of zener diodes with different values of breakdown (stabilizing) voltage is available.

14-13 Fig 10 Zener Diode



CHAPTER 14 - AMPLIFIERS

Introduction

1. Amplification is needed in a system in order to boost weak signals to a level where they can be of some practical use. In the case of a radio receiver, signals at the aerial, of the order of a few microwatts, require amplification before they can be used to drive a speaker system demanding several watts.

2. Amplifiers are classified according to the function they perform. There are four main types of amplifier, each designed to amplify signals in a specific part of the frequency spectrum:

a. **Audio Amplifiers**. These amplify a band of frequencies in the audio range, ie between about 20 and 20,000 Hz. Audio amplifiers are used in radio and television receivers and transmitters, in intercom equipment, CDs, cassette tape decks, and in many other electronic devices.

b. **Video Amplifiers**. These are similar to audio amplifiers except that they cover a much wider frequency band, ranging from about 20 Hz to 6 MHz. Whereas audio amplifiers deal with electrical waveforms corresponding to sounds, video amplifiers handle electrical waveforms corresponding to pictures. They are used in television and in radar apparatus.

c. **RF Amplifiers**. These amplify only a narrow band of frequencies, so narrow that they may be considered to amplify at one frequency only. This frequency may be anywhere in the RF band, ie from 20 kHz to 100 GHz. They are used in radio and television transmitters and receivers, in radar and in weapon guidance systems.

d. **DC Amplifiers**. These direct-coupled amplifiers are used to amplify inputs at frequencies of a few hertz and slow non-regular changes in voltage. They are not used to amplify steady non-varying DC, as this would be pointless. They are used in many types of electronic equipment, particularly in analogue computers.

3. These main categories for amplifiers may be broken down further into the sub-categories of wideband, narrow band, tuned, stagger tuned, voltage and power. Thus, for example, a particular amplifier may be referred to as an AF stagger tuned voltage amplifier.

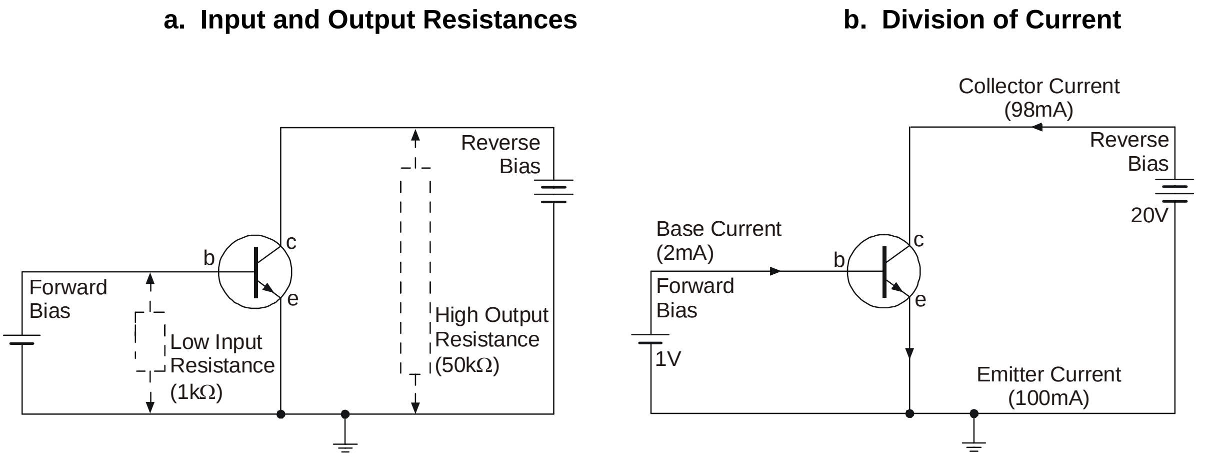
The Transistor as an Amplifier

4. In a properly biased transistor the resistance of the forward biased emitter-base (input) circuit is low whilst that of the reverse biased collector (output) circuit is high. Typical values for the arrangement shown in Fig 1 are one kΩ and 50 kΩ respectively.

5. The forward and reverse-bias voltages are such that certain steady values of current flow in the emitter, base and collector circuits. Ninety-eight percent of the electrons from the emitter reach the collector; therefore if the emitter current is l00 mA, the collector current will be about 98 mA, and the base current about 2 mA as shown in Fig1b. Furthermore, the ratios of these currents remain constant so that if one current changes the other two also change. For example, using the ratios given above, if the emitter current changes by 1 mA, the collector current will change by 0.98 mA and the base current by 0.02 mA. Similarly, if the base current is made to change by 0.02 mA, the collector current will

change by 0.98 mA and the emitter current by 1 mA. In an amplifier, it is these changes in currents and voltages which generate interest.

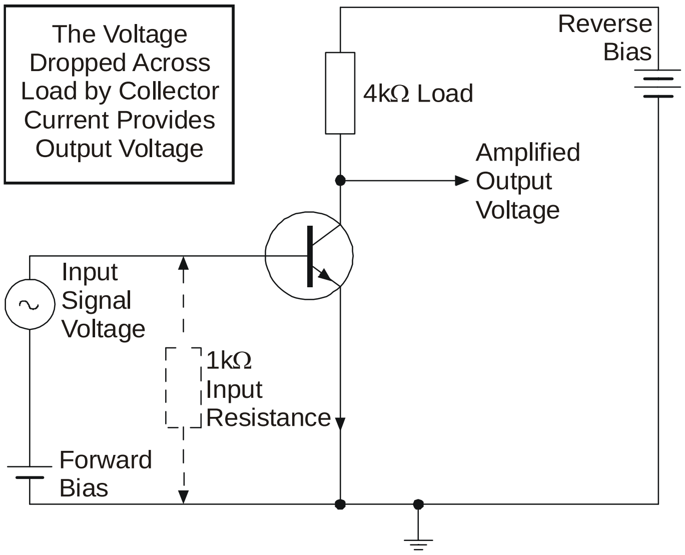
14-14 Fig 1 Transistor Principles



6. If a very low resistance alternating voltage source in the emitter-base circuit is connected in series with the forward bias battery, the voltage applied to the input circuit can be varied. A 4 kΩ load resistor in the collector circuit is also connected in series with the reverse bias battery as shown in Fig 2. The alternating voltage source represents an input signal voltage, eg from a pick-up on a record player. The output voltage is taken between the collector and earth.

7. If the input signal voltage in the circuit of Fig 2 is changed from zero to 0.02 V, the base current flowing in the 1 kΩ input resistance will change by 0.02 mA (ie 0.02V/1 kΩ). This causes the collector current to change by 0.98 mA.

14-14 Fig 2 Transistor Amplifier Circuit



Gain

8. Technically, gain is defined as the increase in power level in the load, ie the ratio of the actual power delivered to that which would be delivered if the source were correctly matched, without loss, to

the load, in the absence of an amplifier. Put more simply, gain in an amplifier is the ratio of output quantity to input quantity. Therefore, using the previous paragraph figures:

$$\text{Current gain} = \frac{\text{Output (collector) current change}}{\text{Input (base) current change}}$$

$$= \frac{0.98\ \text{mA}}{0.02\ \text{mA}} = 49$$

9. The change in output current flowing in the 4 kΩ load resistor produces a change of voltage across the load. This change of voltage represents the output signal voltage, given by IR = 0.98 mA × 4 kΩ = 3.92 V. Therefore:

$$\text{Voltage gain} = \frac{\text{Output voltage change}}{\text{Input voltage change}} = \frac{3.92\ \text{V}}{0.02\ \text{V}} = 196$$

10. There is also a power gain given by:

Current gain × voltage gain = 49 × 196 = 9,604

11. All the output quantities are greater than the input quantities and amplification has been achieved. The large voltage and power gains arise from the fact that the transistor transfers current from a low resistance input circuit to a high resistance output circuit.

Transistor Amplifier Action

12. The action of the circuit of Fig 2 is summarized in Fig 3:

a. The increase in input signal voltage increases the forward bias at the emitter-base junction (battery and signal voltages add).

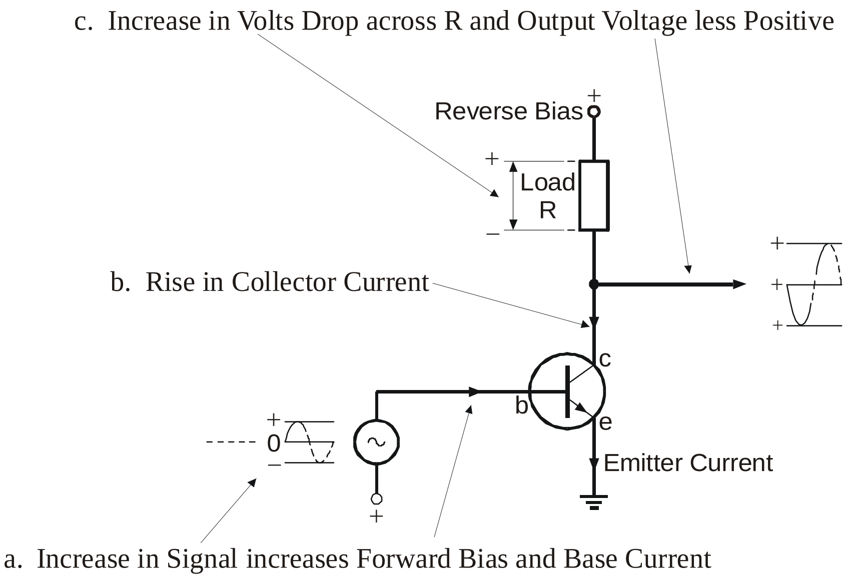
b. The increase in forward bias increases the base current and this causes the collector current to rise also - much more than the base current.

c. The rise in collector current increases the voltage drop across the load resistor and this causes the output (collector-to-earth) voltage to become smaller (less positive).

d. Note the phase relationship between input and output signals. An increase in input signal voltage causes a fall in output voltage. Thus, in this circuit, input and output voltages are 180º out of phase.

Similar results would have been obtained if the above had been applied to a p-n-p transistor amplifier.

14-14 Fig 3 Transistor Amplifier Action

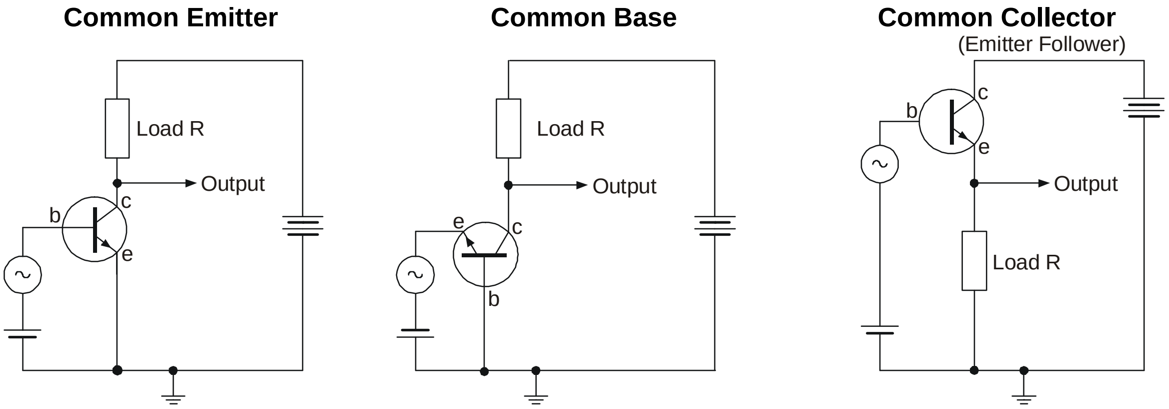

c. Increase in Volts Drop across R and Output Voltage less Positive
Reverse Bias
+
Load R
b. Rise in Collector Current
c
b
e
0
Emitter Current
a. Increase in Signal increases Forward Bias and Base Current

13. The transistor amplifier action has also been covered in some detail, as it is imperative that the principles are understood.

Transistor Circuit Arrangements

14. In the transistor amplifier circuits so far discussed, the emitter is common to both input and output circuits. This arrangement is called a common emitter or grounded emitter amplifier. Although the common emitter amplifier is the arrangement used most often, there are other ways of connecting the input and output circuits to the transistor. In practice, three circuit arrangements are in use as shown in Fig 4.

14-14 Fig 4 Three Circuit Arrangements



15. When dealing with the common emitter circuit several important factors are noted:

a. The input resistance is low and the output resistance high.

b. The current, voltage and power gains are all high.

c. The input and output signals are 180° out of phase.

16. If this procedure is repeated for the other two circuits, several points of difference will be noted. Table 1 compares the important factors for each of the three circuits. Each arrangement is used only for those jobs to which it is suited.

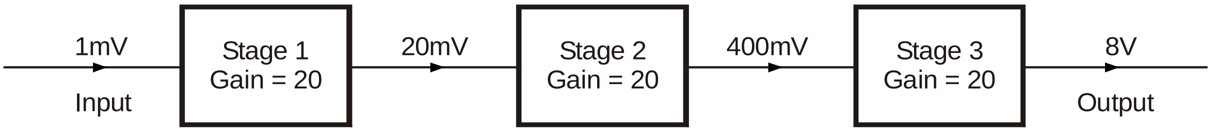
Table 1 Transistor Circuit Comparisons

	Common Emitter	**Common Base**	**Common Collector**
Input Resistance	Low (1 kΩ)	Very low (100 Ω)	High (100 kΩ)
Output Resistance	High (50 kΩ)	Very high (250 kΩ)	Low (200 Ω)
Current Gain	High (50)	Low (less than 1)	High (50)
Voltage Gain	High (50-1,000)	High (50-1,000)	Low (less than 1)
Power Gain	High (100-10,000)	Moderate (10-1,000)	Low (10-100)
Phase Relationship	180º Out of phase	In phase	In phase

Amplifiers in Cascade

17. So far the amplifying device has been considered as a single amplifier stage. In some cases the amplification afforded by one stage is sufficient, but more often several stages are needed, one following after the other, to obtain the required amplification (Fig 5). Therefore, if in a three-stage amplifier each stage has a gain of 20 and the input to the first stage is 1 mV, the output from this stage is 20 mV. This is applied as the input to the second stage which gives an output of 400 mV. This becomes the input to the third stage which provides an output of 8 volts. The overall amplification of the three stages is 20 × 20 × 20 = 8,000. The three amplifier stages are said to be connected in cascade.

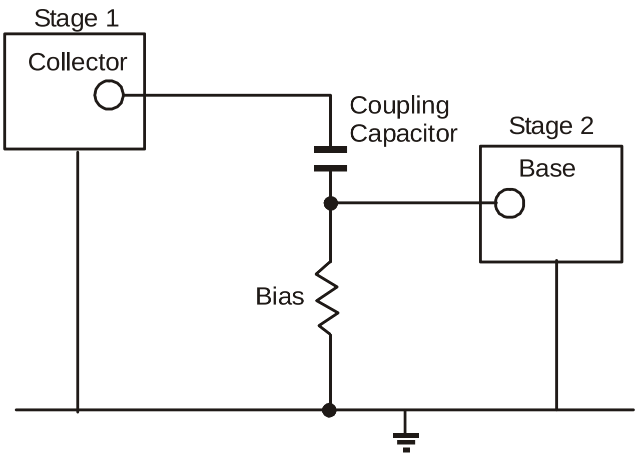
14-14 Fig 5 Cascade-connected Amplifiers



Resistance Capacitance (RC) Coupling

18. In order to couple two amplifier stages together some form of connection between the output of the first stage and the input to the next stage is required. A direct connection between the collector of stage one and the base of stage two, would not do, since the collector DC voltage would be applied to the base and would upset the biasing arrangements. In order to block this DC and still allow the AF variations of collector voltage to be applied to the input of the next stage, a coupling capacitor is needed (Fig 6). This capacitor and the bias, form a potential divider across the output of stage one. All the DC voltage is dropped across the capacitor and most of the AF voltage is developed across the resistor and applied to the input of the next stage. This is a common method of coupling the stages of an AF amplifier.

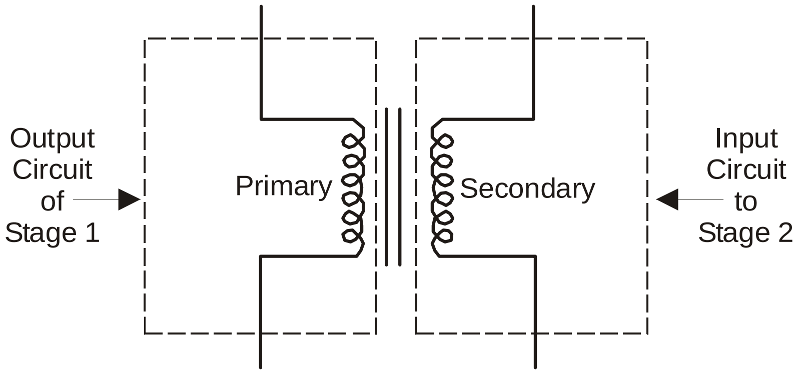
14-14 Fig 6 Coupling Network



Transformer Coupling

19. Another method of coupling the stages of an AF amplifier is by using transformers. The primary winding of the AF transformer forms the output load of the first stage and the secondary winding forms the input circuit of the next stage (Fig 7). No coupling capacitor is required.

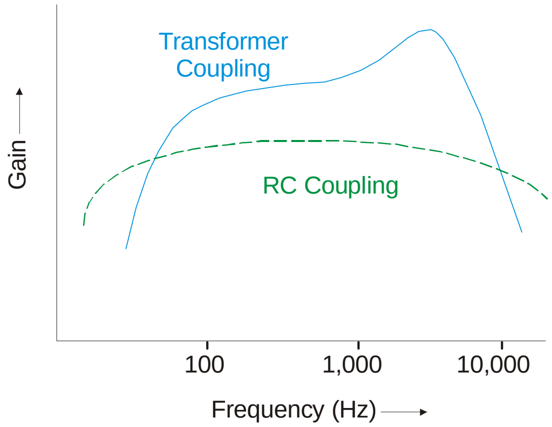
14-14 Fig 7 Transformer Inter-stage Coupling



20. With transformer coupling, a higher stage gain can be obtained. In a transistor amplifier, by choosing the correct turns ratio, the high output impedance of stage one can be matched to the low input impedance of stage two thus giving a considerable increase in gain over RC-coupled stages.

21. The main disadvantage of transformer coupling is the poor frequency response compared with an RC-coupled amplifier. The impedance of the windings varies with frequency; thus at the bottom end of the frequency range the impedance of the primary winding is low and so the voltage developed across it is low. At the high end of the frequency range the impedance is high and so these frequencies receive greater amplification than the middle frequencies. The response drops sharply at the top end of the frequency range due to the shunting effect of stray capacitances. The frequency response curves of transformer and RC-coupled amplifiers are compared in Fig 8. The uneven response of transformer coupling results in distortion.

14-14 Fig 8 Comparison Between Transformer and RC Couplings

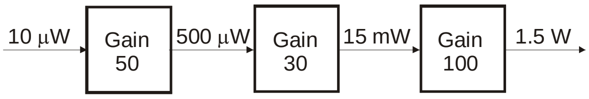


22. For this reason, and because of the size and weight of iron-cored AF transformers, RC inter-stage coupling is preferred. Because they provide a simple way of matching a high impedance to a low impedance, transformers are often used between the output stage and a loudspeaker load.

Decibels

23. The gain of an amplifier (see para 8) is the output divided by the input. We have so far expressed the gain as a number without units and as can be seen from Fig 9 in a multi-stage amplifier this number can become quite large.

14-14 Fig 9 Gain in a Multi-stage Amplifier



24. Radio amplifiers commonly have a power gain of several million and the overall gain figure becomes very cumbersome. Thus it is more convenient to express the power gain of an amplifier as the logarithm, to the base 10, of the output to input power ratio. The basic unit for this measurement is the bel. Therefore:

$$\text{Power gain} = \log_{10}\frac{\text{Power}_{\text{out}}}{\text{Power}_{\text{in}}}\ \text{bels}$$

In practice the bel is a very large unit, so work is done in decibels (dB), a decibel being one-tenth of a bel. Therefore:

$$\text{Power gain in decibels} = 10\log_{10}\frac{\text{Power}_{\text{out}}}{\text{Power}_{\text{in}}}\ \text{dB}$$

25. The amplifier in Fig 9 would have a power gain, expressed in dB, of:

$$10\log_{10} 50 + 10\log_{10} 30 + 10\log_{10} 100\ \text{dB} = (10 \times 1.699) + (10 \times 1.477) + (10 \times 2) = 51.76\ \text{dB}$$

Note that when the gain of each stage is expressed in decibels, the overall gain of the amplifier is found by adding the gain of individual stages.

26. In the case of an attenuator the output power is less than the input power, the power loss being indicated by a minus sign, e.g.

Input power = 160 watts

Output power = 5 watts

$$\text{Attenuation in dB} = 10 \log_{10} \frac{\text{Power}_{\text{out}}}{\text{Power}_{\text{in}}}$$

$$= -10 \log_{10} \frac{\text{Power}_{\text{in}}}{\text{Power}_{\text{out}}}$$

$$= -10 \log_{10} \frac{160}{5}$$

$$= -15 \text{ dB}$$

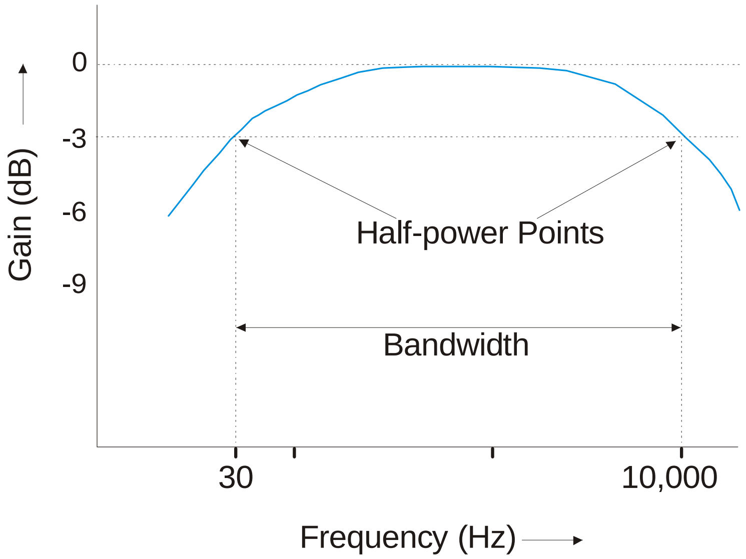
27. Fig 10 shows the frequency response curve for an amplifier; the gain is measured in dB. At the lower and upper ends of the curve where the response falls off, the gain can be expressed as minus so many decibels. A commonly used point is where the power has fallen to half the normal value. Therefore:

$$\text{Power loss} = 10 \log \tfrac{1}{2} = -10 \log 2$$

$$= -3\text{dB, or 3dB down}$$

This point is known as the half-power point and the frequency range between the two half-power points is known as the bandwidth of the amplifier.

14-14 Fig 10 Gain Frequency Curve



28. With voltage amplifiers it is more convenient to express the gain of an amplifier in terms of a voltage ratio.

$$\text{Power} = \frac{V^2}{R}$$

$$\text{Power gain} = \frac{V^2{}_{\text{out}}}{R_{\text{out}}} \div \frac{V^2{}_{\text{in}}}{R_{\text{in}}}$$

$$= \frac{V^2{}_{\text{out}}}{V^2{}_{\text{in}}} \times \frac{R_{\text{in}}}{R_{\text{out}}}$$

If the amplifier has the same value of input and output resistances:

$$\text{Power gain} = \frac{V^2{}_{\text{out}}}{V^2{}_{\text{in}}}$$

Using decibels the power gain becomes:

$$\text{Power gain} = 10 \log \left(\frac{V_{\text{out}}}{V_{\text{in}}}\right)^2 \text{ dB}$$

$$= 20 \log \frac{V_{\text{out}}}{V_{\text{in}}} \text{ dB}$$

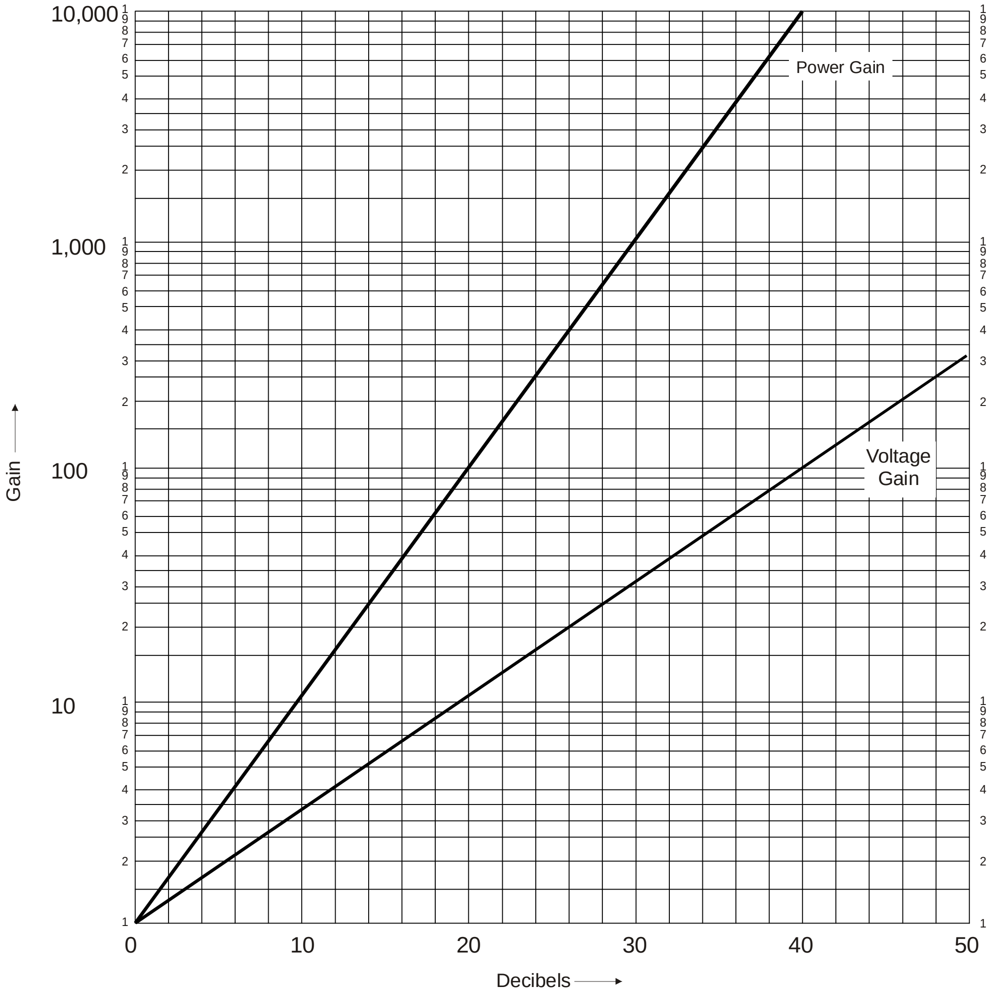
Similarly, since power also equals I^2R, gain can be expressed in terms of a current ratio provided that the values for R in input and output are equal. Therefore:

$$\text{Power gain} = 20 \log \frac{I_{\text{out}}}{I_{\text{in}}} \text{ dB}$$

29. The voltage at the half-power points in Fig 10 compared with the normal voltage of the amplifier is $\sqrt{\frac{1}{2}}$, since power = $\frac{V^2}{R}$. This ratio is 0.707 and so the voltage at the half-power points is approximately 70% of the normal voltage.

30. Fig 11 is a graph relating decibels to voltage and power ratios.

14-14 Fig 11 Conversion to Decibels from Voltage and Power Ratios

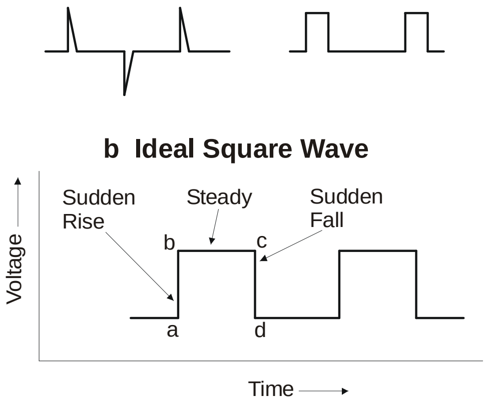


Video Amplifiers

31. Audio frequency signals are electrical representations of waveforms which can be heard. In radar and television, we need to amplify electrical signals which correspond to waveforms which can be seen. These signals are known as video signals.

32. Some video waveforms are shown in Fig 12. The important point to note is that some parts of the waveform consist of sudden changes in voltage and other parts are formed by steady or slowly changing voltages. The ideal square wave shown in Fig 12 is a typical video waveform. It rises to it maximum voltage in a very short time, ideally in zero time, as shown by points a and b. Then it remains at a steady voltage (b to c) and at c falls sharply to its original value (point d). The steady parts indicate a very low frequency; the sudden changes indicate high frequencies. In fact, the square wave is made up of a low fundamental frequency plus many harmonics, or multiples, of this frequency. If we want to amplify the square wave without distorting it, we must amplify all these frequencies by the same amount. This means designing an amplifier which can give equal amplification to a range of frequencies; in theory the range of frequencies is infinite but, in practice, it must cover several MHz. For TV signals, the range is about 30 Hz to 4.5 MHz. This is the major difference between the video amplifier and the audio amplifier.

14-14 Fig 12 Ideal Video Wave-forms



a Example Waveforms

b Ideal Square Wave

Radio Frequency (RF) Amplifiers

33. Amplifiers designed to handle signals from 20 kHz up to many thousand megahertz are called radio frequency amplifiers. They are used in transmitters to amplify the input signal from the receiver aerial. The main difference between RF amplifiers and the amplifiers so far considered is that one particular RF amplifier handles only a very narrow band of frequencies.

34. When a receiver is tuned to a particular radio station the RF amplifier is adjusted so that it amplifies only a narrow band of frequencies; signals from other stations on different frequencies are rejected. If another station is required the RF amplifier is adjusted to respond to the new station’s carrier frequency. Therefore, although the RF amplifier only accepts a narrow band of frequencies at any one time, the band can be chosen from a wide range of frequencies by changing the value of the tuning capacitor.

35. The same principles apply to both television and VHF broadcasting. A TV station may have a carrier frequency of 45 MHz and the signals occupy the band of frequency 42.5 MHz to 47.5 MHz. In this case the band is wider than that for sound broadcasting and it is different again at VHF; but in all types of transmission there is a radio frequency carrier wave with a narrow band of frequencies on either side.

36. The RF voltage induced in the receiving aerial is in the order of microvolts and it must be amplified many times before any use can be made of it. Usually some amplification is carried out at the RF signal frequency, some at a lower radio frequency and some at audio or video frequency. The amplification at the original signal frequency is done by the RF amplifier, but, in addition to amplifying the signal, this stage also selects the desired station. This process of selection is called tuning. Therefore when the tuning dial on a receiver is turned the frequency to which the RF amplifier stage responds is altered.

37. The amplification is done by the transistor and the selection of the narrow band of frequencies to be amplified is performed by the tuned circuits which form part of the amplifier circuit. The tuned circuits consist of coils and capacitors which are designed to resonate at the required frequency. Tuning is the process of bringing the tuned circuit to resonance.

Direct-coupled Amplifiers

38. A direct-coupled amplifier is one in which the output of one stage is connected directly into the input of the next stage and not via a coupling capacitor or transformer. Direct-coupled amplifiers are used to amplify voltages which change in value at a very slow irregular rate, ie voltages at very low frequencies. They are used in radar equipments, in analogue computers and in power supply circuits.

AMPLIFIERS WITH FEEDBACK

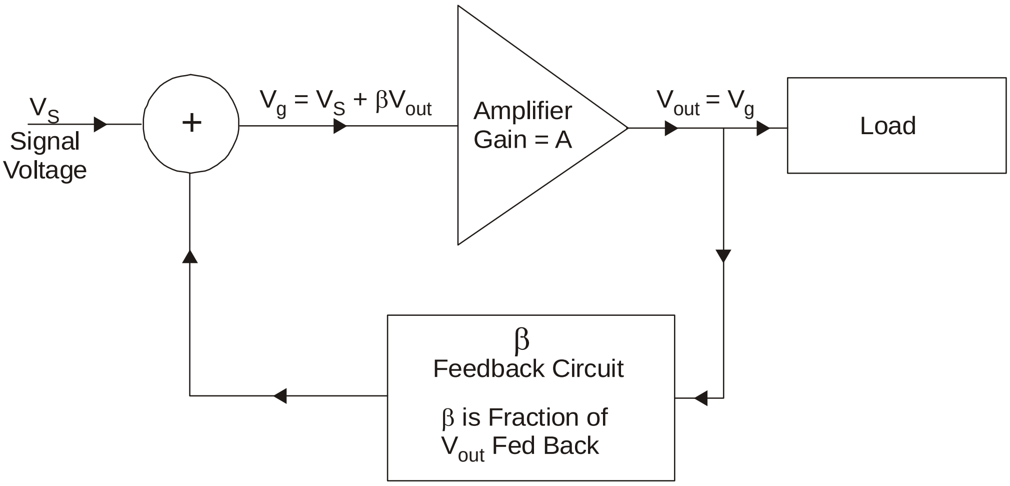
Introduction

39. If a fraction of the output of an amplifying device or amplifying stage is added to its input, then feedback is said to occur. Should this feedback be in-phase with the input then it is classed as positive feedback and oscillation takes place, (this is the subject of Volume 14, Chapter 15). When a part of the output is fed back in antiphase, the feedback is said to be negative.

Principle of Negative Feedback

40. Fig 13 depicts an amplifier with an “open-loop” gain (A), a signal voltage (V_S) and a feedback fraction of β.

14-14 Fig 13 Block Diagram Representation of a Feedback Amplifier



41. The 'closed-loop' gain, ie the gain with feedback, can be expressed by the following equations:

$$V_{out} = AV_g = A(V_s + \beta V_{out}) \text{ .... (in general)}$$

For negative feedback, β is negative,

Hence, $V_{out} = A(V_s - \beta V_{out})$

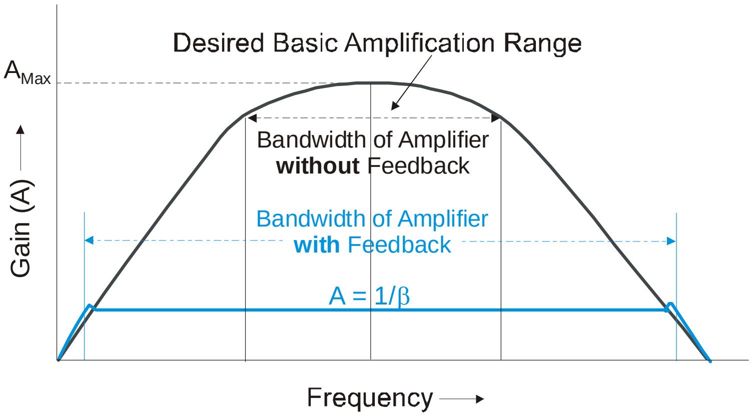
and, $\dfrac{V_{out}}{V_s} = \dfrac{A}{1+\beta A}$

42. If the 'open-loop' gain of the amplifier is very high then the effective gain, with feedback, closely approaches 1/β. The gain of the feedback amplifier is therefore independent of the gain of the basic amplifier. Variations in power supplies, ageing of components and other causes of variation in basic gain, have little effect on an amplifier operating with negative feedback.

The Effect of Feedback on Frequency Response

43. The upper curve of Fig 14 represents the frequency response of a basic amplifier; that is to say the device's ability to amplify over a specified range of frequencies. Ideally the frequency curve should be flat over the desired range in order to provide uniform amplification.

14-14 Fig 14 Frequency Response Curves of an Amplifier with and without Negative Feedback



In practice this does not happen, mainly due to the frequency sensitivity of the passive components associated with the amplifying device, eg capacitors and inductors.

44. The lower curve represents the frequency response of the same amplifier, but with negative feedback applied. The result is a flatter response curve with a central gain of 1/β. The fall-off in gain at the top and bottom ends of the band does not occur until the basic amplifier gain (A) has fallen to such a level as to become insignificant in the feedback equation.

Frequency Selective Feedback

45. The type of feedback discussed so far assumes a constant value of β regardless of the frequency of the input signal; this can be achieved using a simple resistive network. However, there are occasions when, for example, frequencies at the lower end of the band need to be amplified more than the ones at the top end, a form of bass boost. By introducing capacitance into the feedback network β becomes a variable value. High frequencies result in low reactance and a good deal of feedback, while low frequencies result in high reactance and reduced feedback.

Input and Output Impedances with Feedback

46. The application of negative feedback has a marked effect on the input and output impedances of an amplifier, and this is summarized by Table 2. These changes with respect to an amplifier without feedback depend on the method of feedback tapping and injection. Feedback tapping can take the form of voltage or current, and injection can either be series or parallel.

Table 2 Changes to Input and Output Impedances due to Feedback

	Series	Parallel	
Voltage	Higher Lower	Lower Lower	Input Impedance Output Impedance
Current	Higher Higher	Lower Higher	Input Impedance Output Impedance

47. These changes in impedance values may be used to some advantage when matching the output of one stage to the input of another. The emitter follower (Fig 4) is a good example, where the application of series voltage feedback makes the device ideally suited for matching a video output amplifier to a transmission line, or coaxial cable.

POWER AMPLIFIERS

Introduction

48. The amplifiers considered so far have, in the main, been designed to deal with voltage amplification where the reduction of distortion has been a major consideration. Although such amplifiers develop power in their loads, this power is of little importance. Power amplifiers, however, are those in which the power is of chief consideration.

49. Voltage amplifiers normally operate over the linear part of the device's characteristic and with comparatively small voltage swings, well within the supply voltages. On the other hand, power amplifiers must use all the voltage available in order to operate efficiently.

Classes of Operation

50. Voltage amplifiers normally work with a circuit arrangement known as Class A. Under these conditions the amplifier is regarded as being linear and provides low distortion but with low efficiency. The output current flows over the whole of the input cycle.

51. By biasing the input to the amplifier so that output current flows for only half the input cycle, greater efficiencies can be achieved. In order to duplicate the input waveform successfully, two devices working in a push-pull arrangement are required. This class of operation is known as Class B. With push-pull arrangement, Class AB can be used in order to reduce crossover distortion.

52. Radio frequency (RF) power amplifiers normally work under Class C conditions where the output current flows for less than half the input cycle. Although more efficient than other types of amplifier this class of amplifier introduces more distortion.

CHAPTER 15 - OSCILLATORS

Introduction

1. An oscillator is a circuit, containing active and passive components, which converts DC power into AC power at a frequency determined by the values of the components. In many cases, the output waveform is sinusoidal and the oscillator is then referred to as a sine-wave or harmonic generator. Oscillators which generate waveforms such as sawtooth, square-wave and triangular, are known as relaxation oscillators.

2. The main uses of oscillators are:

a. The generation of radio frequency (RF) for transmission.

b. The generation of RF for test equipment.

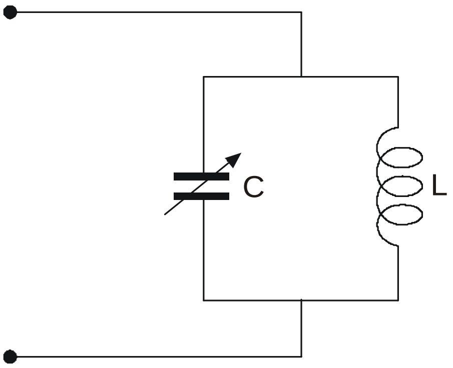
c. The generation of an intermediate frequency (IF) in superheterodyne receivers

d. The production of timing pulses in radar transmitters.

Maintenance of Oscillation

3. If an electrical impulse is applied to the circuit shown in Fig 1, the circuit will oscillate or 'ring' at its natural frequency. However, the resistance in the circuit gradually uses up the power contained in the original impulse and the amplitude of the oscillation decays exponentially. This is known as damping. (The process of natural oscillation is discussed in detail in Volume 14, Chapter 4.)

14-15 Fig 1 A Simple Oscillatory Circuit

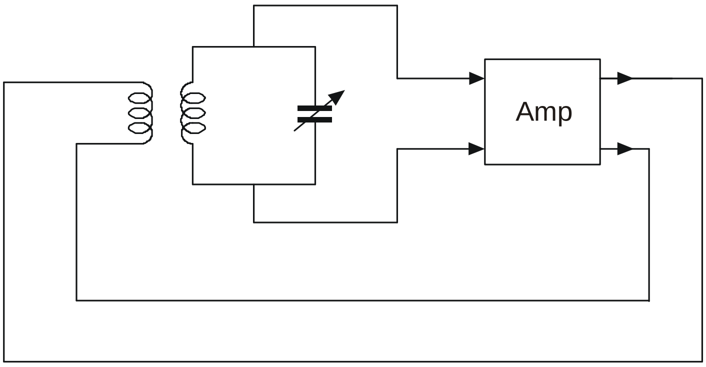


4. If, on the other hand, the circuit can be continually supplied with power at the right frequency and in the correct phase, the oscillations may be maintained at a constant level. This can be achieved by applying the oscillations to the input of an amplifier and then feeding part of the output back to the input, in phase. This is shown diagrammatically in Fig 2. Such an arrangement is called an oscillator, and the feedback required for the maintenance of oscillation is 'positive'. The frequency of oscillation depends on the values of the inductance and the capacitance in the LC 'tank' circuit. It is given, to a very close approximation, by the equation:

$$f_o = \frac{1}{2\pi\sqrt{LC}} \text{ Hz}$$

where L is in henries and C is in farads. This equation shows that the resonant frequency, f_0, increases with a decrease in either L or C. If the feedback is more than enough to compensate for the damping effect, it might be expected that the oscillations would build up to an infinite amplitude. In fact, the oscillations reach a maximum value determined by the flattening of the amplifier's characteristics.

14-15 Fig 2 An Oscillatory Circuit



Frequency Stability of Oscillators

5. One of the most important features of an oscillator is its frequency stability, ie its ability to provide a constant frequency output under varying conditions. Some of the factors which affect the frequency stability of an oscillator, and some of the methods used to counteract frequency drift, are as follows:

a. **Temperature**. Coils and capacitors alter in size, and therefore in value, with temperature changes. This can be compensated for by enclosing the tuned circuit in a temperature controlled compartment, or by using a capacitor whose value decreases with an increase in temperature, thus counteracting the corresponding increase in value of the inductor.

b. **Variations of Load**. The output from an oscillator is coupled to some other device, eg an aerial or an amplifier. Any change in the value of this load causes the oscillator frequency to drift. This effect can be reduced by taking only a fraction of the power which the oscillator could provide, ie loosely coupling the oscillator to its load. Alternatively, a buffer amplifier can be placed between the oscillator and load. This buffer stage screens the oscillator circuit from variations in load impedance.

c. **Changes in Power Supplies**. If the power supplies to the oscillator circuit vary, the output frequency will change. To avoid this, power supplies are carefully stabilized.

d. **Vibration and Shock**. If mobile equipment is subject to vibration, it can cause changes in oscillator frequency. Frequency drift due to this cause is reduced by mounting the equipment on shock absorbers.

e. **Hand Capacitance**. The proximity of the operator's hand or other parts of the body may introduce extra capacitance into the oscillatory circuit. This can be reduced by earthing one side of the tuned circuit or by screening the whole circuit.

Crystal-controlled Oscillators

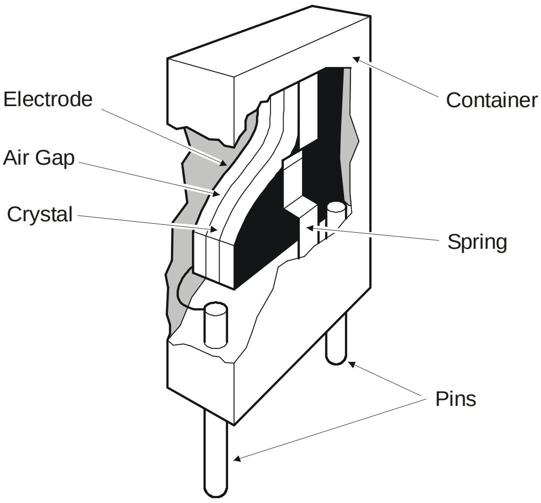
6. A class of oscillator which has exceptionally good frequency stability is the crystal-controlled oscillator. It is widely used when oscillations at one fixed frequency are required, but it cannot be tuned over a range of frequencies as can the LC oscillator. However, by using harmonics generated by several crystal oscillators, a number of spot frequencies in a wide range can be chosen.

7. If a voltage at the same frequency as the natural frequency of the crystal is applied to opposite faces of the crystal, resonance occurs. The crystal's natural frequency is very stable and will remain constant to within one part in 10^6. If temperature compensation and other arrangements are included, stability of 1 in 10^7 can be obtained. The range of frequencies covered by normal crystals is from 50 kHz to 9 MHz.

8. The crystal which controls the frequency is the same type as that used in the crystal microphone; it is made from Rochelle salt or quartz. These crystals exhibit a piezo-electric effect, ie they develop a voltage across opposite faces when they are compressed or expanded and they contract and expand (ie oscillate) when an alternating voltage is applied across them. Such a crystal has a natural frequency of oscillation which depends on its size, its thickness, and the way in which it has been cut; a thick crystal will oscillate at a lower frequency than a thin one.

9. The construction of an oscillator control crystal is shown in Fig 3. A thin crystal wafer is mounted between two metal plates from which electrical connections are taken.

14-15 Fig 3 Crystal Used for Frequency Control



Frequency Multiplication

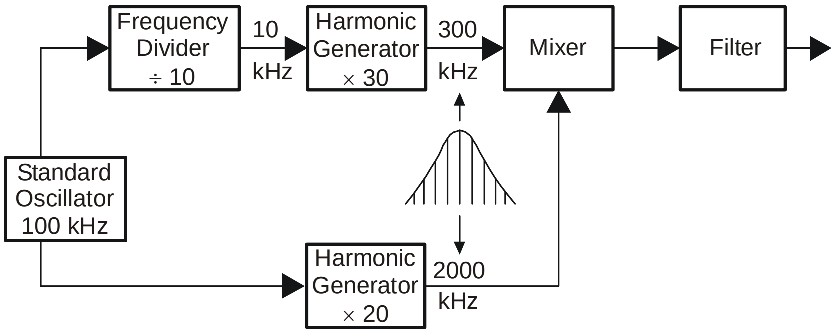
10. For frequencies above 14 MHz, it is not practical for quartz crystals to be ground thinly enough for general usage. If a crystal is used at a relatively low frequency, it can be used to control a non-linear amplifier which generates harmonic multiples of the fundamental. The crystal oscillator is more stable at a lower frequency.

11. If the LC output circuit of an amplifier is tuned to a harmonic of the input then the circuit receives a 'kick' every second or third cycle of its oscillation, which is sufficient to keep it going. In this way, any frequency may be multiplied up as many times as are desired, although it is not usual to multiply by factors greater than three in one stage.

Frequency Synthesizers

12. A frequency synthesizer is basically a circuit in which harmonics and sub-harmonics of a single standard oscillator are combined to provide a multiplicity of output signals which are all harmonically related to a sub-harmonic of the standard oscillator. A simple block diagram is shown in Fig 4.

14-15 Fig 4 Single Crystal Frequency Synthesizer



13. The output from the × 30 harmonic generator is a spectrum of frequencies, centred on 300 kHz, with a separation of 10 kHz. Similarly, the output from the × 20 harmonic generator is a spectrum of frequencies, centred on 2,000 kHz, with a separation of 100 kHz. The desired 'harmonic mix' is selected by the filter. A great advantage of this circuit is that accuracy and stability of the output signal is essentially equal to that of the standard oscillator, which can be tightly controlled.

Relaxation Oscillators

14. Relaxation oscillators generate non-sinusoidal outputs, the most important examples of these being the rectangular or square wave, and the sawtooth wave. Square waves are of immense importance in digital computers and radar, whilst sawtooth generators are used extensively as a source for the time-base waveforms associated with cathode-ray tubes.

15. The most commonly used square wave oscillator is the multivibrator. This generator, in its basic form, consists of two RC cross-coupled amplifiers; the output of one being connected to the input of the other, and vice versa. Normally, one device is conducting while the other is cut-off, and the change-over is determined by the RC time constant. The multivibrator action ensures rapid switching from the 'off' to the 'on' IF state for each amplifier, resulting in a square-wave output with sharp leading and trailing edges.

Microwave Oscillators

16. At frequencies higher than about 1,000 MHz, conventional low frequency devices and resonant circuits become inefficient. Most microwave components and equipment take on a different form. For example, a resistor is replaced by an attenuator, an LC tuned circuit by a resonator, a connecting wire or cable by a waveguide. The cavity magnetron and the reflex klystron are well-established examples of microwave oscillators. These devices, along with other methods of microwave generation, are discussed in Volume 14 Chapter 20.

CHAPTER 16 - TRANSMITTERS

Introduction

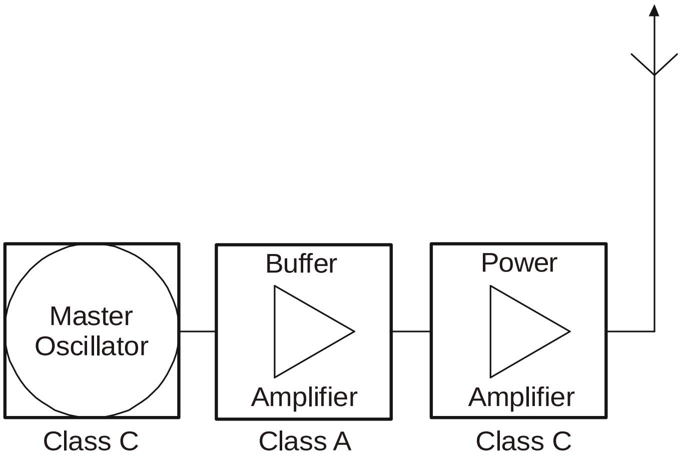
1. Many types of practical transmitters are in use, ranging from low frequency ground installations of considerable size, weight, and power, to miniaturized VHF and UHF equipment for airborne or man-portable applications. Although the appearance and uses of these transmitters may vary considerably, they all use the same basic principles.

2. The simplest form of transmitter comprises an oscillator connected to an aerial. However, this simple arrangement produces only a limited amount of radiated power, and suffers from poor frequency stability, since any oscillator from which more than a very little power is drawn tends to drift in frequency.

3. In a practical transmitter, these drawbacks are overcome by incorporating amplifiers between the oscillator and the aerial. The arrangement is illustrated in Fig 1 and is known as the Master Oscillator Power Amplifier (MOPA) System. The main components of the system and their functions are:

a. **Master Oscillator**. The function of the master oscillator is to produce a radio frequency voltage of good frequency stability. A crystal-controlled oscillator is normally used. To enhance frequency stability, it is normally run at low power and may have its crystal mounted in a thermostatically temperature-controlled enclosure.

14-16 Fig 1 The MOPA Transmitter



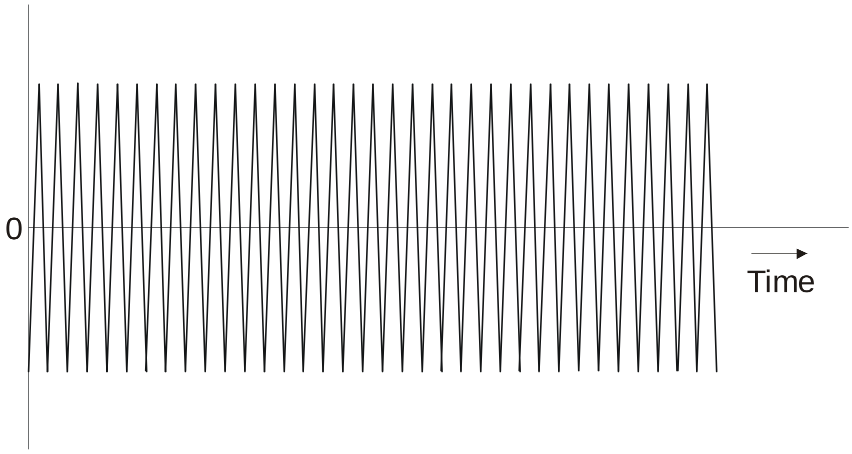
b. **Buffer Amplifier**. The buffer amplifier provides a constant light load for the master oscillator, thereby isolating the oscillator from load variations, which improves frequency stability. In addition, the buffer amplifier amplifies the RF signal from the oscillator to provide the necessary drive to the power amplifier. The buffer amplifier normally operates under Class A conditions (see Volume 14, Chapter 14, Para 50).

c. **Power Amplifier**. The function of the RF power amplifier is to convert the RF drive from the buffer amplifier to a sufficiently high-power level to feed the aerial and provide adequate energy for radiation. The power amplifier operates at a high power and at high efficiency under Class B or C conditions.

4. In many instances, especially at VHF or above, the frequency of the output at the aerial is required to be higher than that at which the master oscillator can efficiently maintain a stable frequency. In this situation, frequency-multiplier stages (doublers or treblers) are inserted between the master oscillator and the final power amplifier stage.

5. The simple MOPA transmitter described would transmit a constant frequency, constant amplitude, RF wave (Fig 2), ie it would carry no information. In order to transmit information, the transmitter output must be altered in some way.

14-16 Fig 2 A Constant Frequency, Constant Amplitude, RF Wave



KEYING AND MODULATION

General

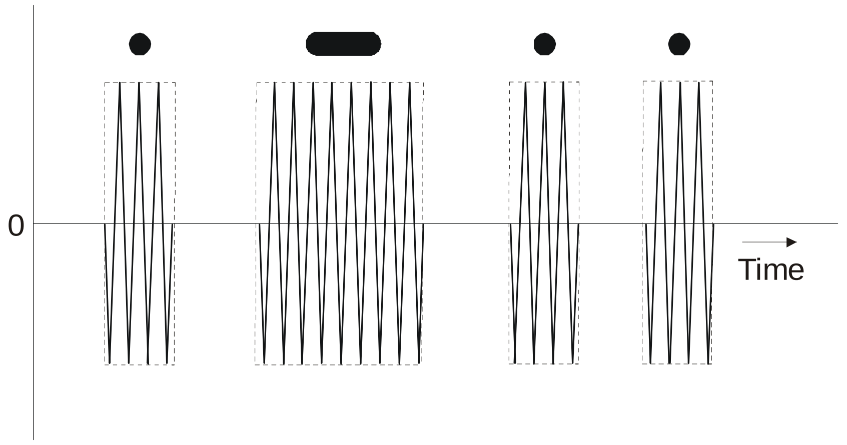
6. The techniques of superimposing information onto a transmitted signal are covered in Volume 14, Chapter 22. However, for convenience, the basic ideas will be reviewed in the following paragraphs, since the technique employed affects the arrangement of a practical transmitter.

7. The information to be transmitted may be either in the form of a quantity which varies continuously with time, e.g. speech, which is termed an analogue signal, or it may be in a form which is only permitted to have discrete values or levels, e.g. Morse code, which is known as a digital signal.

Transmitter Keying

8. In order to transmit Morse (or any similar) code by the continuous wave of Fig 2, it is necessary to switch the transmitter on and off in such a way that radiation from the aerial occurs for short and long periods of time corresponding to the dots and dashes of the code (Fig 3) - the operation is known as 'keying'. One method of interrupting the transmitted output would be to switch the master oscillator on and off, but this is not normally done since it may cause the frequency to drift. Instead, the switching is normally applied to one of the other stages in the transmitter.

14-16 Fig 3 Transmitter Keying - Morse Code

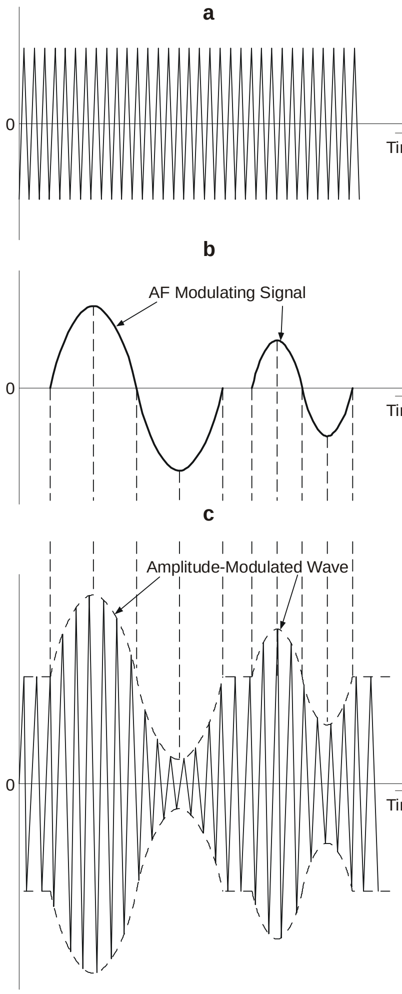


Analogue Modulation

9. Analogue information is transmitted by changing the amplitude, frequency, or phase (which has limited applications and is not considered further here) of the output (carrier) signal, to reflect changes in the information signal. Alternatively, the transmitter radiation may be concentrated into short bursts of RF energy, whose amplitude, length, or interval can be varied (pulse modulation).

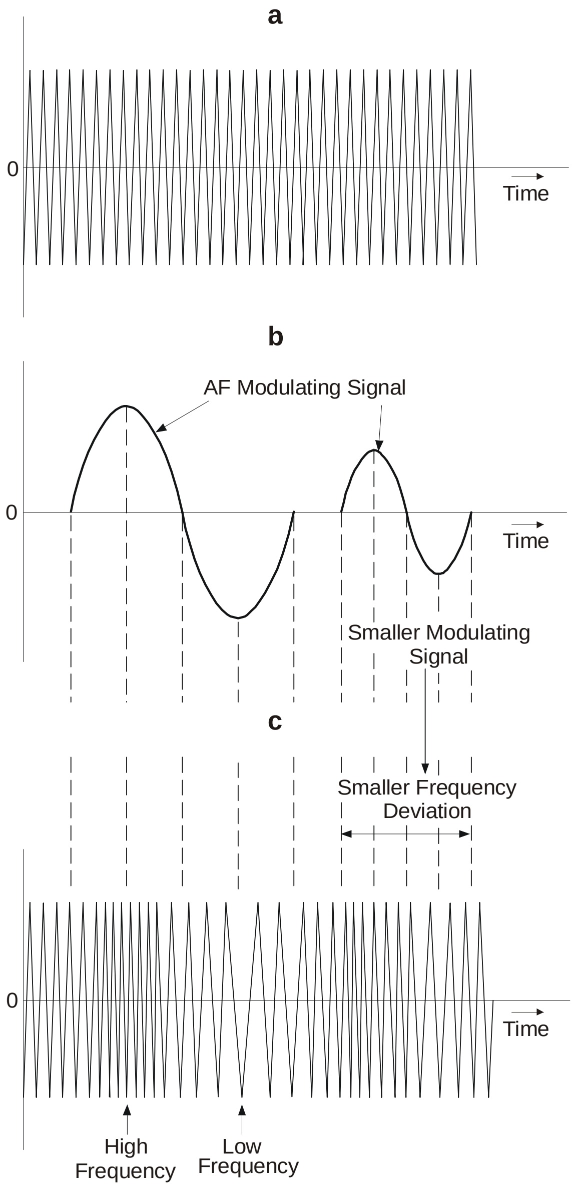
10. **Amplitude Modulation (AM)**. In amplitude modulation, the modulating signal is used to modify the amplitude of the RF carrier signal without affecting its frequency (Fig 4). Fig 4a represents the unmodulated CW carrier signal from the transmitter. Fig 4b shows the audio frequency modulating signal, and Fig 4c shows the change in the amplitude of the carrier, above and below the unmodulated level, proportional to the amplitude and sign of the modulating signal. The rate of change of amplitude of the carrier depends on the modulating frequency. The outline of the modulated wave, known as the modulation envelope, is an exact replica of the modulating signal. Single sideband systems (SSB) are a variation of AM.

14-16 Fig 4 Amplitude Modulation



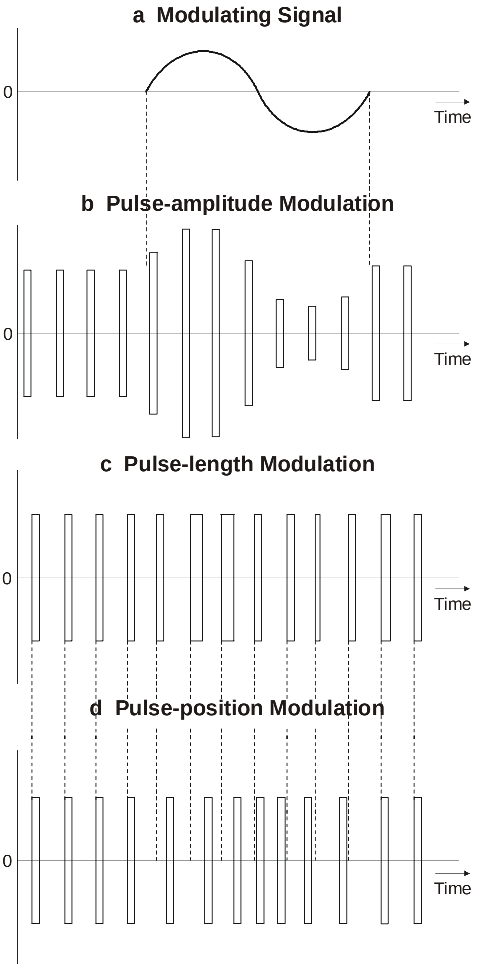
11. **Frequency Modulation (FM)**. In frequency modulation, the modulating signal is used to modify the carrier frequency without affecting its amplitude (Fig 5). Fig 5a represents the unmodulated CW carrier signal from the transmitter. Fig 5b shows the audio frequency modulating signal and Fig 5c shows the changes in the frequency of the carrier, above and below the unmodulated frequency, proportional to the amplitude and sign of the modulating signal. The rate of change of carrier frequency depends on the modulating frequency.

14-16 Fig 5 Frequency Modulation



12. **Pulse Modulation (PM)**. Pulse modulation can be used to convey information by varying the pulse amplitude (Fig 6b), the pulse length (Fig 6c), or the interval between pulses (Fig 6d) in accordance with the modulating signal (Fig 6a).

14-16 Fig 6 Methods of Pulse Modulation

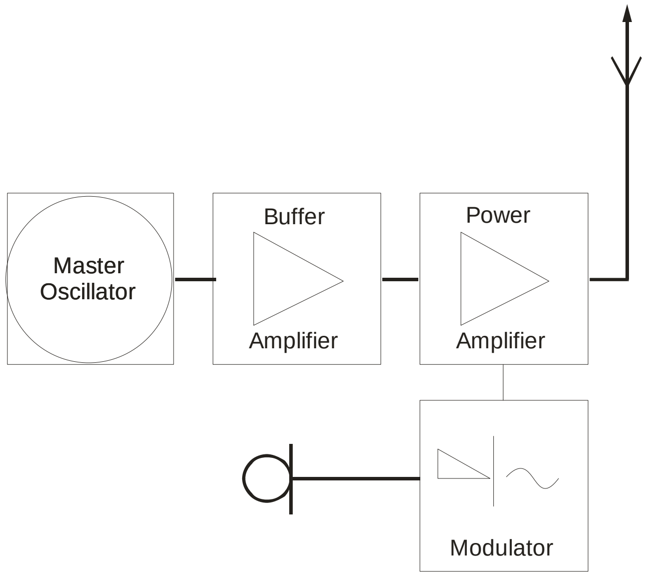

a Modulating Signal
0
Time
b Pulse-amplitude Modulation
0
Time
c Pulse-length Modulation
0
Time
d Pulse-position Modulation
0
Time

TRANSMITTER SYSTEMS

Amplitude Modulated Transmitters

13. The modulating signal in an AM system may be a keyed audio note representing the dots and dashes of the Morse code, or it may be an analogue signal derived, for example, from a microphone. Modulation is said to be high or low, depending on whether it is applied in the transmitter at a point of high or low RF power. If the modulation is applied at an early stage, ie low-level modulation, then the following stages must be operated in Class A or B in order to avoid distorting the modulated RF input. A block diagram of a high-level system is shown in Fig 7. The RF carrier is generated by the oscillator and then amplified to the power level required for radiation from the aerial. The AF amplifiers raise the power of the AF signal from the microphone to the level required for modulation. The AF signal amplitude modulates the RF carrier in the RF power amplifier stage. For a given total output power, low-level modulation would need very little modulation or audio frequency power, but a large pre-amplifier (PA) stage. The PA stage would be inefficient since it would have to be operated in Class A or B, but offsetting the disadvantage, the AF stage would be small without the need for a heavy, high power, AF transformer. High-level modulation uses an efficient Class C amplifier at the PA stage, but uses a bulky, high power, inefficient AF amplifier (the modulator).

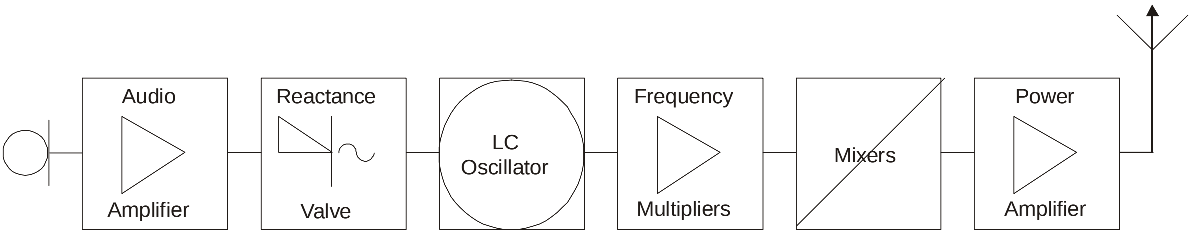
14-16 Fig 7 High-level Modulation AM Transmitter



Frequency Modulated Transmitters

14. A block diagram of a basic FM transmitter is shown in Fig 8. The AF signal is amplified and passed to the reactance device, which is used to vary the frequency produced by the oscillator. The reactance modulator will produce only small frequency deviations, and a practical FM transmitter must raise the frequency deviation by using frequency multipliers. For example, if the output from the oscillator is 2 MHz + 650 Hz, it is necessary to convert this to a VHF output of say 90 MHz with a frequency deviation of 75 kHz. To change the frequency deviation from 650 Hz to 75 kHz requires multiplication by about 115. Since this would produce a carrier of 230 MHz, which is too high, means must be found of bringing the carrier back to 90 MHz without affecting the frequency deviation. This is achieved by the mixer, which would be fed by an RF oscillator at a frequency of 140 MHz (230 MHz minus 90 MHz).

14-16 Fig 8 A Basic FM Transmitter



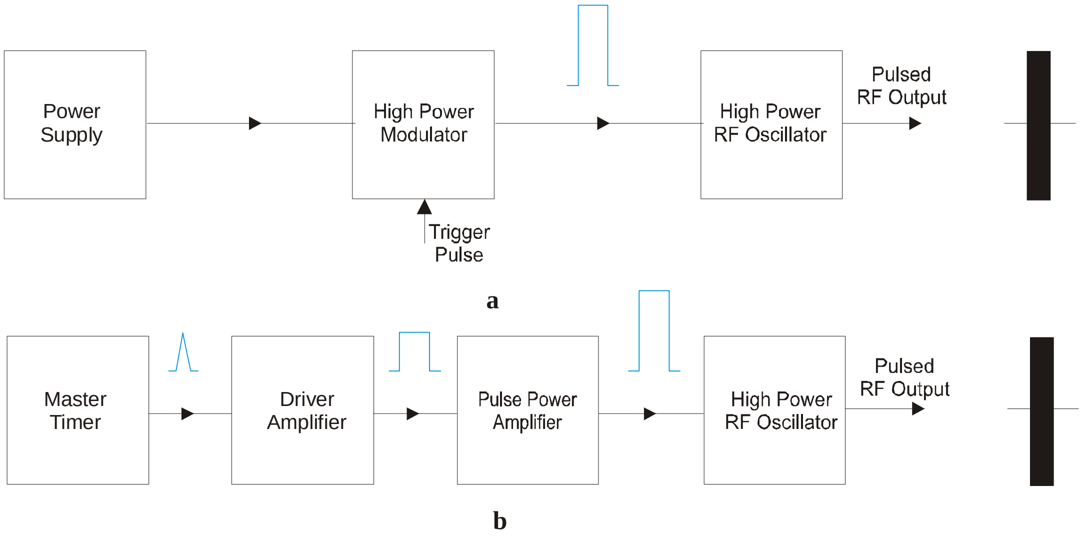
Pulse Modulated Transmitters

15. A pulse-modulated transmitter may be either a higher power oscillator type or a MOPA type. The choice between the two configurations is governed mainly by the application. Transmitters that utilize power oscillators are usually smaller than MOPA transmitters, but the latter are more stable and are usually capable of providing greater mean power. Power oscillators are therefore likely to be found in applications where small size and portability are of the greatest importance, and MOPA transmitters in radar applications in which high power or good MTI performance are desired.

16. **The High-Power Oscillator Transmitter**. The high-power oscillator, typically a magnetron (see Volume 14, Chapter 20), is switched on and off by the modulator. The modulator may be either a high-power stage, as shown in Fig 9a, or it may consist of several stages which amplify and shape the

master timing pulse, as shown in Fig 9b. In both cases, the RF energy of the pulse is generated in the oscillator and forms the final RF output.

14-16 Fig 9 High Power Oscillator Transmitters



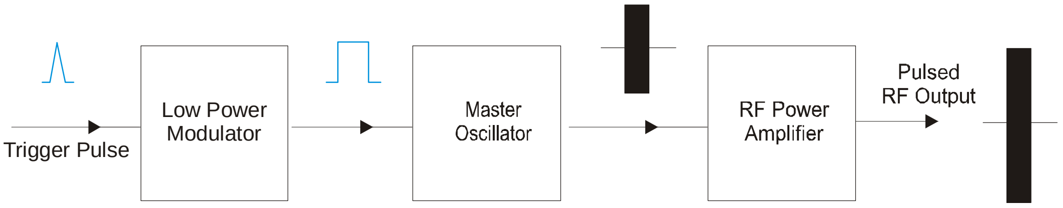
17. **The Master Oscillator Power Amplifier (MOPA)**. In the MOPA type of pulse transmitter (Fig 10), a low power RF oscillator is amplitude modulated, and its output is then amplified by the RF amplifier. The master oscillator may be crystal controlled, or it may be a stable frequency resonant cavity oscillator. The MOPA system has the following advantages over other systems:

a. The frequency stability is much better.

b. A simple modulator can be used because it does not have to provide the high power required to feed a single stage transmitter.

c. Phase coherency between radiated pulses is much easier to obtain.

14-16 Fig 10 A MOPA Pulse Transmitter

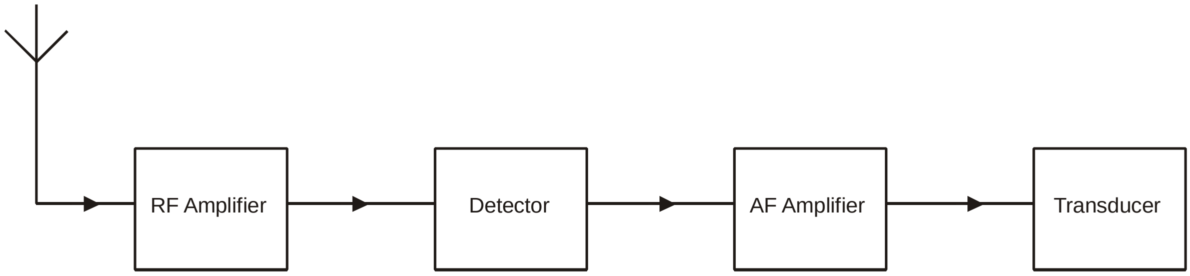


CHAPTER 17 - RECEIVERS

Introduction

1. The task of a radio frequency receiver is to intercept some of the RF energy radiated from the transmitter, to detect the information it contains, and to reproduce it in an acceptable form. The basic receiver shown in block form in Fig 1 uses an RF amplifier to select and amplify the required frequency, a detector to extract the information, and an audio frequency amplifier to produce enough power to operate the loudspeaker (transducer). This type of receiver could be used for single channel operation, but has limited potential for tuning to other channels.

14-17 Fig 1 A Basic Radio Frequency Receiver

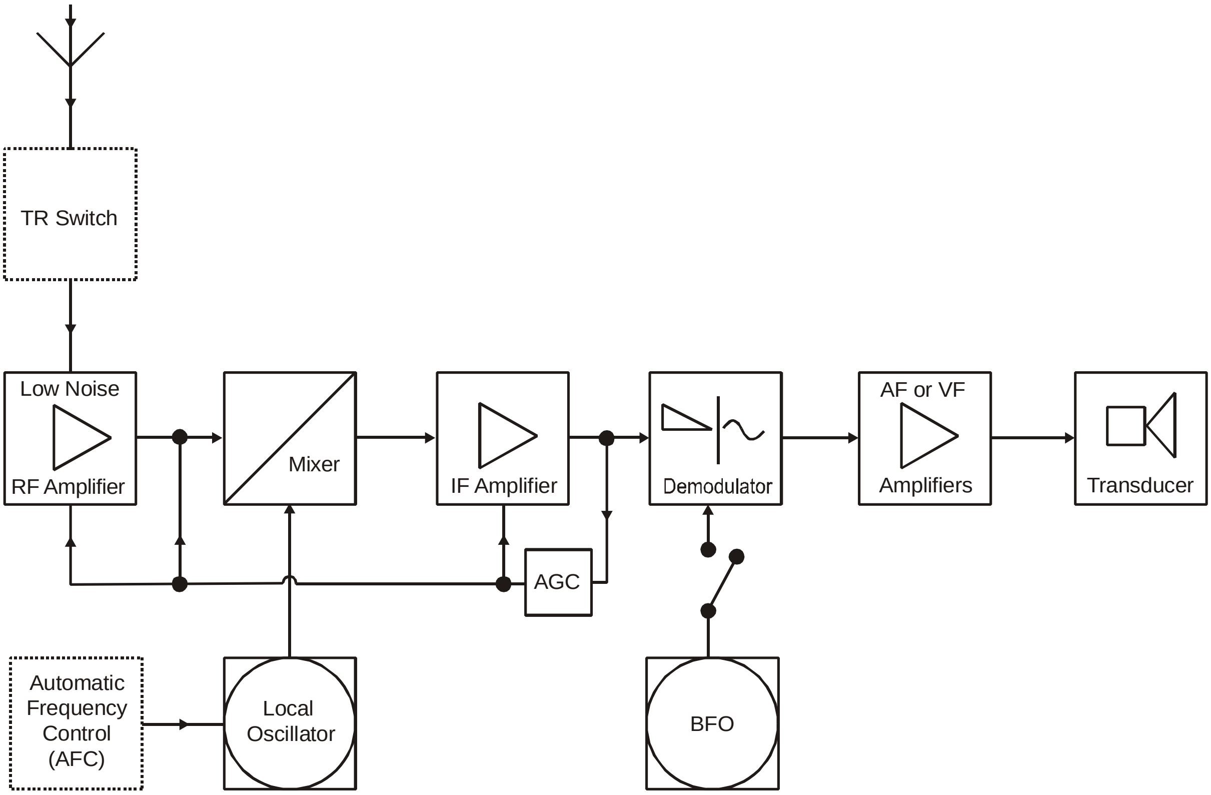


2. A practical receiver uses the superheterodyne (often abbreviated to superhet) principle in which all incoming signals are changed to a fixed frequency (the intermediate frequency - IF) for amplification before demodulation. Most of the amplification within the receiver takes place in the IF stages. Fig 2 depicts a multi-purpose, superhet receiver. The various stages are:

a. **Transmit/Receive (TR) Switch**. The TR switch is used to allow the aerial to act as a radiator or receiving aerial. TR switches are discussed in Volume 14, Chapter 20.

b. **RF Amplifier**. Although not always essential, an RF amplifier is usually included and is an RF voltage amplifier with Class A bias and a tuned circuit collector load. Its functions are to select the wanted signal from a background of other signals and noise, to give some pre-mixer amplification to the signal and hence improve the signal-to-noise ratio at the receiver output, and to isolate the aerial from the local oscillator so as to reduce radiation at the local oscillator frequency.

14-17 Fig 2 A Superheterodyne Receiver



c. **Local Oscillator**. The local oscillator can be any one of the basic oscillator circuits provided that it is stable over the required frequency range. The local oscillator and the RF amplifier tuned circuits are 'ganged' together so that there is always a constant fixed difference between their frequencies. At HF and below, the local oscillator frequency is usually chosen to be higher than that of the incoming signal in order to receive signals of lower frequency than the IF, and to keep the ratio of the maximum and minimum frequencies within the range of a normal tuning capacitor. At VHF and above, the local oscillator operates below the signal frequency in order to improve stability. At these frequencies the ratio of the maximum and minimum frequencies is much lower and well within range of the tuning capacitor.

d. **Mixer (First Detector)**. The inputs to the mixer are the modulated signal and the local oscillator output, the mixer producing the sum and difference of these frequencies. The difference frequency is selected by a tuned circuit and this forms the IF signal which is modulated with the same waveform as the original RF signal. The choice of mixer circuit depends primarily on the frequency of the input signal. At HF and below the multiplicative type is used, but above these frequencies additive mixing is used since this results in less interaction between the two inputs to the mixer. Interaction can be reduced further by the insertion of a buffer amplifier between the local oscillator and the mixer.

e. **IF Amplifier**. The IF amplifier is an RF amplifier operating at a low, fixed, radio frequency and because the frequency is fixed, this stage is always operating under optimum conditions. This permits the sensitivity, selectivity, and stability to be much higher than would be possible with a tuneable RF amplifier. In most receivers there are usually several IF amplifiers in cascade, forming what is known as the IF strip, and it is in this part of the receiver that most of the amplification is achieved. The IF amplifier also sets the bandwidth of the receiver overall.

f. **Demodulator (Second Detector)**. The demodulator is invariably a diode detector since the amplitude of the signal at this stage in a receiver is sufficiently large (a few volts) to allow a diode to operate efficiently and with good linearity.

g. **AF or Video Frequency Amplifier**. The AF or video frequency amplifiers form the final processing stage in a receiver and are necessary to bring the signal to a level suitable for presentation. The number of amplifiers present in this stage is determined by the power output required.

h. **Beat Frequency Oscillator (BFO)**. A beat frequency oscillator is required for the reception of CW signals. It produces a signal at a frequency which combines at the demodulator with the output of the IF stage to produce a component at the difference frequency, usually 1 kHz. This can now be detected and amplified to give an audio output. Since the BFO is not required when full AM signals are being received, an on/off switch is incorporated.

i. **Automatic Gain Control (AGC)**. The automatic gain control circuit reduces the gain of the receiver in proportion to the input signal strength, and the output thus tends to remain constant despite signal fading and changes due to tuning. The circuit achieves this by feeding back a DC voltage (from a diode driven by the last IF stage) which is proportional to the signal strength at that point. This voltage varies the bias on the earlier stages, thus varying the gain of the controlled stages. A simple AGC circuit reduces the gain of the receiver for all signals, even weak ones. This is a disadvantage and so the normal arrangement is to use a delayed AGC circuit which does not feed back an AGC voltage until the signal rises above some predetermined value.

Selectivity and Choice of IF

3. There are two main ways in which an unwanted signal may cause interference with the wanted signal in the superhet:

a. **Adjacent Channel Interference**. Adjacent channel interference is interference from a signal which is close in frequency to the wanted signal. It may be reduced or eliminated by making the tuned circuits of the IF amplifiers highly selective.

b. **Second (or Image) Channel Interference**. The carrier frequency of the wanted signal differs from the local oscillator frequency by the IF. However there is another frequency, also differing from the local oscillator frequency by the IF, but in the opposite sense to the wanted signal, which could cause interference. For example, consider a wanted carrier frequency of 2.2 MHz being received by a receiver with an IF of 500 kHz; the local oscillator would in this situation be tuned to 2.7 MHz. In addition to the wanted signal, a frequency of 3.2 MHz could also mix with the local oscillator to produce a 500 kHz difference frequency. This second channel interference can be prevented by ensuring that the image channel frequency lies well outside the pass band of the RF amplifier.

4. **Choice of IF**. The higher the IF, the further away is the second channel interference frequency, which makes it easier for the RF amplifier tuned circuit to discriminate between them. Conversely, a low value for the IF assists in discriminating between adjacent signals thereby reducing adjacent channel interference. The choice of IF must therefore be a compromise between these conflicting requirements. However, in high quality receivers where this compromise is not acceptable, the problem can be avoided by use of the double superhet receiver technique.

5. **The Double Superhet Receiver**. In the double superhet receiver the incoming signal is first changed to a high IF which assists in second channel rejection. After amplification at this frequency, a second mixer is used to produce a final low IF which improves adjacent channel rejection. Since the input to the second

mixer has a fixed centre frequency, the second local oscillator can be preset and its tuning does not have to be ganged to the tuning capacitors of the first RF and first local oscillator stages.

6. **Typical Output**. Typical signal and intermediate frequencies for various types of input signal are shown in Table 1.

Table 1 Typical Intermediate Frequencies

Type of Signal	Signal Frequency	IF
AM Broadcast	1 MHz	454 kHz
FM Broadcast	90 MHz	10.7 MHz
Television	150 MHz	34 MHz
HF Communication	10 MHz	600 kHz
VHF/UHF Comms	300 MHz	15 MHz - 1st IF 2 MHz - 2nd IF

7. **Functional Differences**. Although most RF receivers use the superhet principle, there are differences between communications and radar receivers, and between types of receiver within each group. These characteristics will be reviewed in the following paragraphs.

COMMUNICATION RECEIVERS

General

8. The essential features of a communication receiver are:

a. **Frequency Coverage**. No one receiver can possibly cover successfully the whole of the RF spectrum and so receivers are designed to operate in a particular band.

b. **Sensitivity**. Sensitivity is a measure of the ability of the receiver to intercept weak signals and extract information from them and it depends upon the amount of RF amplification available at the beginning of the receiver. Sensitivity cannot be increased indefinitely by increasing the number of RF stages because of instability due to interaction and usually only one amplification stage is used.

c. **Selectivity**. Selectivity is a measure of a receiver's ability to intercept only the required signal and to extract the information it carries, even though there may be other information carrying signals on close frequencies. An RF stage on its own is unable to provide the required response for good selectivity and the superhet principle is almost universal since it is easier to obtain the required selectivity at lower frequencies.

d. **Fidelity**. Fidelity is a measure of how well the receiver reproduces the received baseband signal without distortion. Whereas effective sensitivity and selectivity require a narrow bandwidth, good fidelity can only be obtained with a wide bandwidth.

e. **Stability**. As a communication receiver can be remotely operated it is necessary for the local oscillator to have good frequency stability.

Single Sideband (SSB) Receivers

9. The SSB receiver is a superhet arrangement but with the additional requirement of re-inserting the carrier at the correct frequency and amplitude. The carrier must be present in the receiver along with the sideband before demodulation can take place, but if the frequency is incorrect, the signal will be distorted, and if the amplitude is incorrect, the effective depth of modulation will be changed.

10. When the SSB signal is transmitted with a controlled carrier or a pilot carrier, a triple superhet receiver is often used. From the amplified composite signal, the carrier is selected by a filter, and used to produce an automatic frequency control (AFC) voltage (by means of a discriminator and reactance device) with which the frequency of the second local oscillator is controlled. If the transmitter drifts in frequency, the receiver will follow the drift because the pilot or controlled carrier indicates the direction and extent of the drift. A portion of the signal output from the third frequency changer is passed through a filter which rejects the sidebands and passes the carrier frequency only. After amplification and limiting, this is fed into the demodulator stage where it recombines with the sideband frequencies and enables demodulation to take place in the normal way.

11. With the suppressed carrier system there is no carrier present in the receiver to use as a reference, and so the frequency drift in both transmitter and receiver must be negligible. The oscillators are thus usually crystal and thermostatically controlled to provide this high standard of stability. The receiver is usually a double superhet, both local oscillators being controlled by the frequency of a very stable temperature controlled crystal oscillator which is also used to provide the re-insert carrier at the demodulation stage. The frequency at this point is that of the second IF; thus, the difference frequencies produced in the detector are the original audio frequencies, i.e. the modulation.

12. In airborne SSB systems, the receiver forms part of a combined transmitter-receiver, and the stable oscillators used in the frequency translation process in the transmitting function are also used as local oscillators in the receiver function. This helps to ensure the overall frequency stability of the system.

FM Receivers

13. The FM receiver is a superhet arrangement but with wider bandwidth IF amplifiers compared to an AM receiver, and with a limiter and discriminator taking the place of the envelope detector in the demodulation stage. The limiter eliminates unwanted amplitude variations in a signal by cutting off the positive and negative extremities of the waveform. The discriminator is a device which produces a voltage proportional to the instantaneous deviation of the FM signal. This voltage will then be equivalent to the original modulating signal. An FM receiver, operating at VHF and above, is used for short-range communication and there is therefore usually a fairly strong input signal. AGC is normally only necessary in fringe areas and where it is employed the AGC diode would normally take its input from the last IF stage.

Frequency Shift Keying (FSK) Receivers

14. The input to an FSK receiver is frequency modulated with very small deviation and so the basic requirements for the reception of FSK signals are a narrow bandwidth, stringent frequency stability, and selective filtering in the demodulator to separate the mark and space frequencies. The RF amplifier, mixer, and IF amplifier stages are similar to those of a normal communications superhet, although the bandwidths of these stages are reduced. The output from the last IF amplifier is fed to the input filter which accepts the mark and space frequencies but rejects unwanted interfering signals. The limiter passes constant amplitude mark and space signals to selective filters which separate the

mark and space frequencies and pass them to the detectors. The positive (mark) and negative (space) pulse outputs are then amplified in the wideband pulse amplifier, the output of which is well-shaped pulses of + 80 volts suitable for operating a teleprinter relay. An emitter-follower provides a correct match to a 600-ohm line without distorting the pulses. To improve the reliability of FSK systems, two separate aerials and receivers are sometimes used with a common crystal controlled local oscillator used for both receivers.

RADAR RECEIVERS

Introduction

15. Radar receivers are invariably of the superhet type but with certain special characteristics which differentiate them from the normal communications receivers. These desired characteristics are:

a. **High Gain**. High gain is necessary so that the weakest echo may be detected and gains in the order of 150 - 200 dB are common.

b. **Low Noise Figure**. The advantages of high gain will be negated if the receiver noise figure is also high since a reasonable signal-to-noise ratio at the input will be degraded to an unacceptable value at the output. Noise figures for microwave radar receivers should be no greater than about 6 dB.

c. **Bandwidth**. The bandwidth of a radar receiver will inevitably be a compromise since, whereas it must be wide enough to encompass the frequency spectrum of the transmitted signal plus any Doppler shift which might occur, it should also be relatively narrow to minimize noise. In pulse radar, the optimum bandwidth is the reciprocal of the transmission pulselength, and a receiver with such a bandwidth is said to be 'matched' to the pulselength. For CW radar, the optimum bandwidth is the reciprocal of the length of the 'pulse' generated in the Doppler filter as the radar beam scans through the target.

d. **Automatic Frequency Control**. Radar transmitters, especially those using self-excited power oscillators, tend to drift in frequency, and some means of automatic tuning must be incorporated in the receiver so that it follows the drift of the transmitter.

CW Receivers

16. In a CW radar, echoes are received while the transmitter is operating and it is, therefore, necessary (except for very low power systems) to employ separate aerials for transmission and reception, and elaborate measures are sometimes needed to minimize direct leakage of transmitter power into the receiver. However, a controlled amount of leakage entering the receiver along with the echo signal supplies the reference necessary for the detection of the Doppler frequency shift on the received echoes. The amount of isolation required depends on the transmitter power and the accompanying transmitter noise as well as the ruggedness and sensitivity of the receiver.

17. The receiver of a simple CW radar is, in some respects, analogous to a superhet receiver. It is termed a homodyne receiver or superhet receiver with zero IF. The function of the local oscillator is carried out by the leakage signal from the transmitter. This type of receiver suffers from poor sensitivity which is overcome by using a receiver with a non-zero IF. The reference signal in this case is derived from a portion of the transmitted signal mixed with a locally generated signal of frequency equal to that of the receiver IF. Since the output of the mixer consists of two sidebands on either side of the carrier, a narrow band filter selects one of the sidebands as the reference signal. This receiver is, therefore, sometimes called a sideband superhet receiver. The received signal ($f_o \pm f_d$) is mixed with the reference signal ($f_o \pm f_{if}$) to produce a difference frequency of ($f_{if} \pm f_d$). After IF amplification the Doppler signals are

resolved in a Doppler filter bank. This consists of a number of narrow band filters which together cover the range (f_{if} – $f_{d\ max}$) to (f_{if} + $f_{d\ max}$), where $f_{d\ max}$ is the maximum Doppler shift expected. Each filter is provided with its own detector and an output from any one of these indicates a target. An electronic switch is used for rapid examination of each filter detector in turn. The number of targets that the radar can resolve at any one time is equal to the number of Doppler filters.

18. The receivers employed in FMCW radars are similar to those of the simple CW radar. In the homodyne type, the receiver consists of a crystal mixer followed by a low frequency amplifier and a frequency-measuring device. A reference signal is necessary to be able to extract the Doppler shift and the range, and this is usually obtained by direct connection from the transmitter rather than by transmitter leakage, so that its magnitude can be more easily controlled. In the sideband superhet type of FMCW receiver, the output from the mixer is an IF signal of frequency (f_{if} + f_b), where f_b is composed of the range frequency f_r and the Doppler shift f_d. The IF signal is amplified and applied to a balanced detector along with the local oscillator signal f_{if}. The output of the detector contains the beat frequency (range frequency and the Doppler velocity frequency), which is amplified to a level where it can actuate the frequency measuring circuits. The output of the low frequency amplifier is divided into two channels: one feeds an average-frequency counter to determine range and the other feeds a switched frequency counter to determine the Doppler velocity.

Pulse Doppler Receivers

19. In a simple pulse Doppler system, a pulsed coherent transmitter is coupled to a common aerial via a TR switch. As in the CW system, the Doppler signals are extracted at an IF and targets are resolved in velocity by means of IF Doppler filters. However, whereas in the CW system a target signal which appears in the Doppler filters consists of a single frequency lying above or below f_{if} by the Doppler shift, f_d, in the pulse Doppler system the echo is pulsed and is therefore composed of a number of discrete components of pure CW separated in frequency by the PRF. If the echo signal is from a moving target, all components in this spectrum are shifted by the Doppler frequency and in the IF stage the central component has a frequency ($f_{if} \pm f_d$). In order to resolve the target velocity without ambiguity only this component must be allowed to appear in the Doppler filters, and because of this rejection of other than the central frequency component, much of the echo power is lost. The fraction of the total power contained in the central component depends on both pulselength and PRF, an increase in either increasing the fraction. This reduced effective power degrades the signal-to-noise ratio compared to an equivalent CW system.

20. The signal-to-noise ratio of the system described may be restored if the operation of the receiver can be confined to the short period in each pulse cycle when the required echo arrives. The effect of this is to reduce the effective noise power to an extent comparable to the reduction in effective echo power, the net result being that the signal-to-noise ratio is recovered to a value similar to that of an equivalent CW system. This is achieved by the addition of a gating circuit which causes the receiver to be opened for a short period, τ_g, after an interval, t, following the transmission of each pulse. The duration of the range gate, τ_g, is equal to or slightly greater than the pulselength, τ, and the interval t is controlled so as to cause the gate to coincide with the time of arrival of the selected echo.

21. In a single-gate system, the gate must be swept across the inter-pulse period in order to search for a target. This not only increases the search time, but also reduces the target information rate if the system is to be able to deal with multiple targets. An alternative solution is to employ a number of fixed range gates which together cover the entire inter-pulse period. As each gate requires a separate bank of Doppler filters, the cost in complexity can be high. For example, a pulse Doppler radar having a

duty cycle of .1 can have up to 9 range gates, each of which might feed as many as 500 Doppler filters. However, by using Fast Fourier Transform techniques, this number of filters can be accommodated on a single integrated circuit card.

Logarithmic Receivers

22. A requirement of a radar receiver is that it shall not easily saturate, so that for example, a weak target echo will still be visible when superimposed on a strong clutter signal. This requirement may be met by adopting the successive detection principle in which a detector is connected across the output of each IF stage, the outputs of the detectors being added together in a delay line in such a way that the delay introduced between the outputs of successive detectors is equal to the propagation delay through the intervening IF stage. Each stage, in addition to feeding the next stage, thus makes an independent contribution to the final output of the receiver. If the input to such a receiver is steadily increased, the final IF will be the first to saturate. After a further increase in input the penultimate stage will saturate, and so on. In this way, the resistance of the receiver to saturation effects is greatly extended. In such an amplifier there is an approximately logarithmic relation between input and output amplitudes, and a receiver employing an IF amplifier of this type is, therefore, known as a logarithmic amplifier.

CHAPTER 18 - TRANSMISSION LINES

Introduction

1. A transmission line is any system of conductors by means of which electrical energy can be transferred from one point to another with negligible loss. In radio and radar systems, the majority of transmission lines are used to transfer RF energy either from a transmitter to an aerial, or from an aerial to a receiver. The loss in each case must be as small as possible. In the first case, any loss of energy is wasteful since it involves a needless use of power at the transmitter, while in the second case, a loss of RF energy could mean a serious loss of signals at the receiver input. Thus, in any transmission line system, steps must be taken to ensure that energy losses incurred are negligible compared with the magnitude of the energy being transferred.

2. As well as having low losses, a transmission line should not 'pick up' stray external voltages which would degrade the signal/noise ratio. This is particularly important in a receiving system, where the strength of the signal is relatively low.

Types of Transmission Lines

3. All transmission lines consist of a conducting medium and a dielectric medium (which may be air). There will be electric and magnetic fields in both media, and it can be shown that the energy is transmitted along the line mainly by the fields in the dielectric, the conductors merely acting as guides. The four examples of transmission lines shown in Fig 1 represent the main types of line used in radio systems. At frequencies higher than the metric band (which includes most radar systems) waveguides are used, and these will be dealt with in Volume 14, Chapter 20.

4. **Open Twin Wire Feeder**. The open twin wire feeder consists of two parallel wires spaced a small fraction of a wavelength apart (less than $\frac{\lambda}{10}$). The magnetic fields around the two conductors tend to cancel, and losses due to radiation are kept to a minimum. It has the following properties:

a. **Advantages**:

(1) High transmitter power outputs can be handled without danger of breakdown of the air dielectric.

(2) Standing waves (see paras 10 and 11) can be easily measured.

(3) Maintenance of the line is relatively simple.

(4) The line is balanced; that is, the impedance between wire and earth is the same for each wire. This is an important property when considering matching the line to the load.

b. **Limitations**:

(1) It is bulky and rigid and can be used only on static installations.

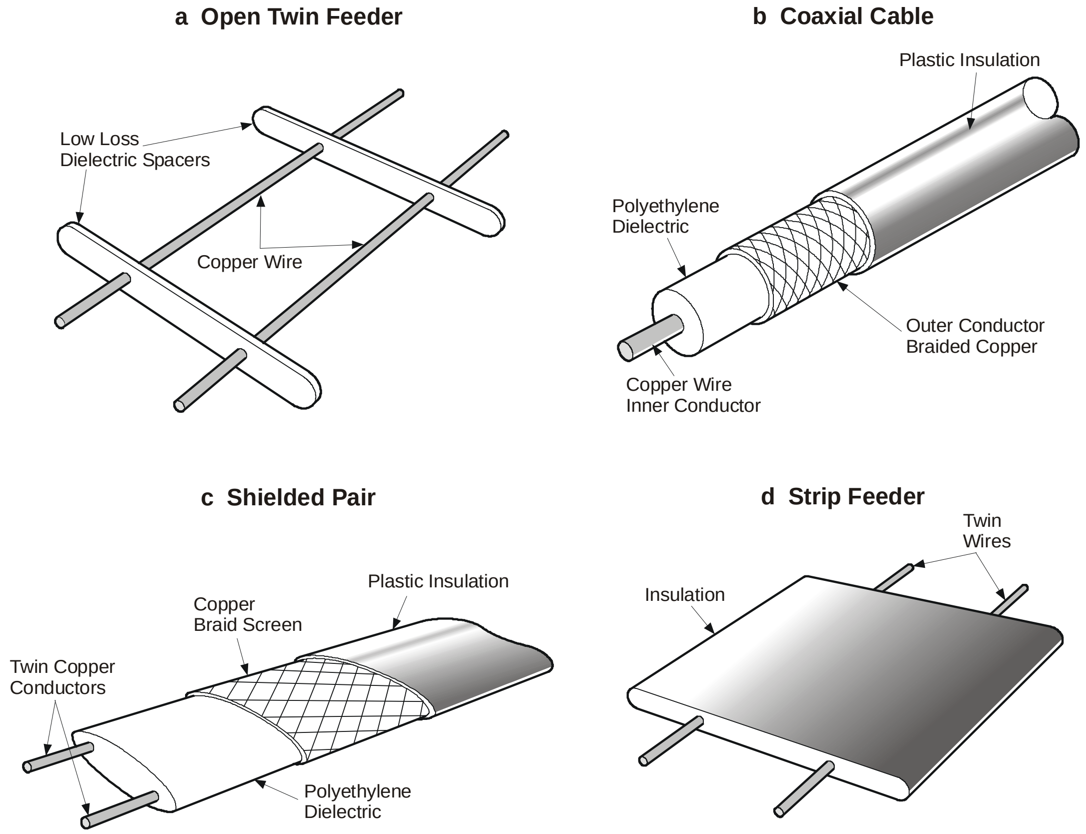
(2) At very high frequencies, the spacing becomes so small that, if high powers are being handled, there is a danger of 'flashover' between the wires. The upper frequency limit for high power installations is about 100 MHz.

(3) It must be kept clear of the ground and walls.

(4) Radiation losses limit the upper frequency to about 400 MHz. In practice, twin feeders cannot be used above 200 MHz.

c. **Applications**. It is the type of line invariably used for feeding balanced aerials for the HF band. It is usually run about 3 m above the ground and supported on insulators at intervals of 70 to 100 ft.

14-18 Fig 1 Types of Tranmission Line



a Open Twin Feeder

b Coaxial Cable

c Shielded Pair

d Strip Feeder

5. **Coaxial Cable**. Coaxial cable is a concentric type of twin wire transmission line. The inner conductor is held in the correct position relative to the outer conductor (braiding) by the use of insulating washers, spaced along the line at frequent intervals, or by completely filling the space between the conductors with a low-loss dielectric, eg polyethylene. Compared with the open wire feeder, its properties are:

a. **Advantages**:

(1) The coaxial line is a screened cable. The fields are confined within the space between the inner conductor and braiding. Since the braiding may be earthed, negligible energy from the outside will penetrate into the cable circuit; there is also little radiation from the cable.

(2) It is flexible and compact and can be buried in the ground.

(3) It can be used at higher frequencies than open wire line since losses due to radiation are negligible.

b. **Limitations**:

(1) Power handling capabilities are less than those of an open wire feeder.

(2) It is an unbalanced line; this introduces additional problems when matching the line to the load.

(3) Although it can be operated at higher frequencies than twin wire lines, it is subject to skin effect and dielectric losses which can cause a half power loss (3 dB) at 3,000 MHz, in three metres of line.

c. **Applications**. It is widely used for radio and radar purposes up to frequencies of the order of 3,000 MHz though waveguides tend to be used over about 1,000 MHz.

6. **Shielded Pair**. The shielded pair (or screened twin wire feeder) consists of two parallel conductors, mounted in a flexible braid screen, suitably insulated from each other and the screen by polyethylene. The metal braiding surrounding the polyethylene acts solely as a screen. In this way, the advantages of the coaxial feeder are combined with one advantage of the open wire feeder. The shielded pair is thus a screened and flexible line as well as being balanced. However, its power handling capacity is relatively small, and it has higher power losses for a given size and weight compared with coaxial cable. It is, therefore, used only for special applications where a balanced line is required.

7. **Strip Feeder**. The strip feeder consists of two conductors mounted along the edges of a strip of insulating material. It is extremely flexible but has a relatively small power carrying capacity. Therefore, it is usually used only for receiving systems. It also has applications in modern radar systems, in printed circuit form, in connection with microwave filters, receiver heads, printed aerial arrays and miniaturized microwave assemblies.

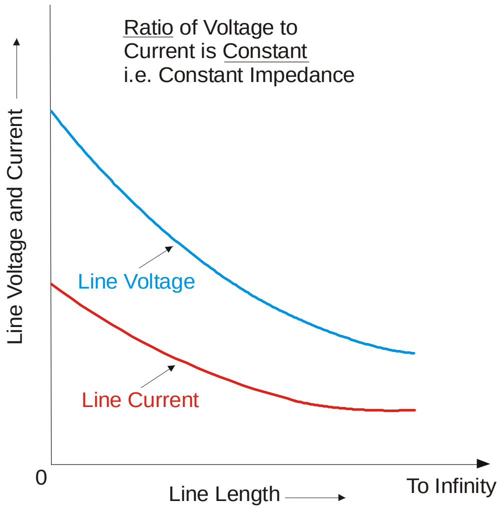
Characteristic Impedance

8. The ratio of voltage to current, measured at intervals along a transmission line of infinite length, is found to be a constant, as shown in Fig 2. This ratio is known as the characteristic impedance (Z0) of the line. It is a complex quantity, but at very high frequencies the characteristic impedence is a pure resistance given by:

$$Z_0 = \sqrt{\frac{L}{C}} \text{ ohms}$$

where L = inductance per unit length, and C = capacitance per unit length.

14-18 Fig 2 Variation of Voltage and Current with Line Length



9. Instead of using inductance and capacitance per unit length to calculate Z0, it is possible to use the physical dimensions of the line as follows:

a. **Twin Feeder**.

$$Z_0 = \frac{276}{\sqrt{k}} \log_{10} \frac{D}{r}$$

where D = distance between the centres of the wires, r = radius of wire,
k = relative permittivity of dielectric (1 for air).

b. **Coaxial Cable**.

$$Z_0 = \frac{138}{\sqrt{k}} \log_{10} \frac{b}{a}$$

where a = radius of inner conductor, b = inner radius of outer conductor.

Standing Waves

10. Although an infinitely long line is a physical impossibility, it is possible to get the characteristics of such a line to exist in a finite length if the practical limit is terminated in a resistance equal to the Z0, of the line. If a transmission line is terminated in a resistance which is not equal to Z0, the load cannot absorb all the energy being sent down the line and some of the energy will be reflected back to the generator. Although there are an infinite number of mismatch conditions, only the two extremes, open and short circuit, need be considered. Neither of these is capable of absorbing any energy, and so the incident wave is totally reflected. In order to reverse the direction of travel of an electromagnetic wave, only one component's direction should be reversed. In the case of a short circuit it is the electric field (voltage) component that is reversed, or phase shifted by 180º, whereas, for an open circuit, it is the magnetic field (current) component that is phase shifted by 180º. Thus, at a short circuit, incident and reflected electric fields cancel out to give a zero electric field, and incident and reflected magnetic fields reinforce to give a maximum magnetic field. The converse applies to the open circuit.

11. **The Standing Wave**. The resultant voltage distribution produced by the combination of the incident and reflected voltage waves is referred to as a standing wave. The positions of its maxima and minima on the line are fixed, and it rises and falls sinusoidally about these fixed positions. The distance between an adjacent maximum (anti-node) and minimum (node) is a quarter wavelength. Voltage nodes coincide with current anti-nodes, and voltage anti-nodes with current nodes.

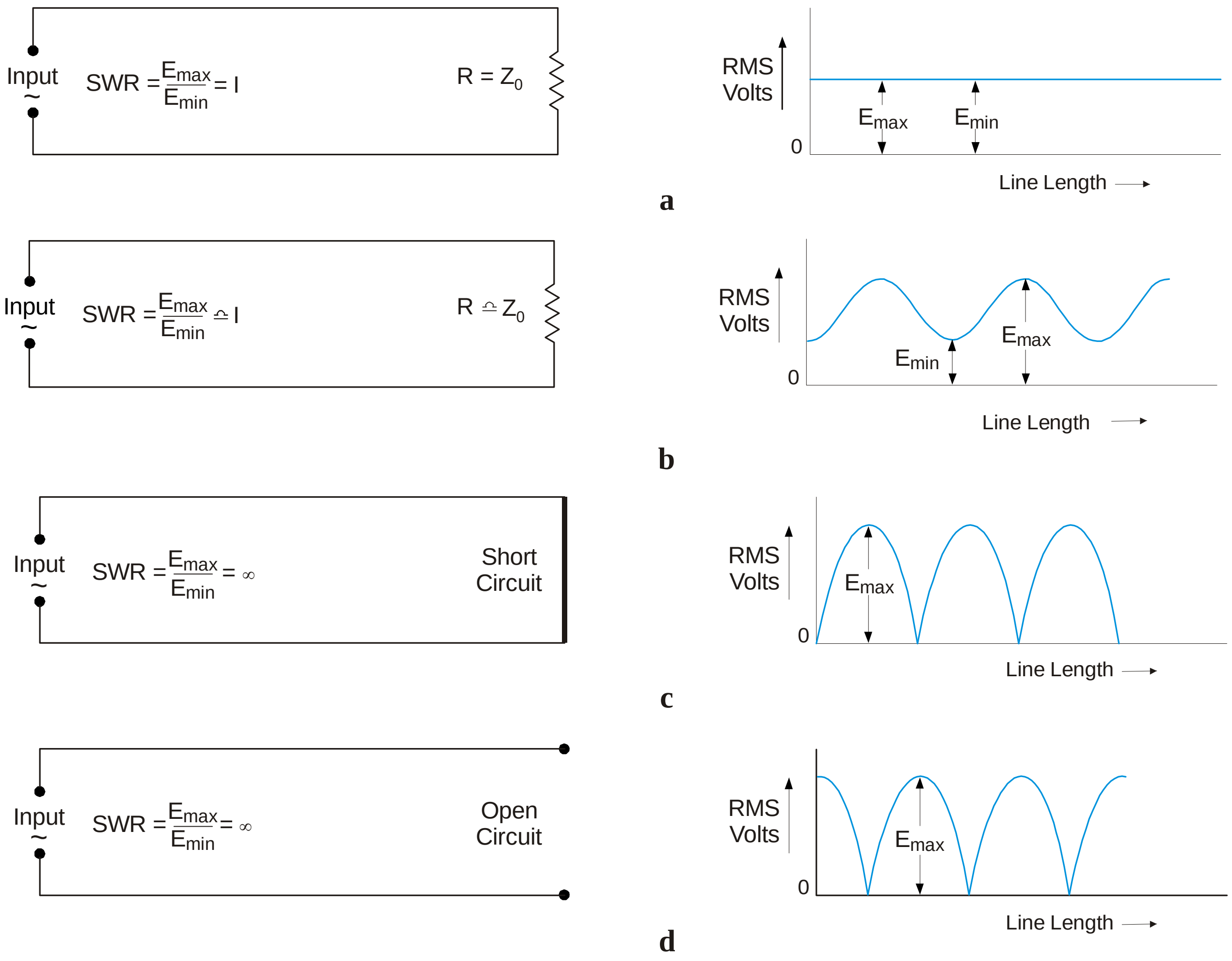
12. **The Standing Wave Ratio (SWR)**. The range of variation between the maximum and minimum values of standing wave current or voltage gives an indication of the degree of mismatch, and is usually expressed in terms of the standing wave ratio as follows:

$$\text{SWR} = \frac{E_{\text{max}}}{E_{\text{min}}} \text{ or } \frac{I_{\text{max}}}{I_{\text{min}}}$$

where E max etc are rms values indicated by a meter.

13. **The Effect of Line Termination**. Fig 3 shows the effect of different line terminations on the standing wave ratio. When the line is correctly terminated in a resistive impedance equal to Z0, there is no standing wave, Emax equals Emin and the SWR is unity (Fig 3a). With a slight mismatch (R slightly less or slightly more than Z0), the reflected wave is small, Emax is only slightly greater than Emin and the SWR is slightly greater than unity (Fig 3b). As the degree of mismatch increases, the SWR increases, and for the short-circuited (Fig 3c) or open-circuited (Fig 3d) transmission line, the SWR is Emax/0 = infinity.

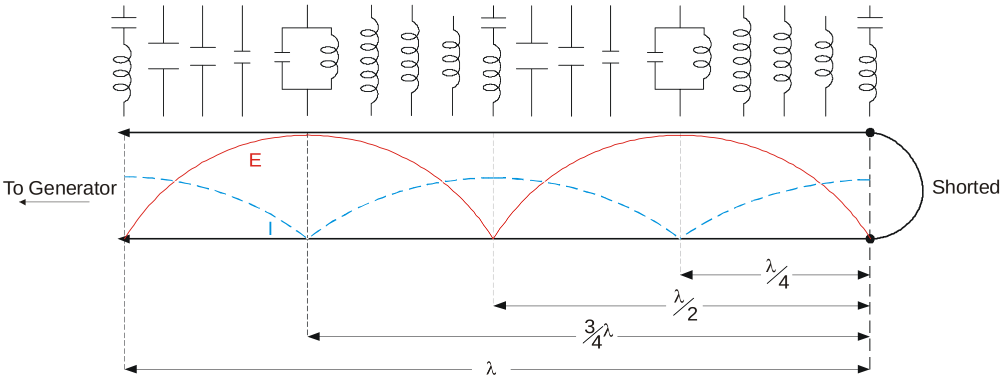
14-18 Fig 3 Effect of Termination on Standing Wave Ratio



Input Impedance

14. For a correctly terminated line, the input impedance is constant, no matter what the length of the line, and is equal to Z0. With a mismatched line, however, the input impedance varies with length. Fig 4 shows a short-circuited line, in which the impedance varies from minimum at the short circuit (corresponding to a series-tuned circuit at resonance), through a region of increasing inductive reactance to a maximum (corresponding to a parallel tuned circuit at reasonance) a quarter wavelength from the short circuit, and then through a region of decreasing capacitive reactance, to a minimum again at the half-wavelength point. A short length of open or short-circuited line is, therefore, a most versatile component, and can act as inductor, capacitor, rejector circuit or acceptor circuit, depending on its length.

14-18 Fig 4 Input Impedance along a Short-circuited Line



Matching

15. Where a load does not match the line directly, a short piece of suitable line can be used as an impedance matching device. It can be shown that a quarter wavelength of line impedance ZT will match a load ZL to the main transmission line of impedance Z0 when:

$$Z_T = \sqrt{Z_0 Z_L}$$

The impedance (ZT) required for this quarter wave transformer is achieved by either varying the diameter of its conductors, or, more commonly, by varying their distance apart.

16. **Stub Matching**. Where the load is reactive as well as resistive, the reactive component may be cancelled by a short-circuited line of equal but opposite reactance. This is termed 'stub matching' and may be used, either alone, or in conjuction with a quarter wave transformer.

Transmission Line Losses

17. In their primary role as carriers of energy, transmission lines are subject to a number of losses. The most important of these are:

a. Radiation loss.

b. Skin effect loss.

c. Dielectric loss.

18. **Radiation Loss**. Radiation loss affects twin feeders. The electromagnetic waves are merely guided by the line in the insulating medium surrounding it, and there is thus a tendency for the waves to escape into space. This is desirable in an aerial, but not in a transmission line. The fields of the wave may also induce currents in nearby conductors, providing another source of loss. The seriousness of radiation loss depends on the spacing between the waves in terms of the wavelength (the spacing should be less than λ/10), and the use of twin feeder is restricted to frequencies below about 400 MHz, or lower if the line has to carry high power. Radiation loss does not affect coaxial line since the waves are contained within a closed reflecting surface and are thus unable to escape. Coaxial line may be used up to about 3,000 MHz and is limited in its use by skin effect and dielectric losses. These two losses also affect twin feeders, but are overshadowed in their effect by radiation loss.

19. **Skin Effect**. Skin effect loss is basically a resistive loss (i.e. ohmic heating). The resistance of a conductor to alternating current increases with increasing frequency, since the current is caused to flow in a progressively decreasing depth of the conductor as a result of the magnetic fields set up by the currents. Thus, at frequencies of about 1 GHz, a thin-walled hollow tube has the same resistance as a solid tube of the same diameter. The effective increase of resistance is more marked for small diameter conductors than for large diameter ones.

20. **Dielectric Loss**. Dielectric loss arises partly from conduction currents, ie imperfect insulators, and partly from dielectric hysteresis. Dielectric hysteresis is the electrical equivalent of magnetic hysteresis and is a lag in electric flux behind the applied electric field. If the field is alternating, there is a heating effect in the insulator similar to the heating of a ferromagnetic material to which an alternating magnetic field is applied. Reduction of dielectric loss may be accomplished in one of two ways:

a. **Improving the Dielectric**. The best dielectric is air, and air-spaced, or semi-air-spaced, lines are extensively employed. If large powers are to be handled, this may pose technical problems since the spacing between lines affects the radiation loss (see para 18).

b. **Reducing the Volume of Dielectric**. Reducing the volume of dielectric reduces the power handling capacity of the line and also increases the skin effect losses.

CHAPTER 19 - AERIALS

Introduction

1. The purpose of an aerial is to act as a transducer between free-space propagation and guided-wave (transmission line) propagation. A transmitting aerial converts the electrical signals from a transmitter (wireless or radar) into an electromagnetic wave, which then radiates from it. A receiving aerial intercepts this wave and converts it back into electrical signals that can be amplified and decoded by a receiver.

2. In this chapter aerials will, in the main, be considered as radiating elements, however, the properties of a transmitting aerial apply equally well to a receiving aerial. This law of reciprocity means that many installations can use, in conjunction with suitable switching, a common aerial for transmission and reception.

3. Aerials used at centimetric wavelengths (microwaves) are dealt with as a separate subject in Volume 14, Chapter 20.

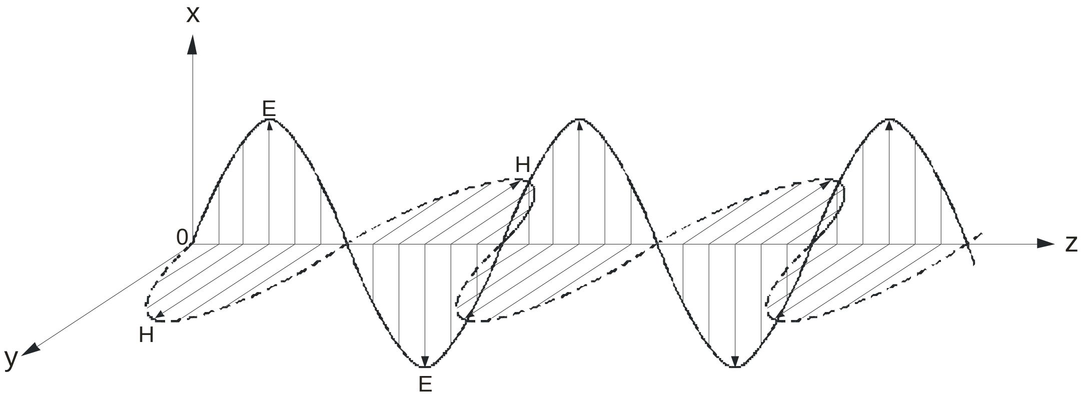
Properties of Electromagnetic Waves

4. An electromagnetic (em) wave consists of mutually sustaining and moving electric (E) and magnetic (H) fields which are at right-angles to the direction of propagation, ie it is a transverse wave. The E and H fields are always at right-angles to each other in space as well as being at right-angles to the direction of propagation.

5. At any point in space the E and H fields are in phase with one another, and their variation as a function of distance in the direction of propagation is as shown in Fig 1. The velocity of propagation is 186,240 miles/sec, or 3×10^8 metres/sec.

6. The ratio of E/H is a constant of 377 ohms, and is known as the impedance of free space.

14-19 Fig 1 Electromagnetic Waves



Polarization

7. The polarization of an electromagnetic wave is defined as the orientation of the E field. This is convenient because the electric component is in the same plane as the linear radiating element of the aerial. Thus, a simple linear radiator orientated horizontally with respect to the earth will emit

horizontally polarized waves. It is for this reason that aerials are often referred to as being horizontally or vertically polarized.

8. There are, however, certain types of aerial which emit a circularly or elliptically polarized wave. Under these conditions the directions of the E and H fields are not constant but rotate round the direction of propagation with constant amplitude, in the case of circularly polarized waves, and with varying amplitude in the case of elliptically polarized waves.

The Half-wave Dipole

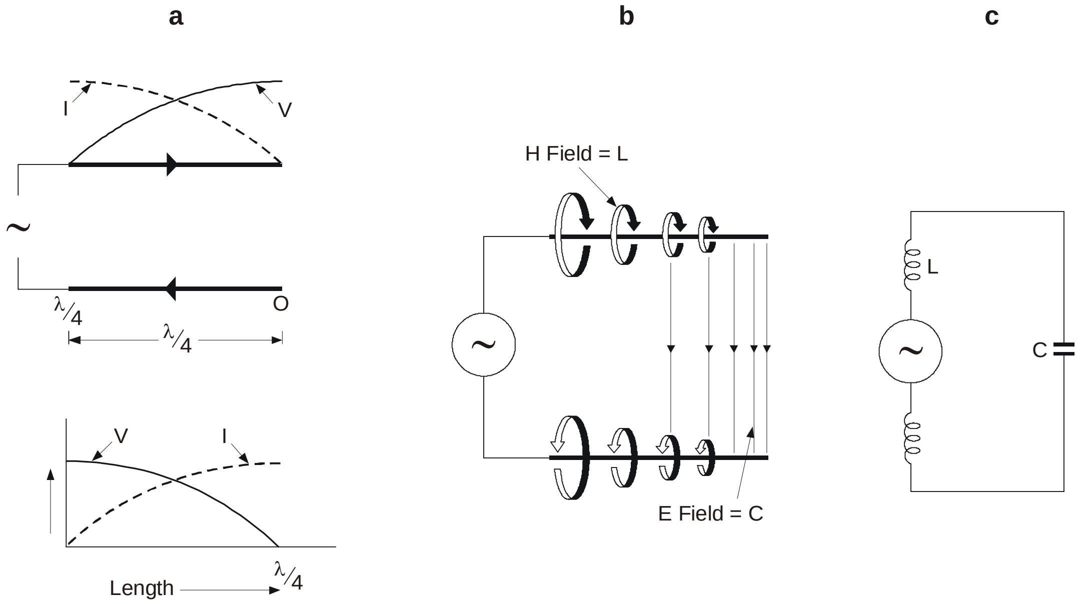
9. A simple transmitting aerial is a conductor, usually a wire, which is designed to radiate em waves as efficiently as possible. The simplest form of aerial is a half-wave dipole. This is a wire which is a half-wavelength long at the frequency of the current being carried by the wire, if the frequency is 30 MHz (10 metres wavelength), a half-wave dipole for use at this frequency will be 5 metres long.

10. A half-wave dipole can be shown to work by comparing it with a quarter-wave section of open-circuited transmission line. Fig 2 shows that such a section may be considered as a series resonant circuit; it has a low input impedance and the current at resonance is maximum with the impedance at a minimum. The current and voltage distribution along the stub is shown in Fig 2a; note that the stub length is measured from the open-circuited end. The positions of maximum magnetic field (equivalent to inductance) and maximum electric field (equivalent to capacitance) are shown in Fig 2b.

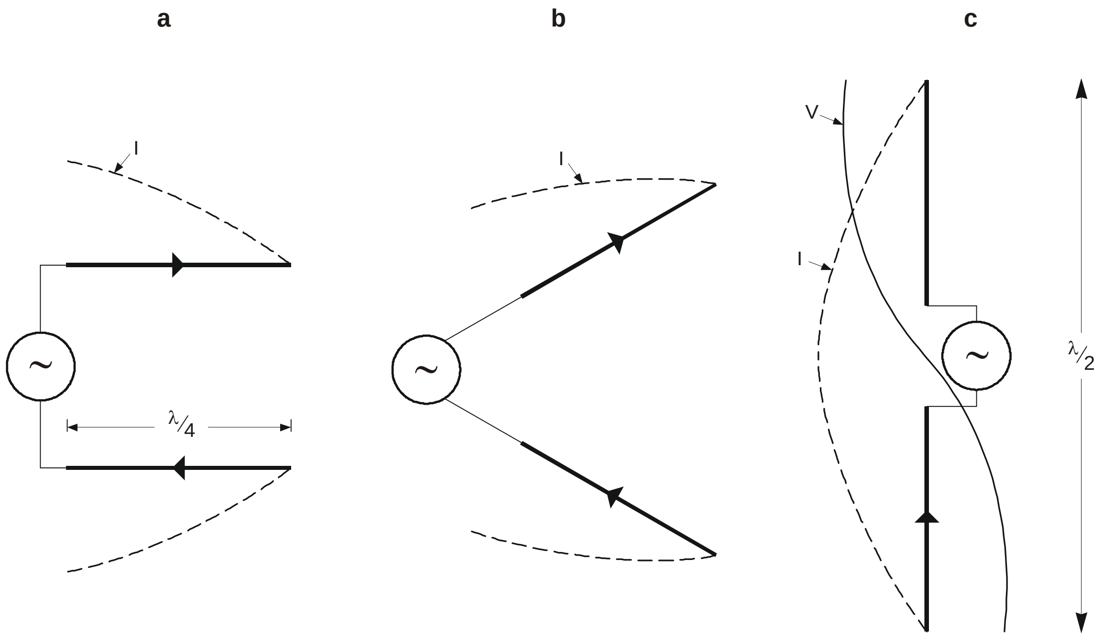
11. The magnetic fields around the two conductors are caused by currents flowing in opposite directions. In the space between the conductors these fields reinforce each other, but away from the conductors the fields cancel; thus very little radiation occurs.

12. However, if the lines are opened out as shown in Figs 3b and c, the currents in the two conductors flow in the same direction and have a maximum value at the centre falling to a minimum value at the ends. There is now an open oscillatory circuit and em waves are radiated into space. This is the dipole aerial. The standing waves of voltage and current along the dipole are shown in Fig 3c. As the aerial is vertical the axes of the graph have moved through 90° compared with those of Fig 2a.

14-19 Fig 2 Open Circuited Stub as a Series Resonant Circuit



14-19 Fig 3 Development of a Dipole



Aerial Matching

13. In transferring energy to the surrounding space in the form of electromagnetic radiation an aerial converts energy from one form to another just as a resistor converts electrical energy into heat energy when a current flows through it. The radiation resistance of an aerial is defined as that resistance which, when connected in place of the aerial, would dissipate the same power. For a half-wave dipole the radiation resistance is approximately constant at 73 ohms. In order to achieve maximum transfer of power to the aerial, the characteristic impedance of the transmission line must be matched to the input impedance of the aerial (for an aerial at the resonant length the input impedance is equal to the radiation resistance) and a number of devices, such as the Delta match, are available for doing this. In the Delta match, two points on the dipole are found where the aerial impedance matches that of the twin feeder transmission line at the point of connection. As the voltage between these two points on the dipole is practically zero, the centre of the aerial can be made continuous. The twin feeder transmission line is opened out to fan the delta. Radiation occurs from the delta which upsets the radiation pattern and reduces aerial efficiency.

Directional Properties of Aerials

14. Aerials do not in general radiate isotropically, ie radiate fields of equal intensities in all directions. In the great majority of cases it is desirable to concentrate power in certain directions and to minimize it in others. An important aspect of aerial design concerns the means of distributing (or receiving) power via the aerial in such a way that the performance of a particular communications system shall be optimized, eg long distance point-to-point communication by HF skywave.

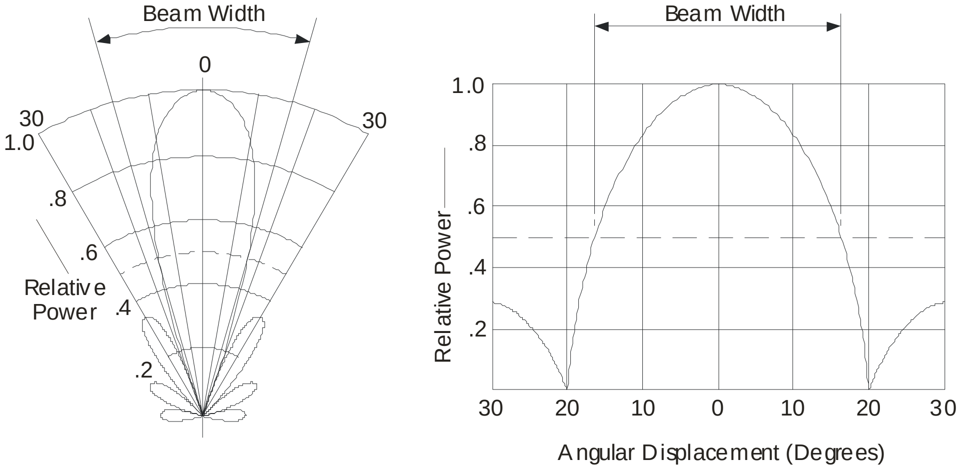
15. The directional properties of aerials may be expressed either in terms of power gain or in terms of angular concentration. The two methods express different aspects of the same property, gain being of primary significance when considering the transmission and reception of power, and angular concentration or beamwidth being significant in applications where direction-defining properties are involved, eg radio direction finding and radar.

a. **Gain**. The power gain of an aerial is normally defined as the ratio of the power it radiates in the direction of maximum concentration to that which would have been radiated from an isotropic source fed into equal power. (More generally, aerial gain may be referred to any defined direction and hence may be greater or less than unity. By definition, the mean gain of any aerial is unity.) Aerial gain is usually expressed in decibels but in calculations involving the communications or radar equations the power ratio must be used.

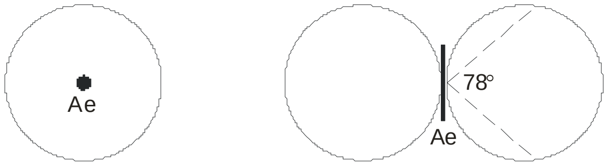
b. **Beamwidth**. The beamwidth of an aerial is defined as being the angle between the two directions in which the power radiated is one-half the power radiated in the direction of maximum concentration. The plane to which the measurement relates must be defined. Referred to field strength, beamwidth is the angle between the directions where the signal is 0.707, ie $1/\sqrt{2}$ of the maximum, since power is proportional to the square of field strength.

16. **Radiation Patterns**. To portray the directional characteristics of an aerial, graphs of relative power (or field strength) as a function of angle are plotted. Polar coordinates are normally used but for highly directional aerials it is difficult to interpret a narrow diagram and cartesian coordinates are used. The two types of radiation pattern graph for an aerial having a beamwidth of 33º are shown in Fig 4. To describe the directional properties fully it is necessary to portray the radiation pattern in two orthogonal planes. The radiation pattern of a dipole, shown in Fig 5, is omni-directional in a plane at right angles to its axis and has a beamwidth of 78º in any plane containing its axis. In most directional aerials the beam has an axis as opposed to a plane, and in radar aerials the beamwidths in orthogonal planes may sometimes be as small as 0.5º.

14-19 Fig 4 Radiation Patterns in Polar and Cartesian Coordinates



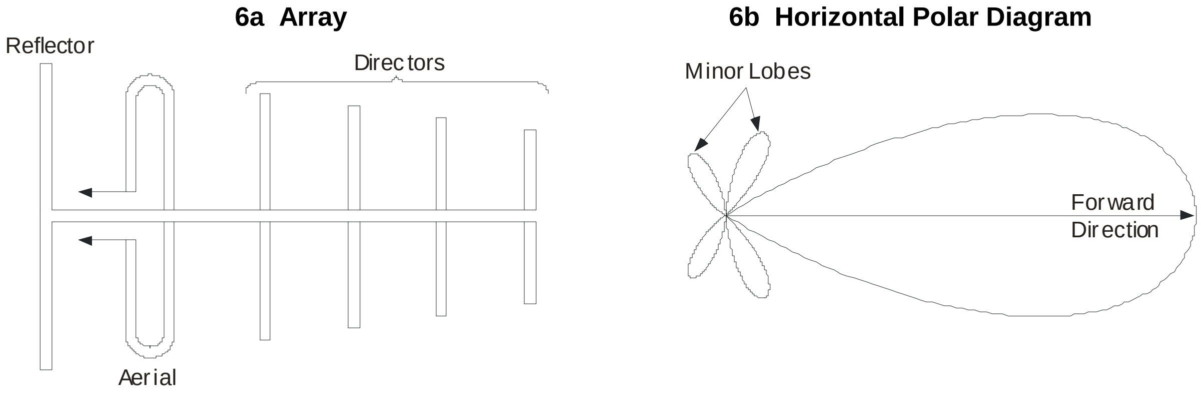
14-19 Fig 5 Field Strength Polar Diagrams of a Half-wave Dipole



17. **Directional Arrays**. The dipole, which has an approximately omni-directional pattern, may be adapted in a number of ways to obtain directional properties. This may be achieved by the addition of parasitic elements, or by the combination of a number of driven dipoles.

a. **Arrays with Parasitic Elements**. Parasitic elements are conducting rods placed near a driven dipole. The transmitted energy of the dipole induces a current flow in the parasitic element, which then re-radiates the energy. By adjusting the lengths and spacings of the parasitic elements, either cancellation or reinforcement may be obtained in a given direction. There are two types of parasitic element, the reflector and the director. The reflector’s length is slightly greater than $\frac{\lambda}{2}$ and when placed at the correct distance from the dipole the reflector cancels the radiation in a direction from the dipole to the reflector, but reinforces the radiation in the opposite direction. Directors are slightly shorter than $\frac{\lambda}{2}$ in length and are placed, relative to the direction of maximum radiation, in front of the dipole. The most common aerial array having parasitic elements is the Yagi array which is formed from a driven dipole with at least one reflector and one director. The beamwidths in the two orthogonal planes are comparable, and by the use of additional directors (normally up to a maximum of about five) beamwidths down to about 30° may be achieved in both planes. The spacing of the elements is theoretically about $\frac{\lambda}{4}$ but in practice is between 0.1λ and 0.15λ. The presence of parasitic elements reduces the radiation resistance of the dipole, and since this is undesirable because of matching problems, the dipole is usually of the ‘folded’ type whose radiation resistance is higher than that of a single dipole, and so the overall radiation resistance is maintained at an acceptable level for convenient matching. A Yagi array with a folded dipole is shown in Fig 6a and its polar diagram in Fig 6b.

14-19 Fig 6 Yagi Array with its Polar Diagram

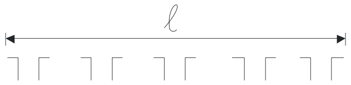
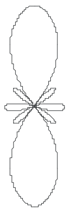
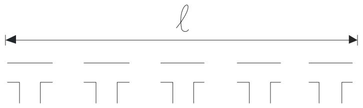
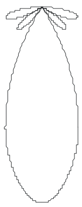
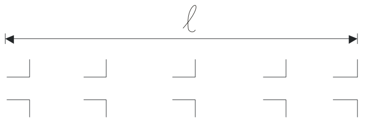
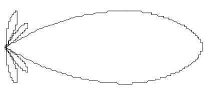
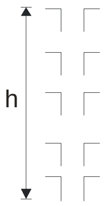
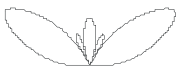


6a Array

6b Horizontal Polar Diagram

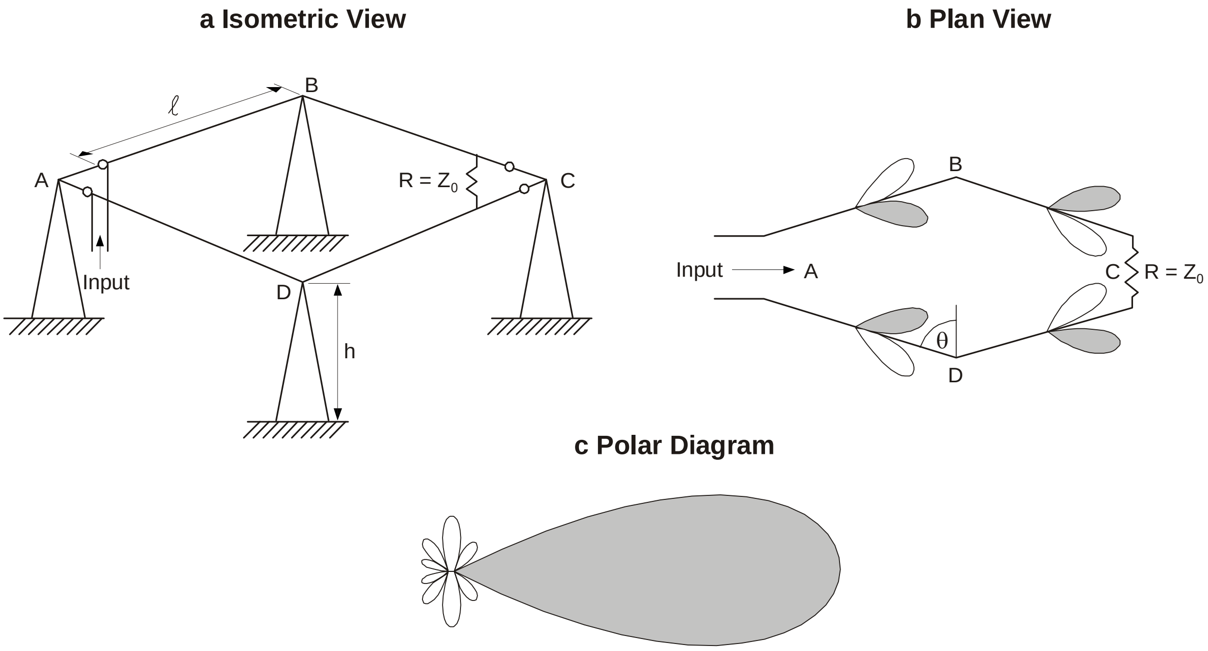
b. **Multiple Driven Arrays**. Directional arrays are often used at HF and above. They have two or more driven elements instead of parasitic elements. Some examples of multi-driven arrays and their characteristics are given in Fig 7. In the broadside array the spacing between the elements is half a wavelength with the result that the phases add in directions at right angles to the array and cancel in directions parallel to the array. The broadside array can be made to radiate in one direction only by placing parasitic reflectors a quarter of a wavelength behind the driven elements. Such arrays give direction only in the plane of the array. It is often desirable to have an array which possesses directional properties in both the horizontal and vertical planes. This can be achieved by stacking a number of broadside arrays one above the other. This was a common technique in early metric radars and is the basis of modern electronically-steered aerials. In the 'end fire' array, the elements are spaced a quarter of a wavelength apart and are fed in phase quadrature. This causes the phases to add in one direction only.

14-19 Fig 7 Characteristics of Driven Arrays

TYPE OF ARRAY	POLAR DIAGRAM (PD)	CHARACTERISTICS
 **a Linear Broadside Array (Plan View)**	 Horizontal PD	Aerials fed in Phase and spaced $\frac{\lambda}{2}$
 **b Linear Broadside Array with Reflectors (Plan View)**	 Horizontal PD	Aerials fed in Phase. Reflectors spaced $\frac{\lambda}{4}$ behind Aerials
 **c End Fire Array (Plan View)**	 Horizontal PD	Aerials fed 90° Phase Lagging and spaced $\frac{\lambda}{4}$
 **d Stacked Array (Side View)**	 Vertical PD	Aerials fed in Phase and spaced $\frac{\lambda}{2}$ Vertically

18. **Travelling Wave Aerials**. The aerials so far considered in these notes have been standing wave (or resonant) aerials, on which the combination of incident and reflected waves forms standing waves. If the wire forming the aerial is terminated in a resistor equal in value to the characteristic impedance of the wire, then there will be no reflected energy and the only waves on the wire will be the incident waves moving towards the termination. These waves are called travelling waves and give this family of aerials its name. Travelling wave aerials are simple to construct and service, they possess good directional properties and, being non-resonant, can be used over a wide band of frequencies without being retuned. This is of considerable advantage when the system employs the sky wave and is required to operate on several widely spaced frequencies during day and night. The most common travelling wave aerial is the 'rhombic' which is widely used by ground stations for transmitting and receiving at HF. It consists of four wires forming a diamond or rhombus shape in the horizontal plane, as shown in Fig 8. It has a very wide bandwidth and a high gain. The direction of maximum gain is along the major axis of the rhombus towards the terminating resistor. The exact shape of both the vertical and the horizontal polar diagrams is determined by the leg length (ℓ), the tilt angle (θ) and the height of the aerial above the ground (h). For example, as leg length increases, the beamwidth narrows and the angle of elevation decreases.

14-19 Fig 8 The Rhombic Aerial



a Isometric View

b Plan View

c Polar Diagram

Aerial Aperture

19. A useful aerial parameter, which is related to gain, is the aerial receiving aperture. Aperture may be regarded as the effective area presented by a receiving aerial to the incident wave. The basic relationship between gain (G) and aperture (A) can be expressed as:

$$\frac{G}{A} = \text{Constant} \;= \frac{4\pi}{\lambda^2}$$

For a half-wave dipole, which has a gain of 1.635, the effective receiving aperture is approximately equal to $\frac{\lambda^2}{8}$.

Image Aerials

20. In the basic description of aerials it is assumed that the aerials have always been placed so far above the earth's surface that any reflected energy from the surface (due to the original radiation from the aerial) is negligible. In practice aerials are seldom so far above the earth's surface, or the skin of an aircraft, and the actual directional characteristics exhibited are due to the vector summation of the direct and the reflected waves.

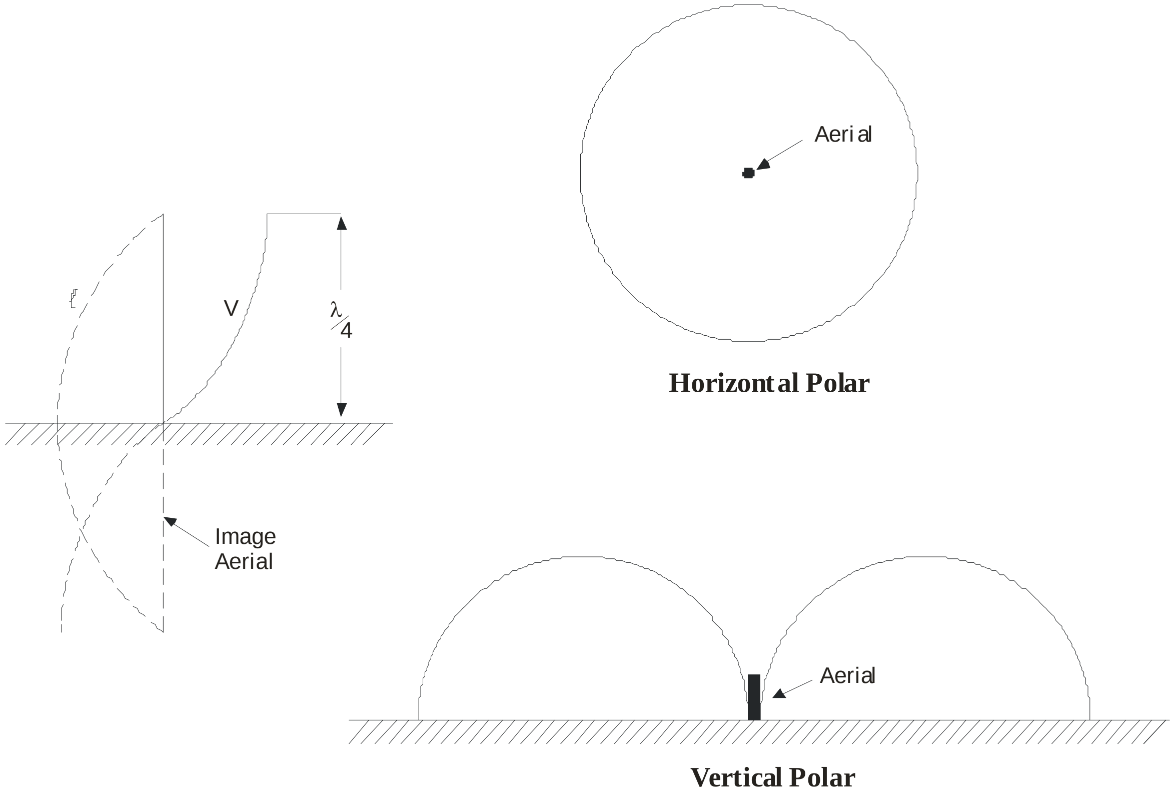
21. For the purpose of calculation, it is convenient to consider that the reflected wave is generated, not by reflection, but by an 'image' aerial located below the surface of the earth. The image aerial is so chosen that the joint action of the actual aerial and its image produces the same conditions in the space above the earth as exists with the actual aerial in the presence of the earth.

22. If the earth is considered to have a reflection coefficient of approximately unity, together with a 180° phase change, then the currents in the corresponding parts of the actual and image aerials are of the same magnitude.

VLF and LF Aerials

23. The major problem encountered in the design of aerials for operation in the VLF and LF bands is the physical size of the elements involved. For example, at 100 kHz the wavelength is 3 km (nearly 10,000 ft) and so a half-wave dipole is out of the question. The Marconi aerial consists of a single conductor a quarter-wavelength long. It is mounted vertically and the transmission line is connected between its lower end and the Earth. The Earth, acting as a reflecting surface, creates an image of the aerial and the whole system behaves as if it were a half-wave dipole. The conductivity of the Earth surrounding a Marconi aerial must be high and if, as in the case of dry sandy soil, this is not so, then a metal mesh round the base of the aerial, called a counter-poise, is necessary. The polar diagrams for this type of aerial are shown in Fig 9. It should be noted that the radiation resistance is about half that of a half-wave dipole, i.e. about 36 ohms.

14-19 Fig 9 Marconi Quarter-wave Aerial



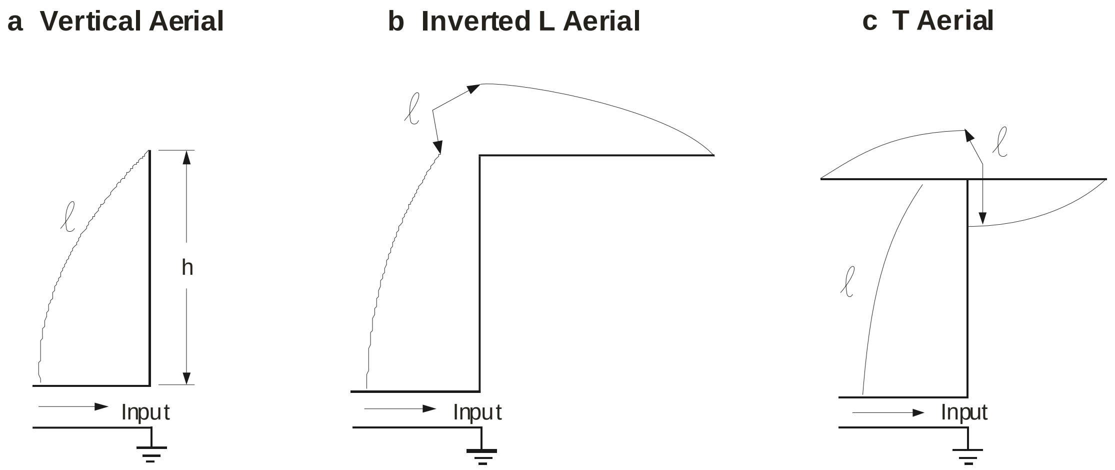
Horizontal Polar

Vertical Polar

24. The Marconi aerial is not the complete answer however at VLF and LF since structural problems restrict heights in general to about 300 ft. Masts for aerials operating in these bands can therefore be only a fraction of a wavelength high and the efficiency of the aerials is correspondingly low. Furthermore, directional arrays would requir massive structures and so omni-directional systems are generally used at these frequencies.

25. One method of increasing the radiation without increasing the mast height is shown in Fig 10. The current distribution for an aerial of height h is shown at a. By connecting a horizontal wire to the top of the aerial, the current distribution in the vertical (and most important) portion is as shown at b and the 'effective' height is thereby increased. This L aerial is very common at VLF and LF. A variation, known as the T aerial, is shown at c.

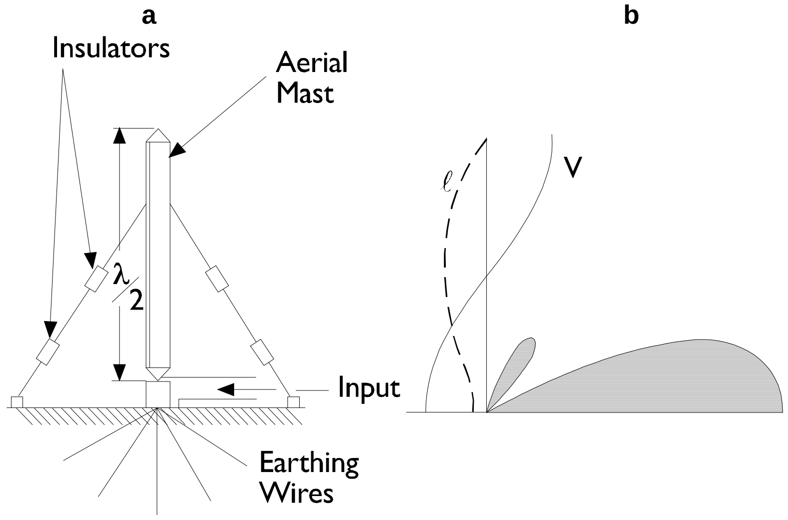
14-19 Fig 10 VLF and LF Aerials



MF Aerials

26. At the lower end of the MF band the techniques employed at VLF and LF have to be used. Towards the upper end of the band, aerials half a wavelength long can be used. Fig 11 shows a typical half-wave vertical MF aerial. It is insulated at its base and energized with respect to an earth conductor system. The current at the base is therefore small and ground losses are low. The radiation resistance is high compared with the total loss resistance and so a high efficiency is obtained. The physical height of the aerial can be reduced by 5 to 10%, without affecting the radiation pattern to any marked degree, by the addition of a capacitance 'hat' on the top of the aerial.

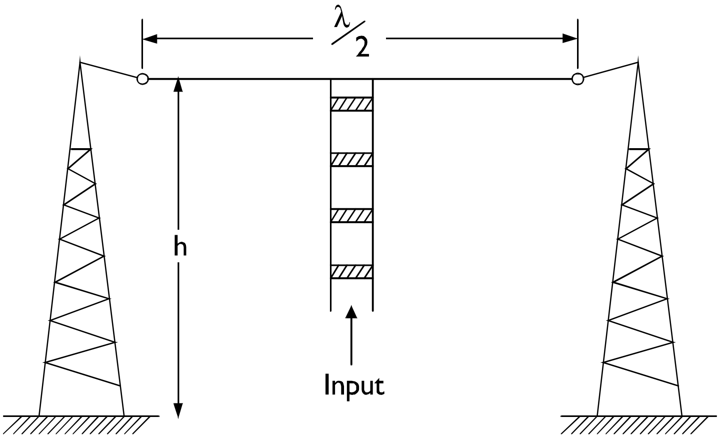
14-19 Fig 11 Half-wave Vertical MF Aerial



HF Aerials

27. Communication in the band of frequencies between 3 and 30 MHz generally employs sky wave propagation. At the lower end of the band the simplest form of aerial used by ground stations is the horizontal half-wave dipole shown in Fig 12. It is rigged between two support masts and uses the Earth as a reflector. Maximum gain is obtained at high angles of elevation and no ground wave is present. This aerial is thus suitable for ionospheric communication using horizontally polarized waves.

14-19 Fig 12 Horizontal Centre-fed Half-wave Dipole for HF



28. In the HF band, increased ranges can be obtained by using highly directional arrays with the main beam directed at a shallow angle to the ionosphere. Rhombic and multi-drive arrays are used for this purpose. An example of the latter is the curtain array which consists of pairs of half-wave end-fed horizontal dipoles stacked vertically, the whole being suspended from two or more lattice masts. Reflectors are usually suspended behind the dipoles.

VHF and UHF Aerials

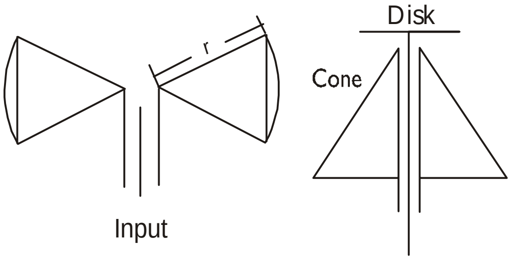
29. Aerials for use in the VHF and UHF bands are generally made of hollow aluminium or copper tubing, and as the diameter of the elements can be increased, so the bandwidth becomes greater. Yagi arrays are in common use in these bands, particularly as receiving aerials, but since the wavelength is relatively small, aerials which are large in terms of wavelength and which often possess elaborate reflecting systems can be used. In these bands there is the problem of interference between the direct and reflected waves from the ground, but this can be minimized by placing the aerial on a mast at a height of several wavelengths from the ground.

30. Transmission in these bands is almost solely by means of the space wave and therefore the range depends on the heights of the transmitter aerials. Horizontal or vertical polarization can be employed. Some of the most common aerials at VHF and UHF are:

a. **Biconical and Discone Aerials**. The biconical and discone aerials, shown in Fig 13, are widely used on UHF ground installations employed on air-to-ground communications. They will operate efficiently over the band for which the cone slant length, r, is between 0.25λand λ, thus giving frequency bandwidths of the order of 4:1 and greater. The cones are sometimes made in the form of a wire cage to reduce weight and wind resistance.

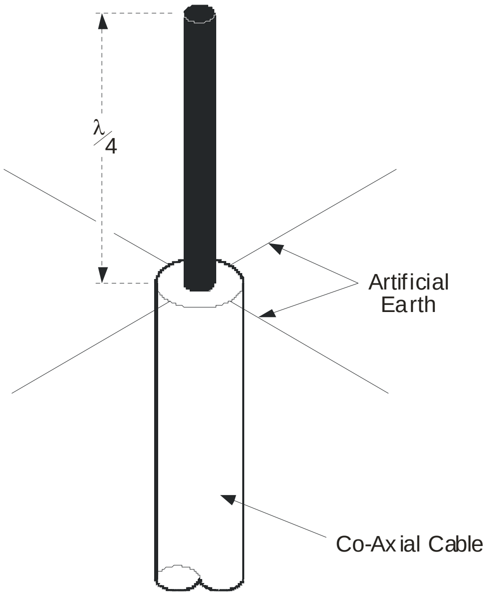
b. **Quarter-wave Ground Plane Aerial**. As shown in Fig 14, this aerial consists of a 0.25λ radiating element with four radial rods joined to the outer conductor of the coaxial feeder and forming an artificial earth. The aerial gives all-round radiation in the horizontal plane, and in the vertical plane the angle of elevation is determined by the length of the rods.

14-19 Fig 13 Wide Band UHF Aerials



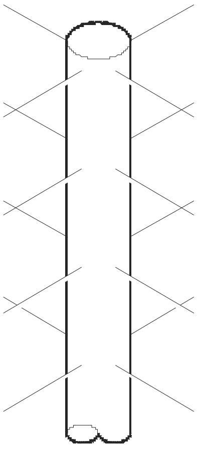
a Biconical Aerial **b Discone Aerial**

14-19 Fig 14 The Quarter -wave Ground Plane Aerial



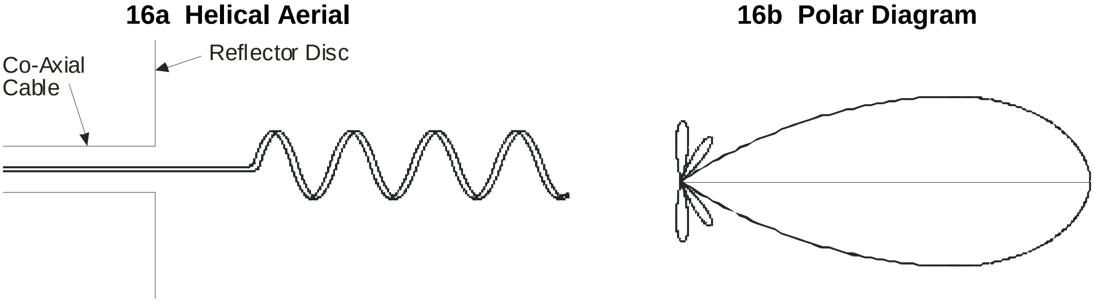
c. **Turnstile Aerial**. The turnstile aerial, shown in Fig 15, gives all-round horizontally-polarized radiation in the horizontal plane suitable for broadcast transmission. The beamwidth in the vertical plane is reduced by stacking the crossed aerials in tiers. This increases horizontal radiation and decreases vertical radiation.

14-19 Fig 15 The Turnstile Aerial



d. **The Helical Aerial**. The helical aerial is of the travelling wave type. It provides a circularly-polarized beam of radiation which may be required in applications where the orientation of the receiving aerial cannot be controlled, e.g. command guidance link to a guided missile. Its construction is shown in Fig 16. The diameter of the helix and the spacing between the turns must be carefully chosen to obtain the required polar diagram.

14-19 Fig 16 The Helical Aerial



Aircraft Aerials

31. **LF, MF and HF**. Aerials employed on aircraft for use up to about 30 MHz are invariably restricted to dimensions which are shorter than one wavelength. In older aircraft carrying HF equipment this was partially overcome by a trailing wire of about 50 metres in length. An alternative is a simple wire aerial such as an inverted L or T rigged between the top of the tail fin and a small mast near the cockpit. When used for transmitting at relatively high altitudes, these wire aerials are prone to corona troubles due to the high voltage gradients combined with the low air pressure. When used for receiving they are liable to have a high noise level due to a phenomenon known as precipitation static. Higher aircraft speeds and the need to reduce precipitation static have necessitated the design of aerials known as 'suppressed aerials'. Suppressed aerials fall into two main classes:

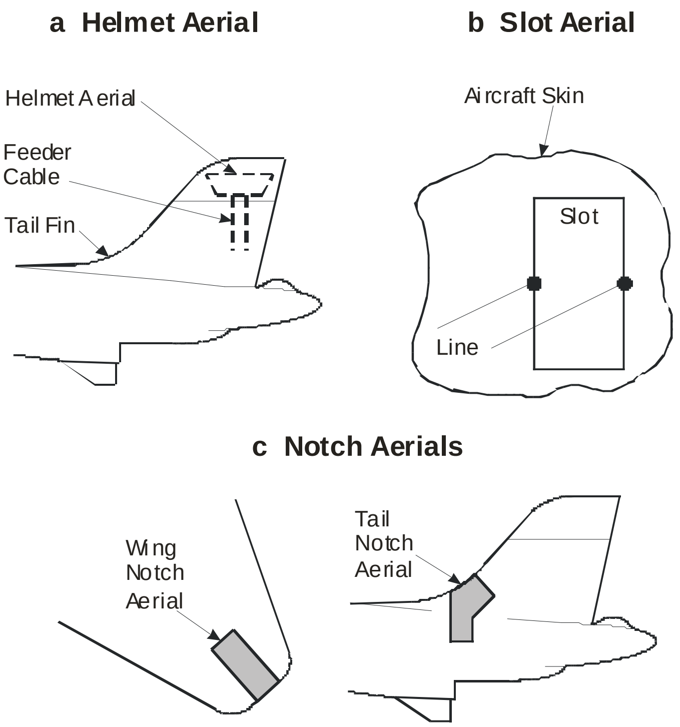
a. **Concealed Aerials**. In many cases these are conventional aerials but they are so constructed and mounted that they are concealed inside the aircraft. Typical of these are radar scanners mounted behind a radome nose section and the general purpose communication aerial of the 'helmet' type mounted on top of the tail fin and enclosed in a plastic cover. A helmet aerial is shown in Fig 17.

b. **Aerials Using the Aircraft Skin**. This type of suppressed aerial forms a resonant cavity within the aircraft and uses the metal skin of the aircraft as the radiating element. It may take one of two forms:

(1) **A Notch Aerial**. This is a notch cut out of the wing or tail and covered with insulating material to prevent the notch affecting the aerofoil. The RF power from the transmitter is fed to both sides of the notch and the RF voltage set up across the notch causes currents to flow in the outside skin of the aircraft and an electromagnetic wave is radiated. The notch aerial is shown in Fig 17.

(2) **A Slot Aerial**. This is a half-wave length long slot cut in the aircraft skin. It differs from the notch in that the slot is resonant and functions in a manner similar to that of a half-wave dipole. The slot aerial is shown in Fig 17.

14-19 Fig 17 Suppressed Aerials



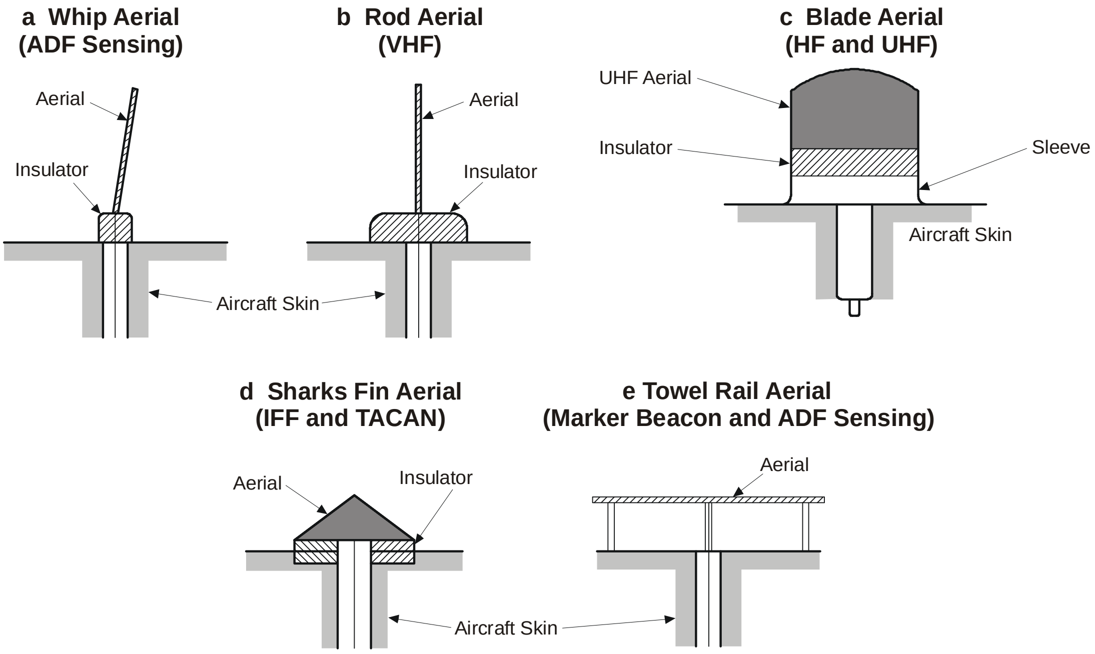
a Helmet Aerial

b Slot Aerial

c Notch Aerials

32. **VHF and UHF Aerials**. Aerials for aircraft use at VHF and UHF are generally easier to design than aerials for lower frequencies, simply because the wavelength is shorter. The most common types, with examples of their uses, are shown at Fig 18.

14-19 Fig 18 VHF and UHF Aerials



a Whip Aerial (ADF Sensing)

b Rod Aerial (VHF)

c Blade Aerial (HF and UHF)

d Sharks Fin Aerial (IFF and TACAN)

e Towel Rail Aerial (Marker Beacon and ADF Sensing)

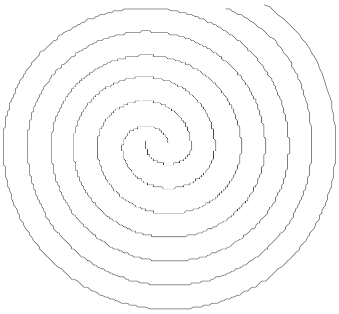
Broad Band Aerials

33. In many applications it is necessary to have an aerial system whose performance remains virtually constant over a wide band of frequencies. Aerials based on conical shapes have a tendency to be broad banded. The discone aerial and the biconical aerial used at UHF (described in para 30) are

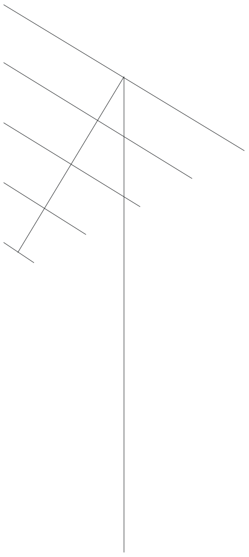
examples of this. Unfortunately, the finite size of such aerials introduces end effects which limit the range of frequency insensitivity. Spiral aerials (Fig 19) are used extensively in passive radar direction finding and semi-active missiles. By far the most successful broad band aerial is the log-periodic type (Fig 20). In the log-periodic aerial the distances between the active elements are designed to conform to a logarithmic expansion. Aerials of this type are in use for long distance HF communications and permit transmission with equal efficiency over the entire HF spectrum. They thus have bandwidths (ratio of highest and lowest frequency) of at least 10:1. Log-periodic aerials also find applications in ECM and radar.

34. The bandwidth of an aerial is also determined by its thickness. The thicker the aerial, the wider is the bandwidth.

14-19 Fig 19 Spiral Aerials



14-19 Fig 20 Log-periodic Aerial



CHAPTER 20 - MICROWAVES

Introduction

1. Microwaves are a form of electromagnetic radiation with wavelengths approximately midway between those of light and radio waves, ie between 30 cm and 1 mm - with equivalent frequencies from 1 to 300 GHz. As with most divisions of the electromagnetic spectrum, these boundaries are somewhat arbitrary with no physical significance. Microwaves exhibit many of the same properties as both light and radio. The principle military applications of microwaves are in communications and radar. In radar systems, good azimuth resolution results from a narrow beamwidth. Short wavelengths mean that this can be achieved with relatively small aerials, which is especially desirable in airborne systems. In communications, the high frequencies allow high bandwidth signals to be transmitted. Microwaves are typically used for satellite communications links.

2. In order to utilize the microwave region of the spectrum, it is necessary to have devices which are capable of generating such frequencies and of amplifying them to useful power levels. Unfortunately, the conventional amplifiers and oscillators fail to perform satisfactorily at frequencies above 1 GHz, and it has been necessary to develop specialized devices which can operate at these frequencies. The properties of resonant cavities and the principle of velocity modulation are fundamental to an understanding of these devices.

Resonant Cavities

3. At UHF, a short-circuited section of transmission line, called a 'lecher bar', may be used as the tuned circuit of an oscillator. Fig 1a shows the instantaneous electric and magnetic fields and the current in an oscillating lecher bar. The directions of the current and the fields change each half cycle. The relationship between the resonant frequency (f_0), the inductance (L), and the capacitance (C) is given by:

$$f_0 \propto \frac{1}{\sqrt{LC}}$$

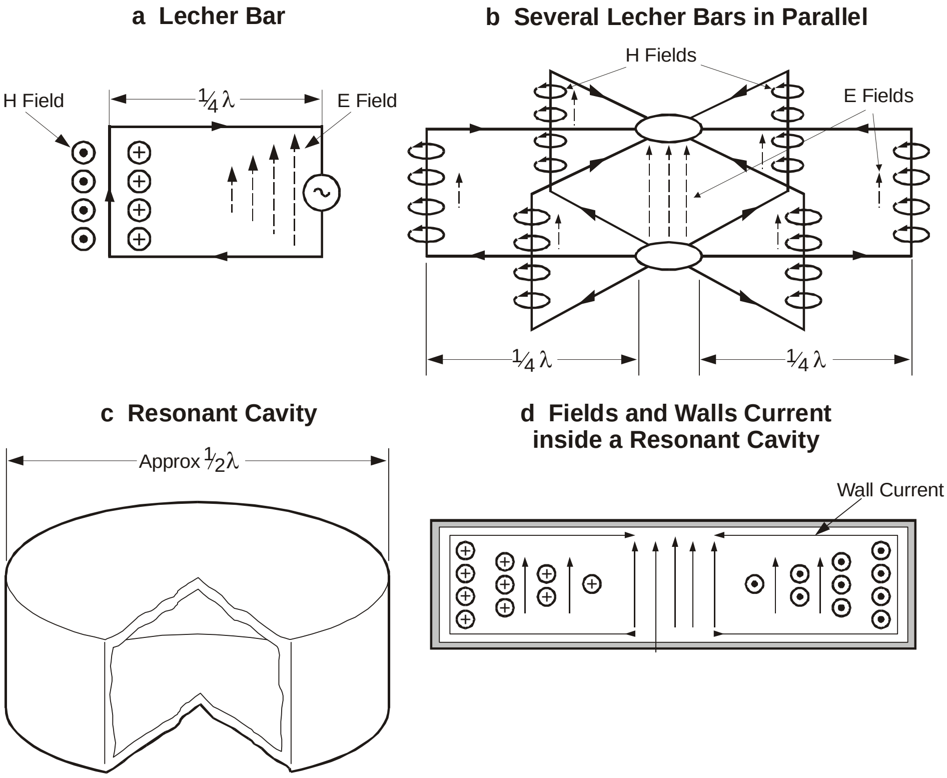
If several lecher bars are connected in parallel (Fig 1b), the inductance is reduced by the same amount as the capacitance is increased; thus, the resonant frequency remains the same as for a single lecher bar. The total resistance is, however, decreased. If an infinite number of lecher bars are placed in parallel, a hollow cylinder or cavity is formed (Fig 1c). Little radiation can occur and, as the skin resistance is small, total losses are small.

4. The internal diameter of the cavity, which should be approximately half a wavelength, determines its resonant frequency. At centimetric wavelengths, the cavity dimensions are small enough for it to form an integral part of an amplifier or oscillator.

5. The instantaneous electric (E) and magnetic (H) lines of force present when a lecher bar oscillates alternate at a high frequency. The E lines may be associated with the circuit capacitance and the H lines with the circuit inductance. In a resonant cavity formed from an infinite number of lecher bars, the individual H lines round each lecher bar combine as in a solenoid and form closed loops, strongest around the circumference of the cavity and weakening to zero at the cavity centre. This H field always lies parallel to the cavity walls. The E lines combine to form a strong E field at the centre of the cavity, decreasing to zero at the cavity walls. During each half cycle the E field builds up to a maximum at the instant that the H field falls to

zero. During the next half cycle the opposite happens, i.e. the E and H fields are 90º out of phase, as are the energies associated with the capacitance and inductance of a conventional tuned circuit.

14-20 Fig 1 Development of a Resonant Cavity

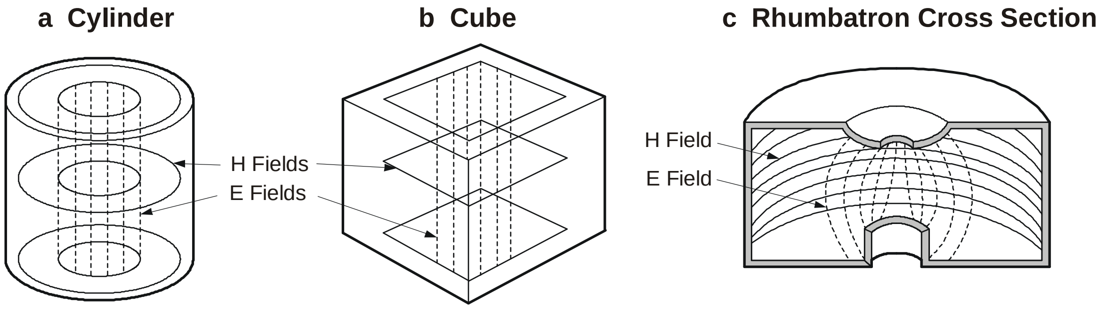


6. The fields shown in the cavity cross-section of Fig 1d are the fields at one instant of time. The electromagnetic energy in the cavity is oscillating at the cavity resonant frequency, the E and H fields changing direction every half cycle. The changing H field induces voltages in the cavity walls which produce wall currents. These wall currents are of the same frequency as, and always flow at right angles to, the H field. Due to skin effect, the currents flow only on the inner surfaces of the cavity and the outer surfaces can therefore be earthed without affecting the cavity operation.

7. A resonant cavity need not be cylindrical. Depending on the function of the cavity, it may be rectangular, spherical, or a modified cylindrical shape. In these cases, the field pattern (or 'mode'), which is formed inside the cavity when it is excited into oscillation, will differ from that shown in Fig 1d. Nevertheless, the E field will always be zero at the cavity walls, and the E and H fields will always be at right angles to each other and 90º out of phase.

8. Some common cavity modes are shown in Fig 2;the rhumbatron cavity (Fig 2c) forms the resonant cavity in the klystron. Electrons are passed through holes in the roof and floor of the cavity and, in order that the transit time of the electrons passing through the E field is short, the shape of the cavity is modified from a simple cylindrical shape, as shown.

14-20 Fig 2 Common Cavity Models



Velocity Modulation

9. In a conventional amplifier, the device current is modulated in response to changes in the voltage existing in the input circuit; thus a sinusoidal input results in a device current that also varies sinusoidally. By making this current pass through a resistance in the external circuit, an output may be obtained which is identical in form to, but larger in amplitude than, the input. This amplitude modulation technique works well at VHF and below, but becomes progressively less effective as frequency increases, such that it is unusable above about 1 GHz. At these higher frequencies, a different technique called 'velocity modulation' is used.

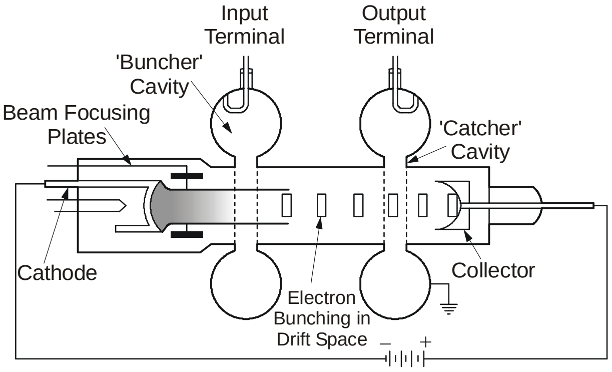
10. The principle behind velocity modulation is that of varying the velocity of the individual charge carriers that make up the device current, rather than their numbers. This results in bunches of charge being formed within the device. Using suitable circuitry, it is possible to use the bunch formation to produce an amplified output in a separate circuit. This technique is used in the klystron.

HIGH POWER AMPLIFIERS

The Klystron Amplifier

11. A schematic diagram of a klystron amplifier is shown in Fig 3. An evacuated tube contains an electron gun at one end. This sends a narrow beam of electrons to an anode collector at the other end. The beam passes through the narrow necks of two cavity resonators. The first cavity is known as the 'buncher', and it is at the same potential as the second cavity, the 'catcher'. Oscillations are induced in the buncher by inputting an RF signal by means of a short coaxial cable.

14-20 Fig 3 Klystron Amplifier – Schematic



12. The oscillating cavity generates an alternating electric field in the region through which the electron beam must pass; this field exerts a force on the electrons in the beam. Since the field is alternating, it accelerates electrons passing through the cavity during one half cycle, and decelerates electrons passing through during the other half cycle, i.e. it imposes velocity modulation. The result of the modulation is that the 'fast' electrons tend to catch up with the 'slow' electrons as they pass down the tube, with the result that electron bunches are formed at some distance from the buncher cavity (Fig 3). The second (catcher) cavity is placed so that the bunches are most pronounced just as they pass through it and, in doing so, they induce in it a strong oscillatory electric field, which has a polarity such as to slow down a bunch, thereby extracting energy from it. Half a cycle later, the direction of the electric field across the gap has reversed, and this field acts as an accelerator, but on the low density, non-bunched portion of the beam. Thus, many more electrons are slowed down than are accelerated, and there is a net exchange of energy from the beam to the catcher's output RF circuit, i.e. amplification.

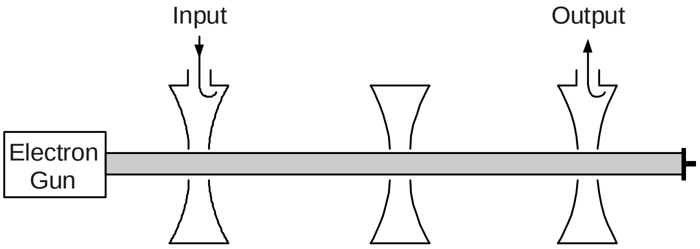
13. Klystrons of this type are readily available at I-band and above. The technique may also be used at frequencies below the microwave region. In the frequency range of 100 to 1,000 MHz, the size of the resonant cavity is such that quite large gap diameters can be used, together with similarly large diameter electron beams; this permits high powers to be achieved.

14. Typically, the two-cavity klystron described achieves power gains of 10 to 15 dBs, exhibits a rather low efficiency of about 40% and, in addition, has a narrow bandwidth (1 to 2%). In order to achieve better gains and efficiency, additional cavities are introduced.

Multi-cavity Klystrons

15. It would, in theory, be possible to achieve higher gains by taking the output from one klystron and using it as the input to a second. This technique is used, in practice, by combining the two klystrons into one tube, as shown in Fig 4. The middle cavity is placed at the optimum distance from the buncher and acts as an additional velocity modulator. A gain of 40 dB is typical for a three-cavity device, but even higher gains can be achieved by introducing further cavities. Although, in principle, the number of cavities could be increased to any number, commonly either 4 or 6 cavities are used, giving gains of 50 and 110 dBs respectively, with efficiencies as high as 65%. The bandwidth may be widened, to a limited extent, by tuning each cavity to a slightly different frequency - a technique known as 'stagger tuning'.

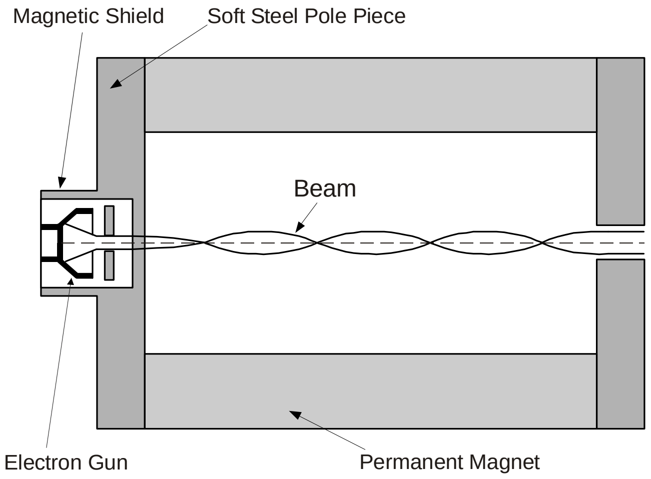
14-20 Fig 4 Three-cavity Klystron Arrangement



16. Multi-cavity klystron amplifiers are available at frequencies from about 400 MHz to 200 GHz. Peak powers produced range from a few kW to about 400 MW; average powers range from a few mW to about 175 kW. High power devices are, however, large and heavy and are not suitable for airborne applications. The kinetic energy of the electrons is converted into heat when they strike the collector, necessitating water cooling in some high power klystron amplifiers. This heat represents wasted DC energy and, to improve efficiency, the collector voltage may be decreased so as to reduce the electrons' kinetic energy.

17. **Beam Focusing**. In airborne applications, the weight of the device is important. The electron beam consists of negative charge, and repulsive forces will tend to spread the beam as it passes down the tube. The traditional way of preventing this is to provide a strong axial magnetic field by using a permanent magnet, as shown in Fig 5. The magnet is larger, and much heavier, than the rest of the klystron, typically 15 kg for a 1.5 kW device operating at 10 GHz. The use of samarium cobalt as the magnetic material is one way of reducing this weight by a factor of as much as 7 or 8; the material is, however, expensive. An alternative is to use a technique known as periodic permanent magnetic (PPM) focusing, which replaces the large magnet with a number of small ones.

14-20 Fig 5 Klystron Beam Focusing using Large Permanent Magnet

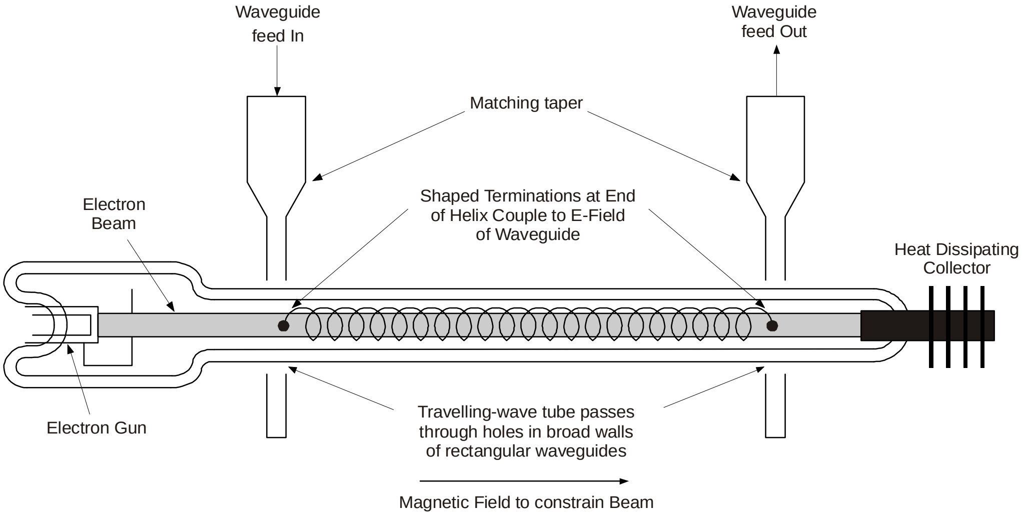


The Travelling Wave Tube (TWT)

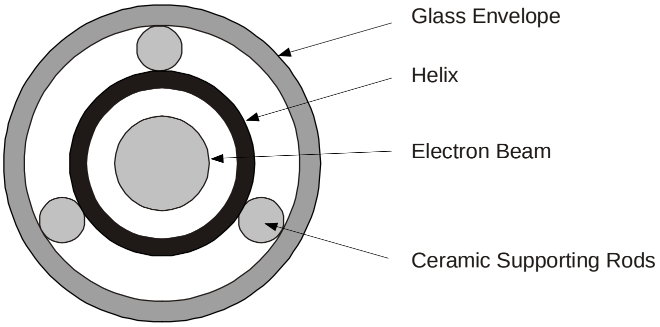
18. The main drawback of the klystron amplifier is its narrow bandwidth, resulting from the use of resonant cavities. The travelling wave tube (TWT) overcomes this problem. The principle of velocity modulation and charge bunching is still used, but the buncher and catcher cavities are replaced by a helix of wire; the arrangement is shown in Fig 6. Amplification is achieved by the exchange of energy between the electron beam and an electromagnetic wave travelling in the metal helix.

14-20 Fig 6 Travelling Wave Tube Arrangement

6a - Schematic


Waveguide feed In
Waveguide feed Out
Matching taper
Electron Beam
Shaped Terminations at End of Helix Couple to E-Field of Waveguide
Heat Dissipating Collector
Electron Gun
Travelling-wave tube passes through holes in broad walls of rectangular waveguides
Magnetic Field to constrain Beam

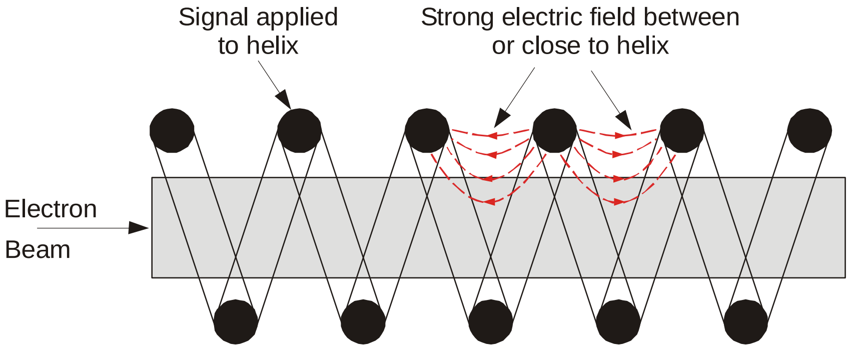
6b – Axial View



19. The electron beam passes down the centre of the tube and is kept in place by an axial magnetic field, generated either by a solenoid or by permanent magnets. The input signal is fed onto the helix at the electron gun end and it moves along the helical wire in the form of a voltage travelling wave; in principle, the same process as the sending of a signal down a transmission line. The signal travels around the wire of the helix at very nearly the speed of light. However, because of its greatly increased path length, its progress along the length of the tube is very much slower. The device is designed such that the travelling wave progresses along the tube at the same rate as the electrons in the electron beam and, since the travelling wave carries with it its own electric fields, the electrons in the beam and the travelling fields are able to interact with each other all the way along the tube.

20. Fig 7 shows a diagram of a portion of the helix and electron beam; examples of the alignments of two of the fields are illustrated. Any electron travelling down the tube will have alternating accelerating and decelerating forces applied to it. As with the klystron, velocity modulation and consequent charge bunching is achieved and, since the electrons travel with the field for the length of the tube, this bunching becomes more marked as the beam progresses.

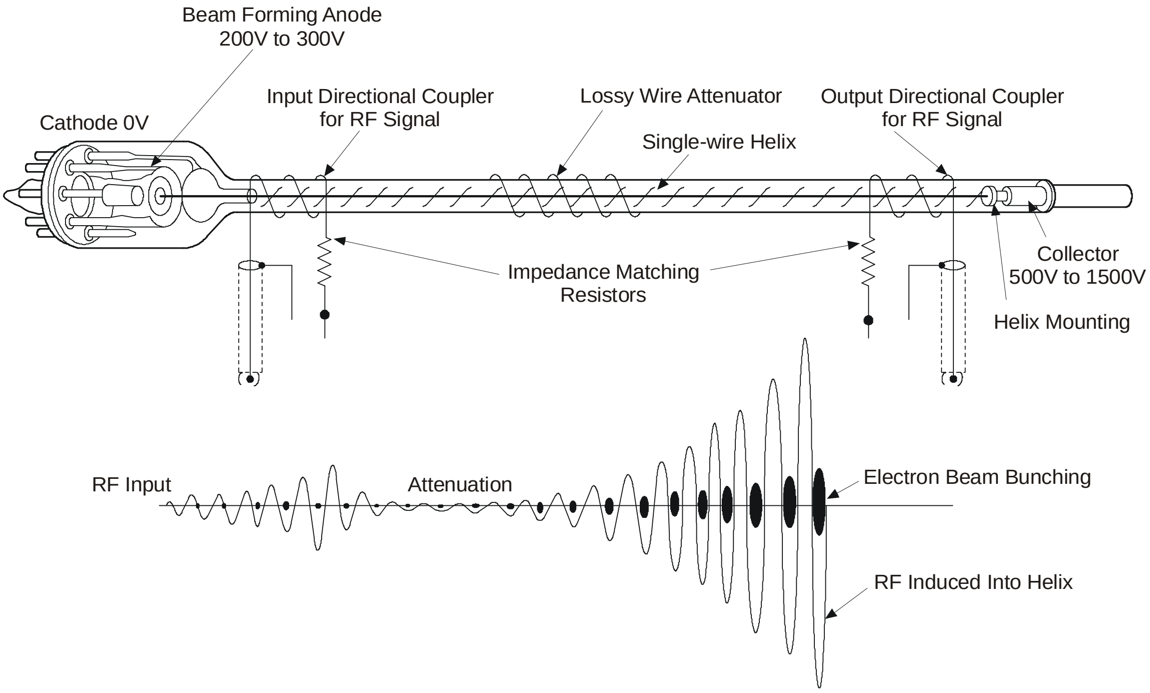
14-20 Fig 7 Travelling Wave Tube - Section of Helix Showing Electric Field Pattern



21. The electron bunches induce electric fields, and hence voltages, into the wire of the helix as they move down the tube. This voltage enhances the signal already present, ie amplification is achieved.

22. **Feedback**. Since the input and output are directly connected, it is possible for part of the output to be reflected back down the helix to the input, which can cause instability. It is, therefore, normal to place an attenuator near the wire to reduce any reflection. Although this also attenuates the proper signal, the electron bunches are not affected and proceed to re-induce a voltage in the wire to amplify the signal (Fig 8).

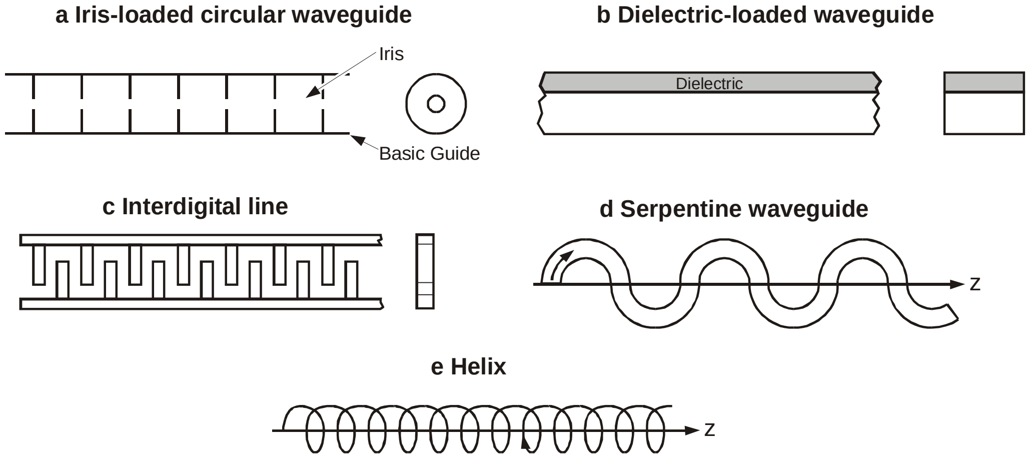
14-20 Fig 8 Travelling Wave Tube with Attenuator



23. **Slow-wave Structures**. The helix of wire in the TWT is designed to slow down the travelling wave to a speed along the tube commensurate with the electron beam. As such, it is an example of a slow-wave structure, and is the commonest structure used, but not the only one. Other forms of slow-wave structures are shown in Fig 9; all work on the principle of extending the path of the travelling wave.

24. **Frequency Range, Bandwidth, and Output Power**. TWTs exist at operating frequencies from 200 MHz to 50 GHz. Output power is very much lower than that of a multi-cavity klystron at, typically, up to a few kW; gains are in the order of 30 dB. The advantages of the TWT lie in its wide bandwidth (typically an octave), small size (length 30 to 60 cm), and low weight (a few pounds with PPM focusing). This makes it a particularly suitable device for airborne use, especially for frequency agile or jamming applications. It is also characterized by low noise and this makes it ideal as a communications amplifier.

14-20 Fig 9 Examples of Slow-wave Structures



The Twystron

25. The multi-cavity klystron is characterized by high power, but narrow bandwidth, whereas the TWT gives wider bandwidth but with much lower power. Some applications, notably wide-band, high-power jamming, require the best features of both of these devices. The twystron aims to provide these characteristics by replacing the output cavity of a klystron with the output section of a TWT. The twystron is, therefore, a hybrid of the two devices and is able to provide peak powers in the MW range, bandwidths of 6 to 15%, and an efficiency of about 30%.

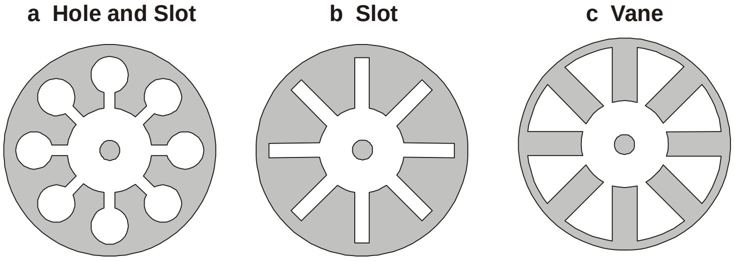
HIGH POWER OSCILLATORS

The Magnetron

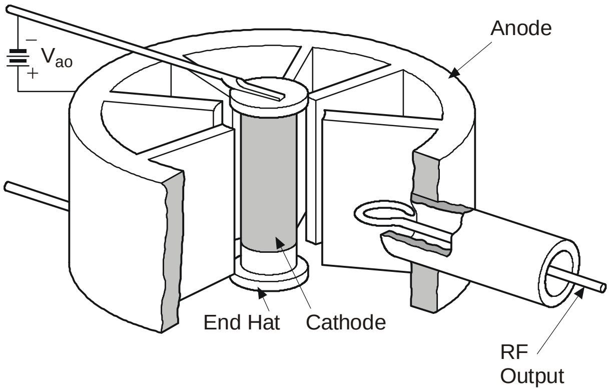
26. The high power oscillator in most radars is a device known as a 'magnetron'. It relies for its operation on the manner in which charged particles move in a magnetic field, experiencing a force at right angles to both the magnetic flux lines and their own motion. In a magnetron, electrons are emitted by a hot, cylindrical cathode, and are attracted by the high positive voltage of the anode, which surrounds the cathode in the form of a hollow cylinder. A magnetic field is established along the axis of the structure, which causes the electrons to move in a curved path towards the anode, the amount of curvature being dependent on the strength of the magnetic field. In practice, the field is arranged so that the electrons reach the anode after six or seven revolutions. It is the resulting motion of the electrons around the cathode that is used to produce oscillations

27. In a practical magnetron, the anode block is cylindrical in form, but is modified to include resonant cavities on its inner surface. Fig 10 shows some typical cavity shapes and Fig 11 shows a cutaway to illustrate the 3-dimensional structure.

14-20 Fig 10 Typical Magnetron Anode Block Structures

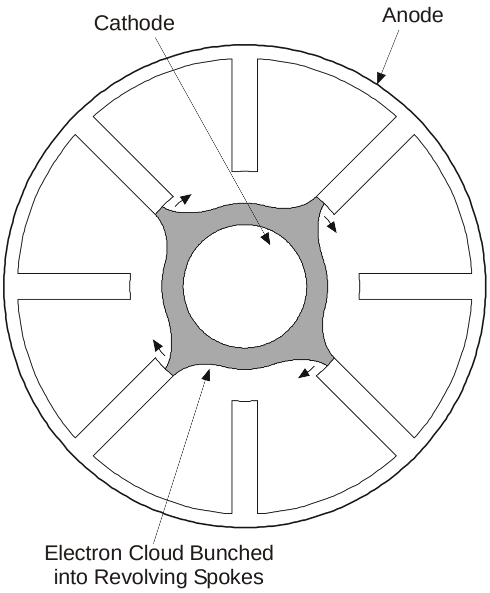


14-20 Fig 11 Cutaway Diagram of Magnetron



28. As the electrons spiral around the cathode, they pass the openings of the cavities and, if one of the cavities is oscillating at its resonant frequency, they will be velocity modulated by the field across its opening. Bunching will be initiated in the revolving electron cloud, and this in turn induces oscillations in the other cavities. The bunching will eventually reach the initially oscillating cavity; its oscillation will be enhanced and the bunching will become more pronounced. This interaction continues all the while that electrons are spiralling round in the device, the bunching coming to resemble spokes, as shown in Fig 12. In practice, there is always some oscillation present in all of the cavities as the result of thermal noise; this is sufficient to start the device. In order to use the power, it is tapped off from one of the cavities by means of a coaxial feed to a waveguide.

14-20 Fig 12 Electron Bunching in Magnetron



29. It is necessary to arrange the device so that the electron bunch reaches the original cavity at the right moment to enhance the oscillations present there. The normal arrangement, achieved by the choice of DC electric field and magnetic field strengths, is to have the bunch travel at such a rate that it moves from one cavity to the next in half a cycle. In this way, the bunch always experiences a retarding field, giving up its energy to the cavity.

30. The magnetic field is essential for the operation of the magnetron. If the field was removed, the electrons would travel directly outwards to the anode. In this situation, oscillation would cease and the current would rise, since the electrons would arrive at the anode in much greater numbers. With no oscillation, all the DC power would be converted to heat; this would be sufficient to melt the device. To obviate this danger, the magnetic field is always produced by a large and powerful permanent magnet, and this is the most bulky and weighty part of the structure.

31. The operating frequency of a magnetron is determined by the resonant frequency of the cavities, but a limited amount of mechanical tuning is possible by changing the cavity volumes with plungers. Inductive plungers in the cavity are used to raise the frequency, and capacitive plungers at the cavity entrances to decrease frequency.

32. Magnetrons are available, for both pulse and CW operation, with efficiencies between 30% and 60% for pulse operation, but lower for CW. In the pulse system, the magnetron is switched on and off by high voltage pulses (typically 20 kV) supplied by a separate circuit known as a 'modulator'. Output pulses are non-coherent and pulse lengths are normally in the range 0.1 to 5 μsec.

33. The rated duty cycle (pulse length × PRF) for pulse magnetrons is normally in the range of 0.005 to 0.00025, although, in some applications, a high duty cycle, approaching 0.5, is used. For low duty cycles, peak power outputs range from about 5 MW at UHF and D-band, to a maximum of about 200 kW at I-band. CW magnetrons generally have lower power outputs, typically 200 W at D-band to 1 W at G-band.

The Carcinotron

34. The carcinotron is a travelling wave oscillator employing a slow wave structure. The field pattern of an electromagnetic wave travels at a certain velocity, known as the 'phase velocity', whereas the energy which it carries travels at a different velocity, known as the 'group velocity'. In the case of the TWT, these two velocities are in the same direction, but it is possible for them to be sensed in opposite directions. This is the case in the carcinotron, which is, accordingly, classified as a backward wave oscillator (BWO).

35. An electron beam is directed along the structure in the same direction as the phase velocity, and oscillations are possible at a frequency at which the phase velocity of the wave is synchronous with the velocity of the beam. The phase velocity of the electromagnetic wave changes with frequency, and so the frequency of oscillation depends on the velocity of the electron beam. Carcinotrons can therefore be tuned over a wide range of frequencies by varying the voltage of the electron beam. There are two types of carcinotron:

a. **O-type**. The O-type of carcinotron is similar to the TWT, and the slow wave structure is usually a helix. It is normally a low power, low efficiency tube, principally used for receiver and test equipment applications.

b. **M-type**. In an M-type carcinotron, the electron beam moves under the influence of orthogonal electric and magnetic fields. The device resembles a magnetron, in that it has a circular, rather than linear, geometry in order to reduce the size of the magnet. The efficiency of M-type carcinotrons is relatively high, and this, combined with wide range electronic tuning, makes them attractive for countermeasures or jamming applications.

LOW/MEDIUM POWER SOLID STATE DEVICES

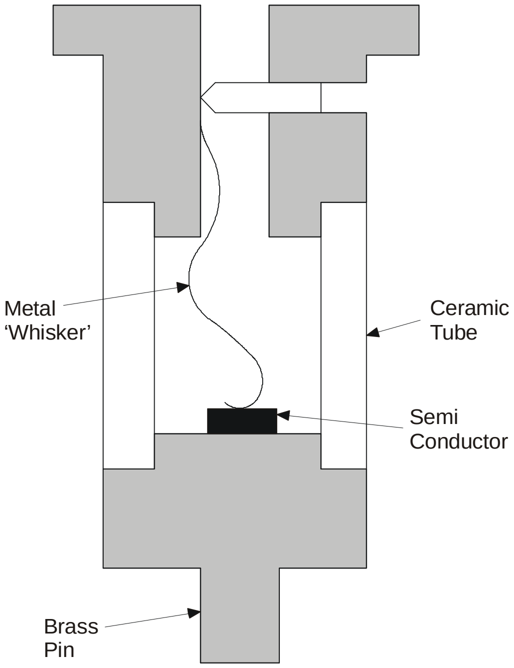
Introduction

36. Whereas the high power requirements for the output stages of radar transmitters are catered for by the thermionic amplifiers and oscillators described above, radar receivers and microwave links in the low and medium power region normally use solid-state devices. These are far less noisy, have lower DC power requirements, and are also smaller, lighter, more reliable, and more efficient.

Point-contact Diode

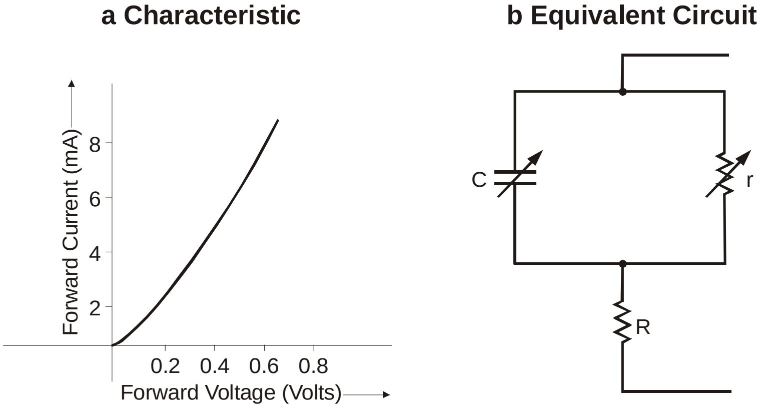
37. The point-contact diode, illustrated schematically in Fig 13, consists of a metal 'whisker' which presses on a small crystal of p- or n-type semiconductor. A diode made of p-type silicon with a tungsten whisker can handle very small signals, and is easily damaged by high power. In some cases, welded or gold bonded contacts are used. Crystals of gallium arsenide can withstand high temperatures and can operate at frequencies up to 100 GHz.

14-20 Fig 13 Point-contact Diode



38. Fig 14 shows the rectifying properties of a point-contact diode and its equivalent circuit. The resistance (r) is the variable 'barrier' resistance of the contact and varies with the applied voltage, thus enabling the diode to be used as a mixer. The contact capacitance (C) also varies with the applied voltage and is very low (approximately 0.2 pF) so that, at microwave frequencies, it does not short circuit r. The series resistance (R) is the constant ohmic resistance of the semiconductor.

14-20 Fig 14 Properties of a Point-contact Diode


a Characteristic

b Equivalent Circuit

39. Point-contact diodes are fairly robust but, when used as microwave mixers, their life depends on the effectiveness of the TR device which protects them from the high transmitter power. The TR cell takes a short time to break down completely after the transmitter pulse has started, and a spike of energy leaks to the mixer diode during this period. If the power in the spike is too high, the metal to semiconductor contact may overheat and burn out. Over a period of time, even comparatively small leakage powers can cause the contact to change its rectifying properties, and the diode becomes inefficient and noisy. This can be detected by a deterioration in the overall performance of the radar, or by a decrease in the back-to-front resistance ratio of the diode.

40. **Backward Diode**. The backward diode is a form of point-contact diode in which the whisker is coated with semiconductor material and bonded to a crystal of n- or p-type germanium. This semiconductor-to-semiconductor contact operates better than the point-contact diode at low signal levels.

Schottky-Barrier Diode

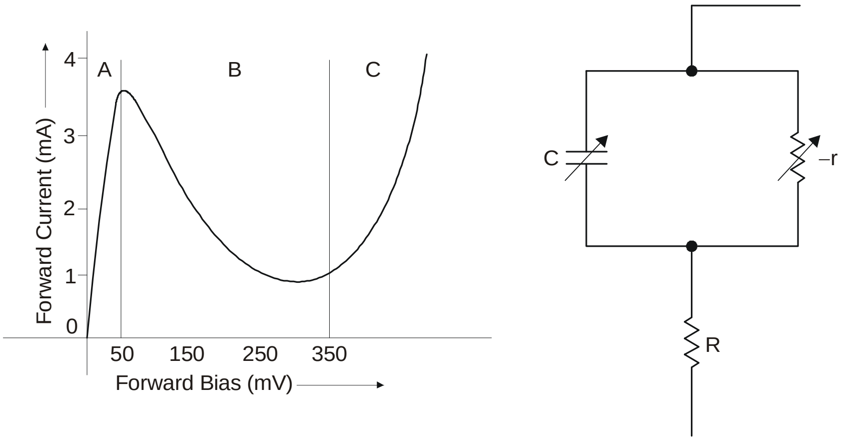
41. The Schottky-Barrier diode uses a metal-to-semiconductor junction, and consists of a thin metal film deposited on a layer of n- or p-type semiconductor. Because of the larger junction area, its capacitance is greater than that of the point-contact or backward diodes (about 1 pF), but it is not easily damaged by high power. Its characteristics are fast switching speed, low forward turn-on voltage, low stored charge, low reverse leakage currents, and high rectification efficiency; hence it is used in circuits for high and low level detection, mixing, UHF modulation, pulse-shaping, voltage damping, and pico-second switching (in computers).

Tunnel Diode

42. The tunnel diode is a semiconductor junction made of heavily doped germanium. It has a very narrow depletion layer between the p- and n-type materials and current-carrying charges can 'tunnel' through the layer at low values of forward bias. From about 0 to about 50 mV, the forward current due to these 'tunnelling' carriers rises fairly sharply (region A in Fig 15a) but, as the voltage is increased above about 50 mV, the tunnelling effect decreases and the forward current falls (region B). At about 350 mV, normal semiconductor action takes over, and the current again increases with an increase in forward voltage (region C).

14-20 Fig 15 Some Properties of a Tunnel Diode

a Characteristic **b Equivalent Circuit**



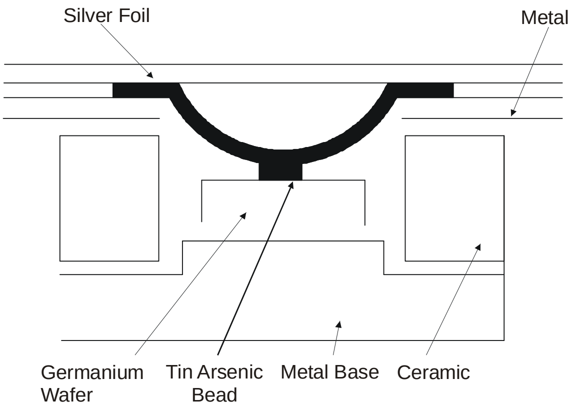
43. By operating the diode in the region where the voltage/current curve is negative, it can be used as a negative resistance oscillator or amplifier. When employed in an oscillator circuit, the positive resistance of the circuit is completely cancelled by the negative resistance of the diode. In an amplifier circuit, the positive resistance is partly cancelled, and the input applied to the circuit is amplified.

44. The equivalent circuit of a tunnel diode (Fig 15b), is similar to that of a normal semiconductor diode, but the variable barrier resistance becomes negative (–r). The constant resistance (R) is due to the semiconductor material and the capacitance (C) depends on the junction area and width. Thus, C is fairly high (about 10 pF) and, at microwave frequencies, the tunnel diode presents a very low impedance.

45. The transit time of the carriers through the very narrow depletion layer is very short (about 10^{-13} seconds), so the tunnel diode can be used at the highest radar frequencies.

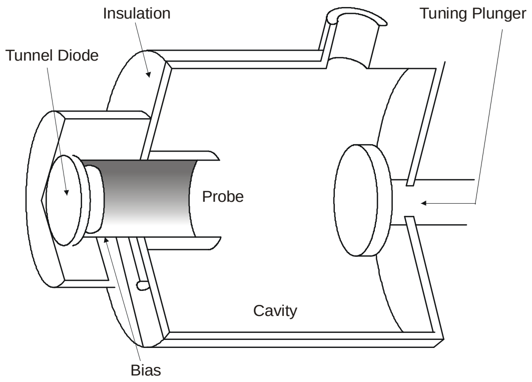
46. A typical construction of a tunnel diode is shown in cross-section in Fig 16. The p-n junction is formed by alloying a small n-type tin-arsenic bead into the surface of a p-type germanium wafer. The wafer is soldered into the casing, which forms one contact, and the n-type bead is connected to the other end of the casing via a strip of silver foil. The silver foil is soldered to the bead to act as a support as well as a contact. The bead is very small (about 0.025 mm diameter) to ensure a low junction capacitance and a high cut-off frequency.

14-20 Fig 16 Conventional Tunnel Diode Construction



47. **Tunnel Diode Oscillator**. Since a suitably biased tunnel diode presents a negative resistance to its terminals at all frequencies up to cut-off, it is capable of resonating at the resonant frequency of any high Q series-tuned circuit of smaller resistance connected across is terminals. It is possible to construct a resonant cavity tunnel diode oscillator capable of being tuned over a broad frequency range. The outline of a simple system is shown in Fig 17. The oscillation frequency is that of the cavity, and may be varied by altering the setting of the tuning plunger. Tunnel diode oscillators have been developed to operate satisfactorily at 100 GHz, and the cut-off frequency of available tunnel diodes is usually quoted as being in excess of 50 GHz. At this frequency, output powers of 200 μW are possible.

14-20 Fig 17 Resonant-cavity Tunnel Diode Tuneable Oscillator



48. **Tunnel Diode Amplifier**. The tunnel diode is a 'majority carrier' device and is not limited in frequency by the low diffusion velocity of minority carriers. Its high frequency performance is limited only by its CR product and by the inductance of the package. By using small area junctions, and semiconductor materials with high charge-carrier mobility, the CR product may be kept low to produce cut-off frequencies in excess of 50 GHz. The useful upper frequency limit for tunnel diode amplifiers is of the order of 20 GHz. Up to this frequency, the tunnel diode amplifier can provide gains as high as 15 dB at noise figures as low as 3.5 dB. Other advantages of the tunnel diode amplifier are its reliability, and its small size and weight. By using strip-line components, very compact RF heads for radar receivers operating at frequencies around 10 GHz have been developed.

P-I-N Diode

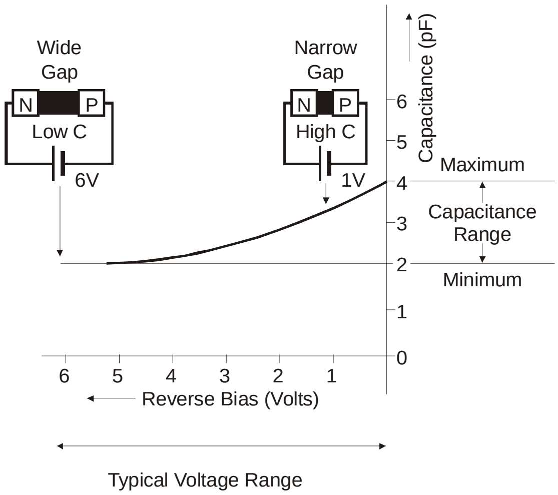
49. The p-i-n diode consists of heavily doped layers of p- and n- type silicon, separated by a thin slice of high resistance undoped (intrinsic) silicon. p-i-n diodes have been designed for use as switches at frequencies up to 40 GHz, with insertion losses as low as 0.1 dB and isolation better than 30 dB. This insertion loss means that about 98% of the incident power passes to the output when 'on', and the isolation is such that only about 0.1% of the incident power reaches the output when 'off'. For some applications, high peak power handling capability and high reverse breakdown voltage characteristics are required, and p-i-n diodes capable of handling up to 10 kW peak power with breakdown voltages of 1500 volts are now available.

50. It is possible to have an array of p-i-n diodes suitably spaced along the length of a transmission line. This gives an improvement in the attenuation and control of RF energy being passed, and also improves the reliability of the system, since the failure of any one device would result in only a small reduction in overall effectiveness.

Varactor and Step Recovery Diodes

51. In a semiconductor diode, a capacitance is formed by the opposite charges either side of the junction. When the diode is reverse biased, the charges move further apart and the junction capacitance decreases (capacitance is inversely proportional to the distance between opposite charges). Therefore, by varying the value of reverse bias voltage applied to the diode, the capacitance can be varied, as shown in Fig 18.

14-20 Fig 18 Variation of Capacitance in Varactor Diode



52. p-n junction diodes that have been specially doped to provide a wide variation of capacitance with applied reverse bias are called 'varactor diodes'. They have many uses, eg providing a voltage controlled variable capacitance for tuning tuned circuits, in parametric amplifiers, and as electronic switches. One of their main uses is in frequency multiplier circuits, where the output frequency is a multiple of the input frequency.

53. The 'step-recovery' diode is a close relative of the varactor diode, differing in the degree of doping of the n and p regions, and being particularly suitable for frequency multiplication.

Metal Base Transistor

54. In the metal base transistor, the usual n- or p-type base region is replaced by a thin metal film of molybdenum, which has three effects:

a. The 'energy gap' between the emitter and the base is increased, so that the electrons from an n-type emitter are injected into the base region with high energy.

b. The base resistance is reduced.

c. There are no minority carriers.

55. Electrons are injected over the emitter-base barrier and collected after crossing the reverse-biased base-collector junction. Control is effected by the base voltage; the current is not diffusion limited as there are no minority carriers in the base region. The emitter current is more like the thermionic emission current in a valve; the mobility of electrons through the base is high, transit time effects are negligible, and the low base resistance of the metal base transistor further improves high frequency performance.

56. Metal base transistor amplifiers have been developed to give 20 dB of gain at 10 GHz, and there are indications that they will be used to provide amplifiers and oscillators in the 50 GHz to 120 GHz range.

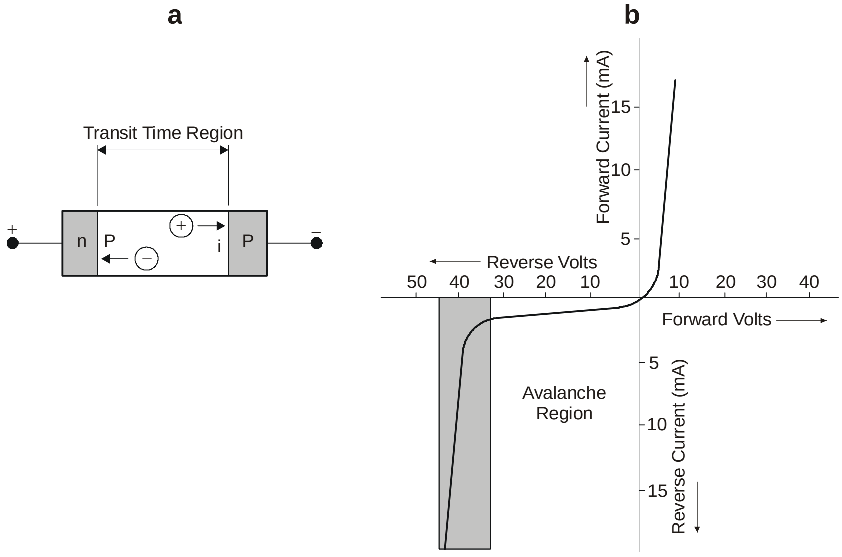
IMPATT Devices

57. The term IMPATT (**IMP**act **A**valanche and **T**ransit **T**ime device) is used to describe a broad class of oscillating devices which use impact ionization to create an avalanche of charge carriers which, drifting in a transit time region, produce a negative-resistance oscillation.

58. IMPATT devices include those semiconductor diodes which exhibit negative resistance characteristics, and which also depend for their operation on transit time effects. Such diodes include the silicon avalanche diode and the 'Read' diode. In general terms, an IMPATT device contains at least one semiconductor junction which is reverse biased to such a value that the resulting electric field is sufficient to spontaneously generate electron-hole pairs by a form of internal secondary emission (avalanche or multiplication process).

59. Fig 19 shows the schematic outline of a reverse biased Read diode, and illustrates a typical characteristic. The reverse bias creates a strong electric field across the transit time region. The maximum field exists at the p-n junction, and is large enough to create avalanche conditions, producing a pulse of electron-hole pairs. The generated electrons are attracted by the applied field to the nearby n-region, whilst the positive charge carriers (holes) move through the transit time region to the p-electrode. When the holes reach the p-electrode, another avalanche pulse is generated, the frequency of the pulses depending upon the transit time of the charge carriers. The phase relationship between current and voltage is such that the device exhibits negative resistance characteristics. The diode can, therefore, be used as a negative-resistance oscillator.

14-20 Fig 19 Basic Read Diode and Characteristic



60. The usable frequency range is partly determined by the transit time of the current carriers. Thus, for operation at 10 GHz, the transit time region would be very thin (about 25 × 10^{-6} mm). However, since the basic IMPATT device produces pulses, sinusoidal oscillations can be achieved only by mounting the diode in a microwave cavity. The same diode can be used with several different cavities to produce oscillations over a very wide frequency range. An experimental Read diode has produced oscillations in the 2 to 4 GHz band in a coaxial system, in the 7 to 12 GHz band in an I-band waveguide, and at 50 GHz in a millimetric waveguide.

61. A silicon avalanche diode has been developed to provide outputs of 1W CW at 12 GHz with 8% efficiency, and a few milliwatts CW at 50 GHz with 2% efficiency. It seems likely that these figures can be improved upon to provide perhaps 20W in the 5 to 20 GHz range, with efficiencies of some 30%. As an amplifier, IMPATT devices have produced gains of 20 dB at 10 GHz with 30 MHZ bandwidth, although the high noise of 50 dB prohibits its use as an RF amplifier.

TRAPATT (Trapped Plasma Avalanche Transit Time) Diode

62. Whilst trying to improve the efficiency of the IMPATT by experimenting with the diode/cavity geometry, it was accidentally discovered that an anomalous mode of operation is possible. If the diode is installed in a cavity whose resonant frequency is approximately half of the IMPATT operating frequency, and a voltage approximately twice the breakdown voltage is applied, unusually high power and high efficiency can be achieved. The overvoltage produces an avalanche shock front, which propagates through the diode at a velocity that is greater than the velocity of the charge carriers. Behind this front, a high-density plasma is trapped in the low field region that has been created. These trapped carriers then move relatively slowly across the diode and maintain a high current in the external circuit, whilst the applied voltage is low. Diodes operating in the TRAPATT mode can achieve up to 20 W pulsed at an efficiency of up to 30%. Although they are currently limited in operating frequency to about 16 GHz, this is quite high enough for airborne applications, such as radar transponders and TACAN equipment.

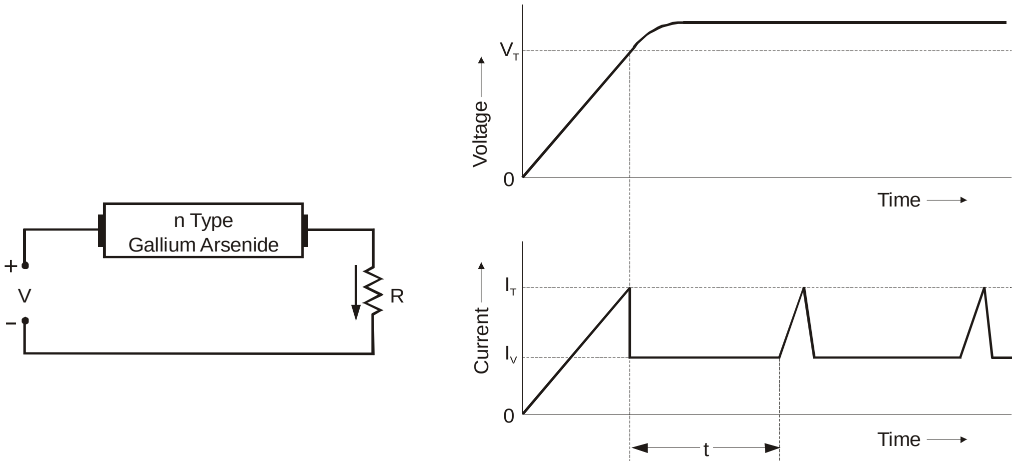
Gunn Effect Oscillator

63. When the voltage applied to a thin slice of gallium arsenide semiconductor exceeds a critical threshold value, periodic fluctuations of current result and, if the slice is thin enough (4 to 25 × 10^{-6} mm), the frequency of oscillation is in the microwave band. The effect is known as the 'Gunn' effect, after its discoverer. A Gunn effect semiconductor is called a 'bulk' device since there are no junctions.

64. Fig 20a shows a basic circuit of a gallium arsenide slice connected, in series with a small resistor, to a source of dc voltage. If the current through the resistor is measured, it is found that it is proportional to voltage up to a certain critical threshold value of voltage (V_T). However, when the voltage is increased above V_T, the current drops rapidly from its value (I_T) to a valley current value (I_V). The current then remains at the value I_V for a time (t) before rising to its original value (I_T) (Fig 20b). The period of oscillation is related to the time (t), which depends upon the slice thickness, but is independent of the applied voltage.

14-20 Fig 20 Gunn Effect in Gallium Arsenide Semiconductor

a Circuit **b Characteristics**



65. Gunn oscillators can be tuned to operate over a band of frequencies by mounting the diode in a microwave resonant cavity of variable length. Operated in a pulse mode, the output power ranges from 200 W at 1 GHz to 1 W at 10 GHz. CW outputs between tens and hundreds of mW are available. These devices are simple and reliable, and find applications in receivers, as local oscillators, and in low-power transmitters.

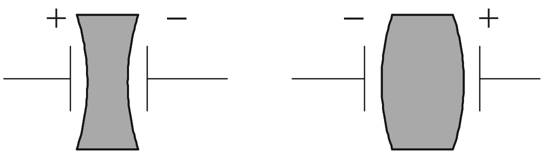
SURFACE ACOUSTIC WAVE (SAW) DEVICES

Introduction

66. Surface acoustic wave devices are becoming increasingly important as components in radar systems where, typically, they may be used as delay lines, pulse generators, and to produce frequency modulated pulse compression pulses. The underlying phenomenon is that of the piezo-electric effect.

67. **Piezo-Electric Effect**. The piezo-electric effect is exhibited by a number of crystals, such as quartz. It manifests itself as a change in physical shape in response to the application of a voltage across a slab of material made from such crystals. In effect, the material expands or contracts depending on the polarity of the voltage, as shown in Fig 21. Therefore, if an oscillating voltage is applied, the material will vibrate at the frequency of the applied voltage. The effect is reversible; thus, if a piezo-electric crystal is alternately compressed and expanded, an oscillating voltage will appear across its face.

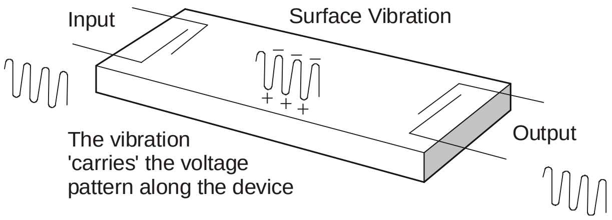
14-20 Fig 21 Piezo-Electric Effect



Basic SAW Device Construction

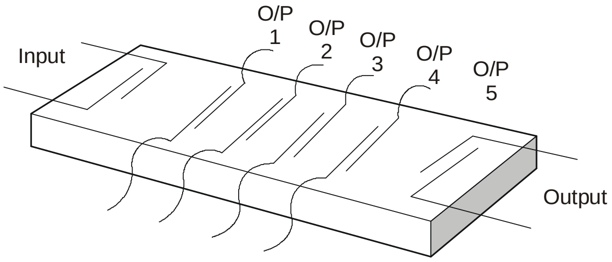
68. A basic SAW device consists of a slab of piezo-electric crystal material with metal inter-digital fingers deposited on one surface, as shown in Fig 22. When an input oscillating signal is applied, a voltage is created across the input finger. This produces a mechanical vibration in the surface material. This physical vibration travels along the crystal slab, mimicking the electrical input signal and, as the vibrations pass the output fingers, an oscillating voltage is generated.

14-20 Fig 22 Basic SAW Device



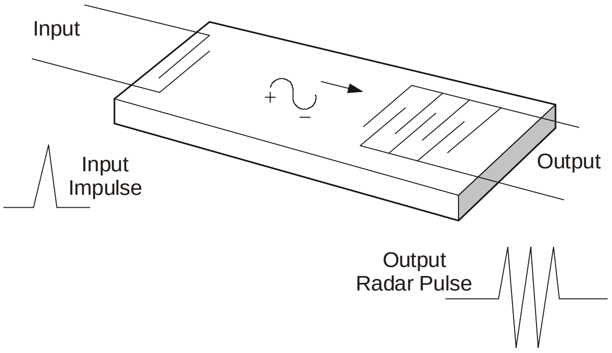
69. **Delay Line**. The speed at which the mechanical vibration travels in the slab is some 10^5 times slower than the speed of electromagnetic waves in space; so, the simple device of Fig 22 could act as a delay line. For a 1µs delay, the device need only be 3 mm long, compared to 100 ft of cable to achieve the same effect. The delay can be varied by placing a series of inter-digital fingers at different positions along the length of the device, so that the output can be tapped after any desired delay (Fig 23).

14-20 Fig 23 Tapped Delay Line



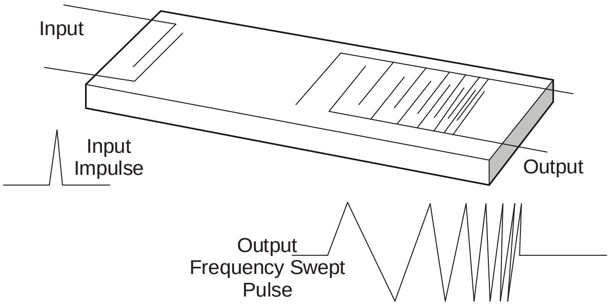
70. **Pulse Generation**. If a single voltage spike is applied at the input of the simple device of Fig 22, the output will also be a single spike. However, if the output contacts are modified to be a series of inter-digital fingers (Fig 24), an oscillating pulse is produced. As the mechanical vibration pattern passes the first two fingers, a positive spike is generated and, as it passes the second pair, a negative spike is produced, and so on. The number of cycles in the pulse depends on the number of inter-digital fingers at the output, and the wavelength on the spacing of the fingers.

14-20 Fig 24 SAW Pulse Coding

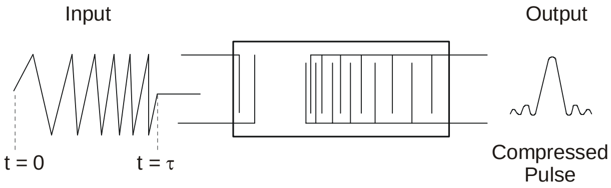


71. **Production of an FM Pulse Compression Pulse**. The device illustrated in Fig 24 can be modified to produce an output pulse with a varying frequency by varying the spacing of the output inter-digital fingers, as shown in Fig 25. By reversing the device, it can act as a detector for such a frequency modulated pulse. In Fig 26, the FM pulse is applied to the input. The resulting surface wave will have an identical pattern, and will only produce a large output when the pattern exactly overlaps the output contacts, ie the device compresses the pulse into a narrow width.

14-20 Fig 25 FM SAW Device



14-20 Fig 26 A SAW Device for Pulse Compression



LOW NOISE AMPLIFIERS

Introduction

72. If there was no noise present in a receiver, it would be possible to detect any signal, given sufficient amplification. The noise generated within a practical receiver, compared with a 'noiseless' receiver, is expressed by the noise figure (F) where:

$$F = \frac{S_{in}/N_{in}}{S_{out}/N_{out}}$$

where S_{in} = input signal power

S_{out} = output signal power

N_{in} = input noise power

N_{out} = output noise power

73. Amplifiers in a superhet receiver generate sufficient noise to negate the gain they provide. However, low noise amplifiers have been developed, such as:

a. The travelling-wave tube (paras 18 - 24).

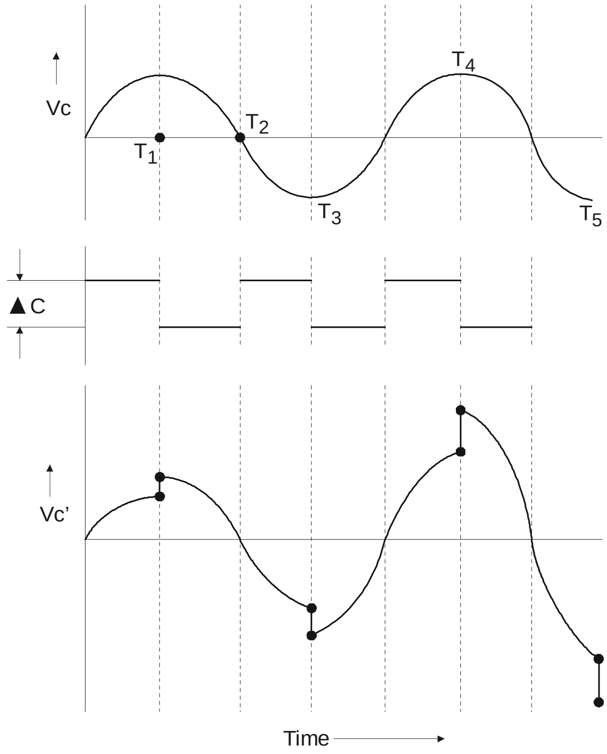
b. The parametric amplifier.

c. The maser.

The Parametric Amplifier

74. The parametric (or reactance) amplifier derives its name from the fact that the equation governing its operation has one or more parameters that vary with time. The principle of operation may be explained by considering a simple resonant circuit having an inductance and a capacitance oscillating at the resonant frequency. If the capacitor plates were to be pulled apart at the instant when the oscillating voltage was at a positive maximum, then, since for a fixed charge the voltage is inversely proportional to the capacitance, the voltage would increase. If the plates were returned to their original position as the voltage passed through zero, then, since there would be no charge on the plates, the voltage would remain unchanged. If this process was repeated, separating the plates every time the voltage passed through a positive or negative maximum, and returning them to the original position as the voltage passed through zero, then a signal at the resonant frequency would be amplified. The principle is illustrated in Fig 27 where, at time T_1, corresponding to a voltage maximum, the capacitance decreases, leading to a step increase in voltage. There is no change at T_2, where the voltage is zero, but there is a further step at T_3, where the voltage is at a negative maximum. The process is repeated at T_4, T_5 and so on. The variable capacitor element of a parametric amplifier is normally provided by a varactor diode (see para 51), in which the depletion layer at the junction may be considered as the dielectric between the plates of a capacitor. Under reverse bias, the width of the depletion layer varies with the applied electric field and hence, if an oscillating voltage is applied across a varactor diode, the capacitance will vary at the oscillating frequency.

14-20 Fig 27 Principle of Parametric Amplifier

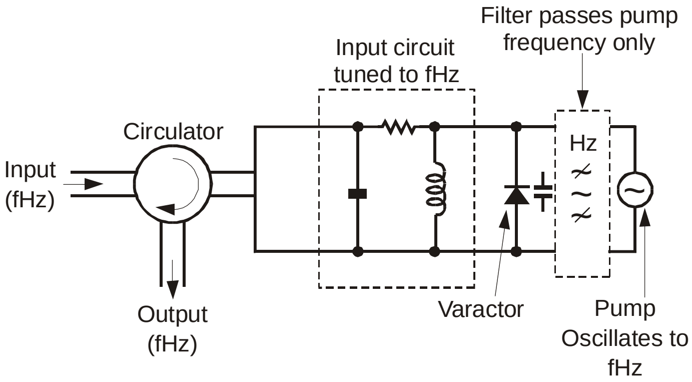


Vc = Input Signal Vc' = Amplified Output

▲C = Change in Capacitance

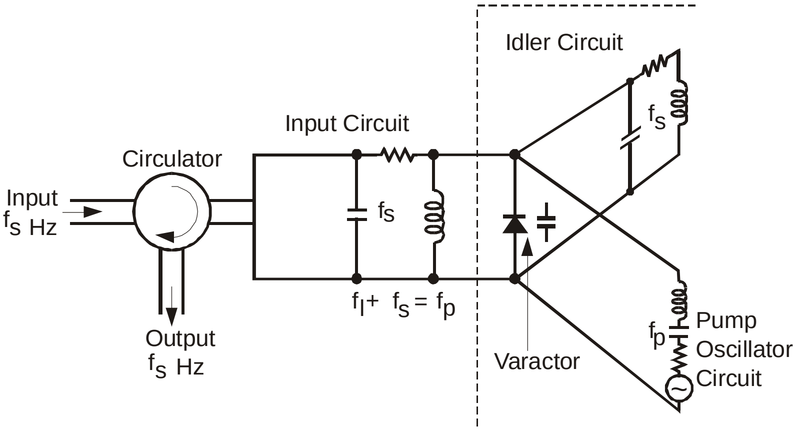
75. **Degenerate Parametric Amplifier**. The simple parametric amplifier described above requires that the variation in the capacitance occurs at a frequency, called the 'pump' frequency, which is twice that of the resonant frequency of the resonant circuit. This mode of operation is known as the 'degenerate mode', and amplification takes place only if the phase relationship between the pump frequency and the signal frequency is correct. The essentials of a basic degenerate parametric amplifier are shown in Fig 28.

14-20 Fig 28 Degenerate Parametric Amplifier



76. **Non-degenerate Parametric Amplifier**. The phase sensitivity of the degenerate parametric amplifier can be overcome by operating the pump at some frequency other than twice the resonant frequency. The nature of the parametric amplifier is such that a third frequency, known as the 'idling frequency', is produced. The idling frequency is equal to the difference between the pump frequency and the signal frequency. The phase and frequency limitations can be relaxed if an idler circuit, resonant at the idling frequency, is added, as shown in Fig 29. In effect, this idler circuit acts as an energy reservoir, accepting energy from the pump or signal circuits, and storing it until needed, then releasing it at the proper time and phase, to provide power gain in the signal circuit. The non-degenerate parametric amplifier, in which $f_p = f_s + f_i$ (frequencies of pump, signal, and idler circuits respectively), is a negative resistance device; it has limited gain and, like any other negative resistance amplifier, has a tendency towards instability.

14-20 Fig 29 Non-degenerate Parametric Amplifier



77. **Parametric Up-converter**. The problem of instability can be solved by a parametric amplifier which presents a positive resistance to the signal circuit, by making $f_p = f_i – f_s$. This is called an 'up-converter', and is completely stable. The output is at the idling frequency, which is higher than the signal frequency, and this amplifier can be followed by a conventional crystal mixer receiver. The up-converter has a maximum gain which is proportional to f_i/f_s, so is primarily useful at UHF frequencies or lower.

78. **Other Parametric Devices**. If $f_p = f_s – f_i$, the device becomes a 'positive resistance down-converter'. It has a gain of less than unity, and is used as a mixer in preference to the crystal diode. Negative resistance up- and down-converters are also available; these have high gains but have tendencies towards instability.

79. **Properties of the Parametric Amplifier**. The parametric amplifier has a very low noise figure, typically about 2 to 3 dB at room temperature; this can be improved further by refrigeration using, for example, liquid nitrogen at 77 K. This low noise is mainly the result of not employing a noisy electron beam generated by a hot cathode. A parametric amplifier can be used as a signal frequency amplifier in radar receivers to improve the receiver sensitivity, and hence increase the range of the radar. It is mounted between the TR cell and the crystal mixer and is a 'fail safe' device, in that failure of the parametric amplifier does not result in failure of the radar, although there will be a decrease in signal-to-noise ratio. The noise present in a parametric amplifier is due to the noise introduced by the varactor and the thermal noise generated in leads, circulators, waveguide walls etc; the noise figure increases with increasing operating frequency. The gain of a parametric amplifier depends on the pump power supplied, and the pump frequency, and is limited by possible instability. The output from a klystron pump oscillator is often fed to the varactor through an attenuator in order to give the required gain consistent with stability. Pump power is normally a few mW. For a given input frequency, a high gain can be obtained by using a high pump frequency; however, it is difficult to generate useful power at very short wavelengths. A typical gain is 20 dB, and, to obtain this with an input frequency of 600 MHz, 20 mW of pump power is used at a frequency of 8.9 GHz.

Masers

80. The maser is an extremely low noise amplifier, based on quantum mechanical principles, and capable of noise temperatures in the order of a few degrees Kelvin. The name is an acronym formed from **M**icrowave **A**mplification by **S**timulated **E**mission of **R**adiation. A maser requires complicated external circuitry and a magnetic field and, furthermore, it has a narrow bandwidth and must be operated at liquid helium temperatures (about 1 to 4 K). The principle of operation of the maser is identical to that of the laser, in that the electrons in the atoms of a material are pumped to a higher than normal energy level from which they are stimulated to jump back to a lower level with the emission of energy. The frequency of the emission is dependent on the difference in energy levels and is given by:

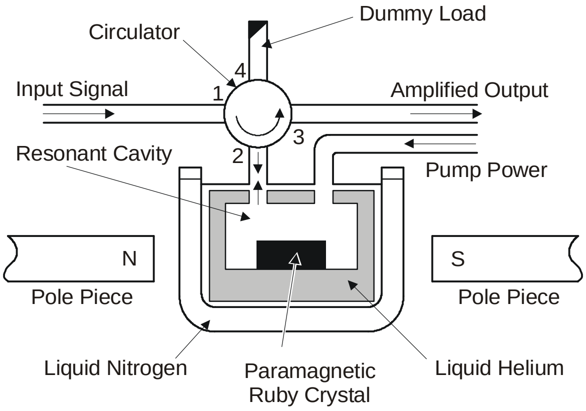
$$f = \frac{\text{Difference in energy levels}}{\text{h (Planck's constant)}}$$

There are several forms of maser and those of chief interest in radar applications are solid-state devices using paramagnetic crystals placed in magnetic fields.

81. **Three-level Cavity Maser**. The three-level cavity maser, in which a ruby maser material has three closely spaced energy states (E_1, E_2, and E_3), is illustrated in Fig 30. E_2 and E_3 are separated by an energy difference which would lead to radiation at the frequency of the signal requiring amplification. A strong RF pump signal is applied to the system at a frequency of $f_p = (E_3 – E_2)/h$. By exciting electrons into higher energy levels, this forces atoms from the lower energy state E_1 to the higher state E_3 until the populations in the two levels are equal. There are now more atoms in E_3 than in E_2, and the application of the signal which requires amplification ($f_s = (E_3 – E_2)/h$) acts as a stimulus to cause atoms to drop to the E_2 level, with an accompanying emission of radiation at frequency f_s, thereby amplifying the signal. The separation of the energy states depends upon the strong magnetic field in which the paramagnetic material is placed. Thermal processes can, however, mask this maser effect, hence the need to cool the material to liquid helium temperature. Low noise is an important outcome of this low temperature operation. Since the cavity maser is a one-port device, the amplified signal must be taken from the cavity via the same line as it entered, and a circulator is used to separate the input signal from the amplified output signal. A typical three-level cavity maser, operating at 1.5 K, amplifies a signal of 2.8 GHz with a gain of 20 dB. The pump power is about 10 mW at 9.4 GHz, the bandwidth 20 MHz, and noise figure

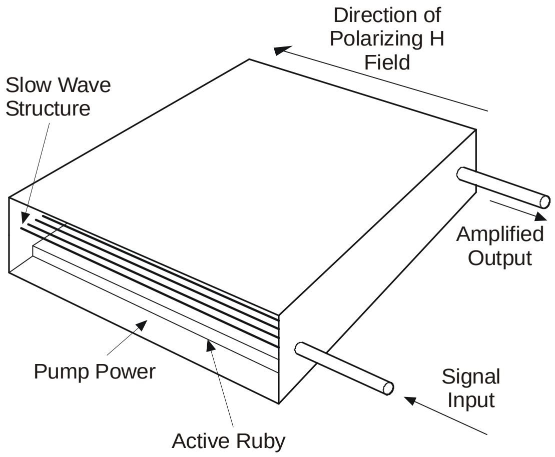
about 0.3 dB. An I-band maser amplifying a signal of 9.4 GHz, requires a pump frequency of 24 GHz. The disadvantages of this type of maser are its tendency to oscillate and its narrow bandwidth, both of which are overcome in the travelling wave maser.

14-20 Fig 30 Three Level Cavity Maser



82. **Travelling Wave Maser**. The use of a travelling wave structure, rather than a cavity, allows bandwidths to be improved to a reasonable figure. The basic arrangement, shown in Fig 31, consists of a ruby loaded comb structure mounted in a waveguide. Pump power is fed down the waveguide and this causes a population inversion in the ruby crystal. The input signal to the slow-wave structure causes stimulated emission, and is thus amplified. A magnetic field, in the direction shown, is provided and the whole structure is immersed in a liquid helium bath. A typical travelling wave maser, operating at 19 GHz, has a gain of 20 dB, a bandwidth of 55 MHz, and a noise figure of 0.16 dB.

14-20 Fig 31 Travelling-wave Maser



WAVEGUIDES

Introduction

83. Above about 1 GHz, dielectric and skin losses in coaxial cables become so great that only very short lengths may be used. The greatest contribution to the loss comes from the inner conductor, and the solution to the problem is to remove it. Thus, a waveguide is essentially a hollow air-filled pipe made of a conducting material; it may be considered as a medium of one dielectric constant (air) enclosed in another medium of different dielectric constant (metal tube). In theory, the cross-section of the waveguide may take any form but, in practice, it is usually rectangular, although circular

waveguides are used for some specialized applications. The dimensions of the cross-section are dependent on the frequency of the wave which is to be transmitted along the waveguide and, as a general rule, the dimensions decrease as the operating frequency increases.

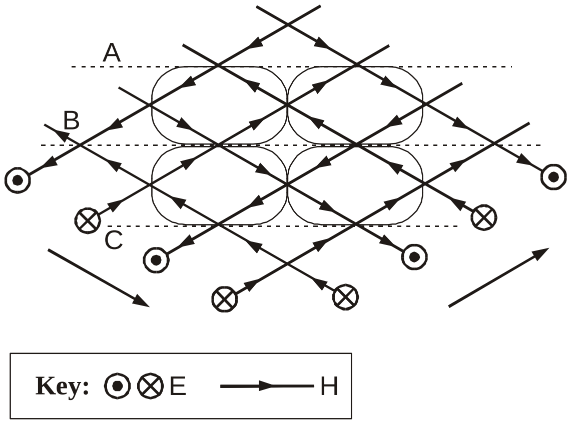
Propagation in Waveguides

84. Where an electromagnetic wave passes close to a conducting surface, certain boundary conditions must be fulfilled. These are that the electric field (E) must be normal to the conducting surface, or zero, and the magnetic field (H) must be parallel to the conducting surface, or zero. If an attempt is made to pass a transverse electromagnetic (TEM) wave straight down the axis of a rectangular waveguide, the boundary conditions will not be satisfied at one pair of the conducting surfaces. Thus, a TEM wave is not capable of being propagated straight down the waveguide.

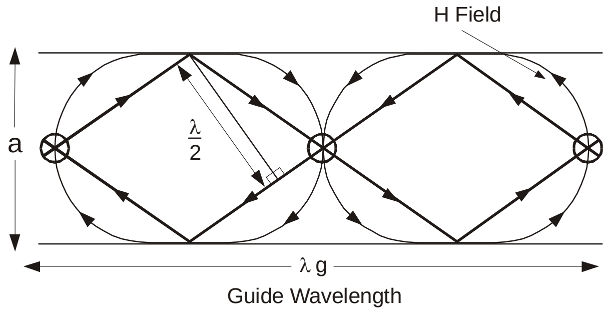
85. If a TEM is passed into a waveguide, such that it satisfies the boundary conditions at the broad faces, but is travelling at an angle to the axis of the waveguide, it will be reflected by the narrow faces (at which the boundary conditions will also be met).

86. Fig 32 shows how two such waves, at an angle to each other, would interact. Along the planes A, B, and C, perpendicular to the page, the resultant E field is zero, and the H field between these planes forms loops as shown. Thus, conducting plates could be placed along these planes and the boundary conditions would be satisfied. A further pair of plates can be added to complete a rectangular waveguide, and these too will satisfy the boundary conditions. Fig 33 shows the resulting pattern which, since the fields change polarity at the signal frequency, propagates along the waveguide. The wavelength of the pattern, known as the 'guide wavelength' (λ_g), is longer than the free space wavelength. Fig 34 shows the arrangement in three dimensions.

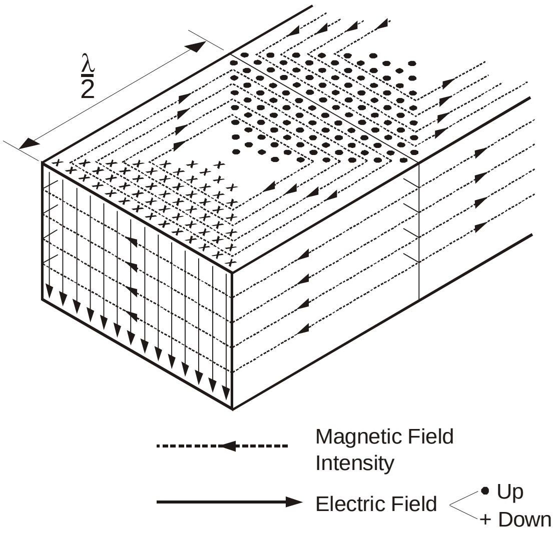
14-20 Fig 32 Interaction of Two TEM Waves



14-20 Fig 33 Field Pattern Resulting from Two TEM Waves



14-20 Fig 34 Three-dimensional Illustration of H_{10} Mode



87. The pattern described is known as the H_{10} mode; other modes are possible but the H_{10} is the most efficient. The particular mode which is established depends upon the dimensions of the waveguide but, if the broad dimension of the waveguide (a) lies between λ (the free space wavelength) and λ/2, only the H_{10} mode will be propagated. If a = λ/2, a wave will not propagate, and there is, therefore, a critical wavelength (λ_c = 2a) corresponding to a cut-off frequency, $f_c = \frac{3\times10^8}{\lambda_c}$, below which the waveguide will not work.

88. The next highest mode of propagation is established if λ = a or 2b (where b is the narrow waveguide dimension), whichever is the greater. In order to prevent arcing, and to allow maximum power handling, b should be as large as possible and is normally made equal to a/2. In summary, for only H_{10} propagation:

$$\text{a (or 2b)} < \lambda < \text{2a, or } \frac{3\times10^8}{a} > f > \frac{3\times10^8}{2a}$$

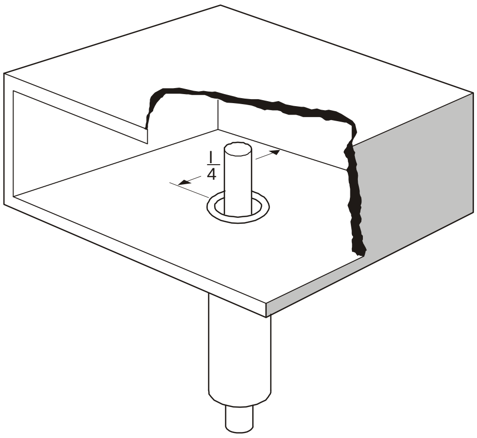
89. **Power Carrying Capacity**. There is a limit to the power which can be carried by any transmission line. In an air-filled waveguide, there is no loss in the air and the power dissipated in the walls is seldom a limiting factor. The limit is set by voltage breakdown in the air. Dry air, at atmospheric pressure, breaks down when the electric field strength reaches about 30 KV/cm. For the practical power rating of waveguides, it is usual to take the safe field strength limit as 15 KV/cm. Since the dimensions of a waveguide are a function of wavelength, their power-carrying capacity decreases as frequency increases. Changes in pressure also affect breakdown, and it is found that the breakdown voltage is approximately proportional to pressure over a wide range. This presents an obvious limitation for airborne equipment. For example, at 35,000 ft, where the air pressure is approximately a quarter of that at ground level, the power carrying capacity of the waveguide is reduced by a factor of 16 (since power is proportional to the square of voltage). This effect can be obviated by sealing off the waveguide system so that the interior is maintained at, or near, standard atmospheric pressure. Conversely, the power-carrying capacity can be increased above the normal by increasing the pressure inside the guide above standard pressure.

Waveguide Techniques and Components

90. **Launching the Wave**. Two methods are commonly employed to launch the waves into the guide. The first method employs a suitable probe, connected to a short run of coaxial line, which is fed by the source of the waves. In its simplest form, the probe is merely the inner conductor of the

coaxial line, extended across the narrow dimension of the guide, the outer conductor of the line being connected to the waveguide itself (Fig 35). The inner conductor acts as a small aerial and radiates the required component TEM waves of the waveguide mode. The waveguide short circuit is used to reflect the waves in the required direction, and is placed $\lambda_g/4$ behind the probe, so that the correct phase relationship is obtained. The other method of launching the wave is by connection to a resonant cavity by means of a slot.

14-20 Fig 35 Launching from a Coaxial Line by a Simple Probe



91. **Extracting the Wave**. When a waveguide is used to connect an aerial to a receiver, the wave is often extracted by means of a probe, as used for launching the wave. When the waveguide is connecting a transmitter to an aerial, the method of extraction is usually one of the following:

a. **A Matching Horn**. If the waveguide terminated in a simple open end, some energy would leak out, but some would also be reflected. To prevent the reflection, the open end of the waveguide is flared out to form a horn, effectively matching the impedance of the guide to that of free space; the impedance of the guide gradually changing over the flared section.

b. **Coupling Slots**. The H field in a waveguide causes currents to flow in the inside walls, these currents being at right angles to the H field. If a slot is cut in the wall of the waveguide, such that it interrupts the flow of wall current, it will radiate energy. Fig 36a shows slots cut in the broad and narrow sides of a waveguide. Slots 1 and 2 will interrupt the wall current at right angles, so maximum coupling occurs; slots 3 and 4 are parallel to the wall current and no energy is coupled; slot 5 is inclined to the wall current and slot 6 is displaced from the centre of the broad side. As both slots 5 and 6 interrupt the current, they will couple energy from the guide, but with reduced effectiveness compared to slots 1 and 2. For least disruption of the waveguide field pattern, the slots should be half a wavelength long. In practice, a series of half-wave slots is cut in the wall of the waveguide, arranged so that they each radiate an equal amount of in-phase energy. Two systems used are:

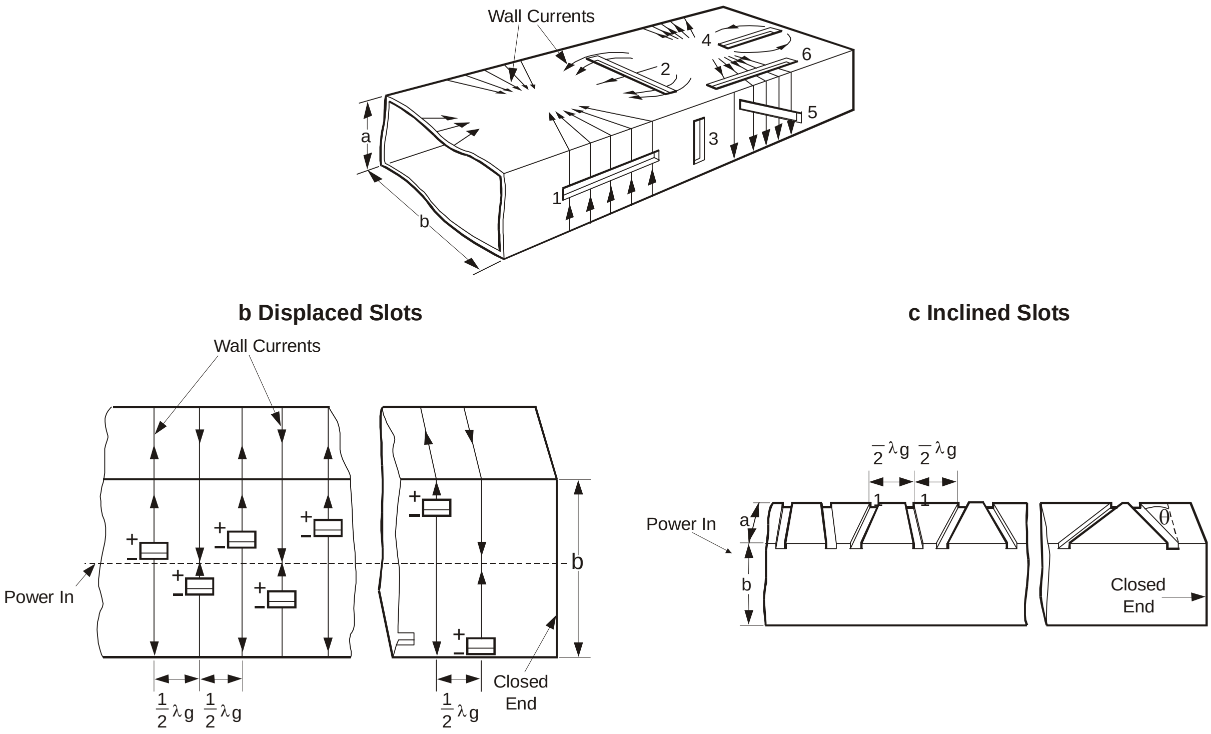
(1) **Displaced Slots**. Displaced slots are slots of length $\lambda_g/2$ cut parallel to, but not along, the centre line of the broad side (e.g. slot 6 in Fig 36a). As the wall current reverses phase every half a guide wavelength, the slots are spaced with their mid-points $\lambda_g/2$ apart on alternate sides of the centre line, successive slots being further removed from the centre line to give equal radiation from each slot (Fig 36b).

(2) **Inclined Slots**. Inclined slots are slots of length $\lambda_g/2$ cut at an angle in the narrow side of a waveguide, e.g. slot 5 in Fig 36a. As the angle of inclination (θ in Fig 36c) is increased, the

coupling increases. With an array of slots, the angle of inclination is progressively increased to compensate for the reduction of energy in the guide, so that each slot radiates equal energy. If the slots were spaced λ$_g$/2 apart, and inclined in the same direction, adjacent slots would radiate in anti-phase. By alternating the slopes of adjacent slots, as in Fig 36c, the slots radiate in phase, and a beam at 90º to the line of the array is formed. As the slots are half a wavelength long, they continue into the broad side of the waveguide.

14-20 Fig 36 Coupling Slots

a Slot Positions in Relation to Wall Current

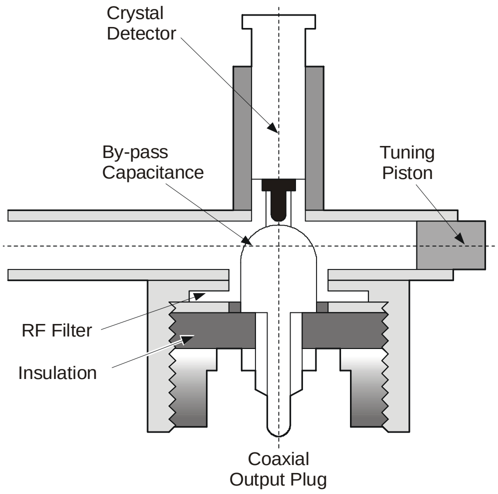


b Displaced Slots

c Inclined Slots

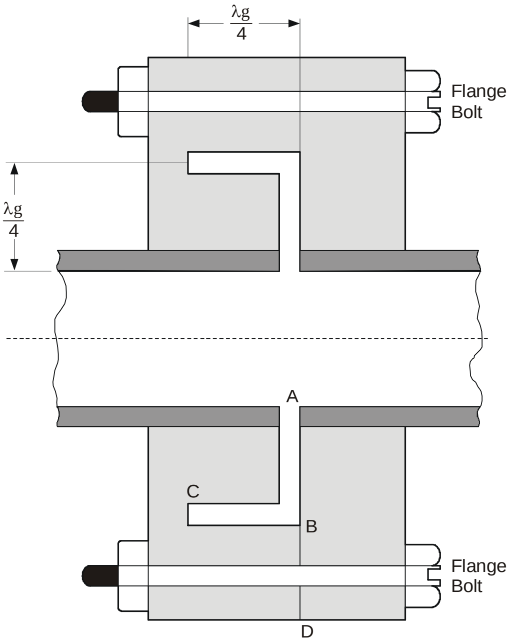
92. **Detecting the Wave**. To detect signals travelling down a waveguide (e.g. for test purposes), it is usual to mount the detector directly on, or in, the waveguide, the physical size of the normal crystal detectors being such that this is readily possible. Fig 37 shows a coaxial-type crystal, mounted on a 'doorknob' transformer, acting as a detector in a rectangular waveguide.

14-20 Fig 37 Crystal Detector Mounted on a Waveguide

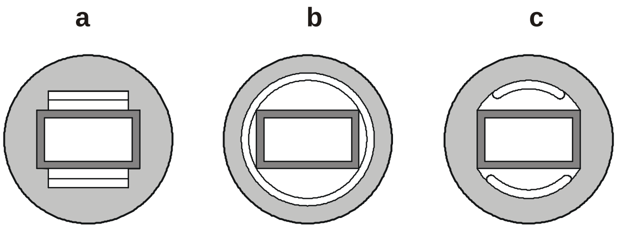


93. **Choke Joints**. Problems with installation and maintenance demand that waveguide circuits be made in a number of rigid sections which must then be joined together. Early systems used two plain flanges on the ends of the pieces to be joined, which were then clamped mechanically in order to obtain good electrical continuity. A better system is that used in the resonant, short-circuit, choke joint, shown in section in Fig 38 and end-on in Fig 39a. One flange is still plain, but the other is machined to form an effective half wavelength of transmission line when bolted to the plain flange. Due to standing waves, the short circuit at C sets up an effective short circuit at A, on the waveguide wall, thus giving good continuity. Since the point B is at a region of high impedance (and hence low current), poor contact between the coupled sections will not have any appreciable effect. It should be noted that good continuity is not required on the narrow faces of the guide since the electric field there is zero. For this reason, the choke BC is often machined in circular fashion, as shown in Figs 39b and c.

14-20 Fig 38 Section of Choke Coupling

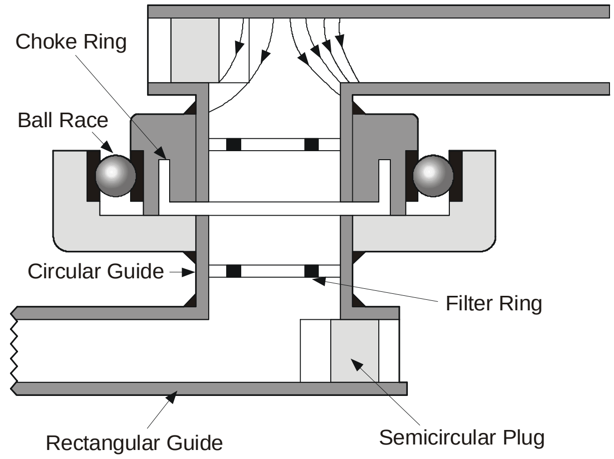


14-20 Fig 39 Various Types of Choke Flanges



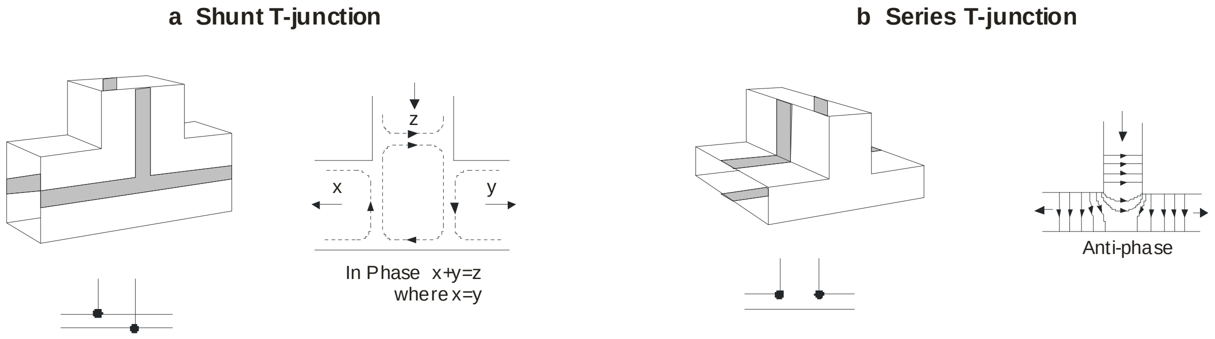
94. **Rotating Joints**. In many radar and communications equipments, it is necessary to rotate the aerial with respect to the rest of the equipment and, for this purpose, a rotating joint must be provided in the waveguide feed. A section through a typical rotating joint is shown in Fig 40. A circular waveguide is used for the rotating section. The wave travels along the circular guide in circular mode, which has a radially symmetrical field pattern. The input to, and output from, the joint is by rectangular guide. Transformation from the rectangular mode to the dominant circular mode is assisted by the semicircular plugs in the stub ends of the rectangular guide. Unfortunately, this dominant mode does not possess circular symmetry. To suppress it, ring filters are fitted above and below the choke joint, producing a field which possesses circular symmetry. This allows rotation of one part of the joint relative to the other, without disruption of field patterns.

14-20 Fig 40 Waveguide Rotating Joint



95. **T-junction**. There is frequently a necessity to split the energy carried in a waveguide into two portions, or combine the energy carried in two waveguides into a single guide. This is usually carried out by T-junctions, the two basic forms of which, together with their transmission line equivalents and field patterns, are shown in Fig 41. If the junction is set into the narrow face of the waveguide, it is an H-plane or shunt junction, whereas, if it is made in the broad face of the guide, it is an E-plane or series junction. The two types of junction are similar in that, if power is fed into the vertical arm of the junction, it is split evenly between the other two arms of the junction. However, they differ in that the resultant waves are in phase for the shunt junction, but in anti-phase for the series junction.

14-20 Fig 41 T-junctions

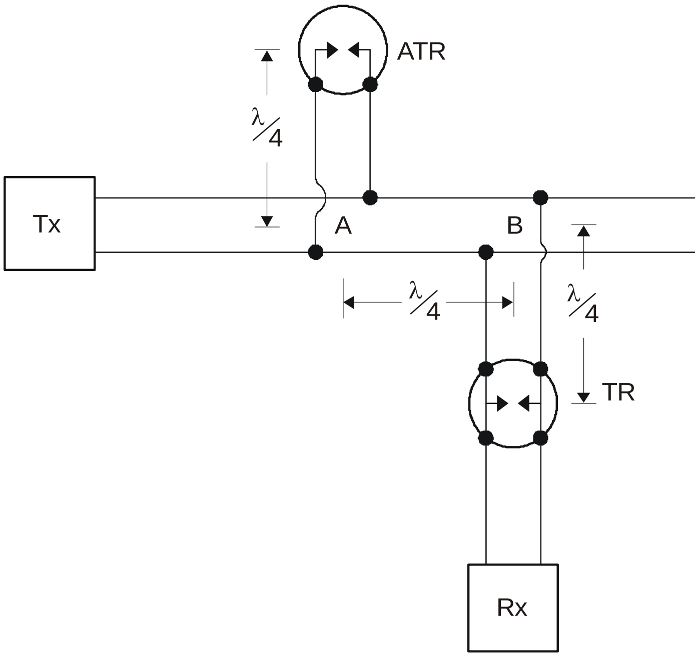

a Shunt T-junction

b Series T-junction

96. **TR Switches**. In a pulse radar system which uses the same aerial for transmission and reception, it is necessary to include in the waveguide system a transmit-receive (TR) switch (also known as a 'Duplexer'). This automatically couples the transmitter to the aerial for the duration of each pulse, while at the same time protecting the receiver from burn-out or damage. It also couples the aerial to the receiver during the receiving period. The most common types are:

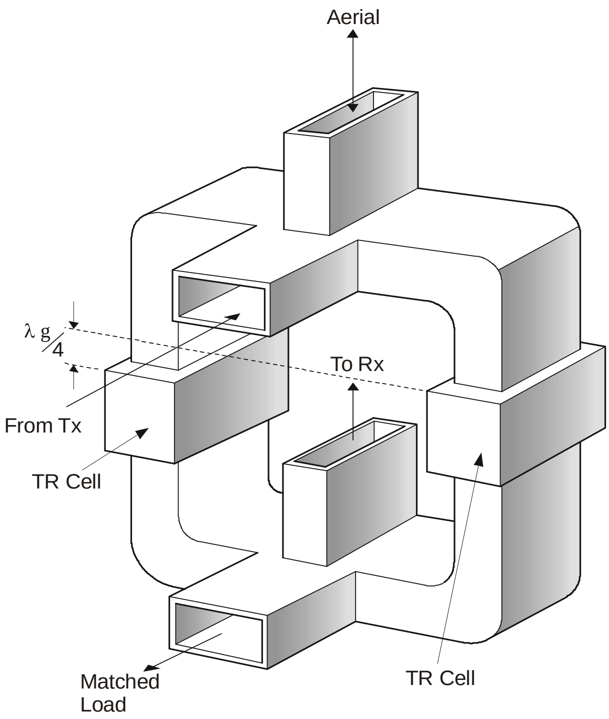
a. **Branched Duplexers**. The principle of the branched duplexer is illustrated in Fig 42. It consists of a TR cell and an anti-transmit-receive (ATR) cell placed on the waveguide (or transmission line), as shown. The ATR cell is properly termed a 'transmitter blocker' (TB) cell. Both of the cells are special gas-discharge valves, such as the soft rhumbatron, which ionize and provide a near short-circuit when the transmitter is turned on. Since both cells are situated $\lambda_g/4$ from the main feeder in shunt arms, the short-circuits at the cells are transformed to very high impedances across the main feeder at points A and B. Thus, none of the transmitter power flows to the receiver. When the transmitter pulse is ended, the switches rapidly de-ionize and offer a high impedance. This is transformed by the $\lambda_g/4$ sections to represent low impedances at points A and B. The low impedance at A is further transformed by the $\lambda_g/4$ section A-B to a high impedance at B. Received signals therefore pass down the receiver branch and do not pass B towards the transmitter.

14-20 Fig 42 Branched Duplexer Principles



b. **Balanced Duplexers**. There are many circuit arrangements of TR and ATR cells for switching. The branched duplexer is probably the simplest configuration, but it is not inherently broadband. A more broadband arrangement is the balanced duplexer whose bandwidth is limited only by that of the waveguide. A combination of series and shunt T-junctions (see Fig 41) is known as a 'hybrid T-junction' (sometimes called a magic T-junction). Hybrid T-junctions are used in conjunction with the TR cells, and a common arrangement is shown in Fig 43. The transmitter power enters the magic-T by the arm shown, and splits equally left and right in phase. The two signals cause the TR cells, which should be identical and acquired in matched pairs, to ionize, thus acting as short circuits. The two signals are totally reflected, but since their path lengths differ by $\lambda_g/2$ by the time they arrive back at the top magic-T, they leave by the arm which is connected to the aerial. If any small fraction of the transmitter power manages to penetrate the cells, two signals arrive at the bottom magic-T in phase and leave by the shunt arm, which is connected to a matched load where they are dissipated. The receiver is thus completely isolated from the transmitter power. On receive, the signal arriving from the aerial splits equally left and right in anti-phase, and none of it enters the transmitter. The two signals are not sufficiently strong to ionize the TR cells and so pass straight through. They arrive at the bottom magic-T, still in anti-phase, and so combine to leave by the arm connected to the receiver. In addition to this arrangement, a low power TR cell is often placed between the balanced duplexer and the receiver to safeguard the receiver against random pulses from nearby radar equipments. These pulses may be too weak to ionize the cells, but could be strong enough to damage the receiver.

14-20 Fig 43 Balanced Duplexer



97. **Miscellaneous Waveguide Components**. The foregoing list of waveguide components is by no means exhaustive and the scope of this chapter will permit only a brief mention of the following components:

a. **Bends and Twists**. Often, a change of direction is necessary in a waveguide run. This can be done simply by the use of a length of guide, bent as required. The overriding principle in such a device is that any change must be made gradually to avoid reflections.

b. **Attenuators**. In making measurements on waveguide components, and for other purposes, it is often desirable to absorb some or all of the incident energy. This can be achieved by the use of attenuators, of which there are several types.

c. **Matching Devices**. Most waveguide components introduce a certain amount of mismatch, which has to be removed by some matching device. Matching can be carried out by means of waveguide stubs of adjustable length, in the same way as matching stubs are used with two-conductor transmission lines, or by means of lumped reactances, known as 'irises', placed in the waveguide.

d. **Components for Circular Polarization**. Circularly polarized waves have many radar applications (eg in discriminating against rain). One of the most popular devices for producing such waves is the dielectric quarter-wave plate, which consists of a longitudinal dielectric slab in a section of circular waveguide, placed at 45º to the plane wave. This introduces the correct amount of phase delay in one component of the wave to cause the emerging wave to be circularly polarized.

MICROWAVE AERIALS

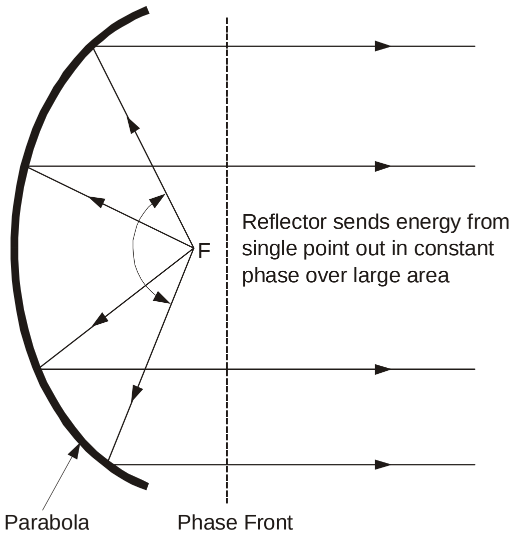
Introduction

98. One of the principle advantages of operating in the microwave region of the spectrum is that it is possible to obtain a narrow beam width from a small aerial, allowing accurate bearings to be measured, and permitting rapid movement of the aerial, where necessary. The narrow beam may be formed by a reflector or a lens (in a similar manner to light), or by an array of radiating elements.

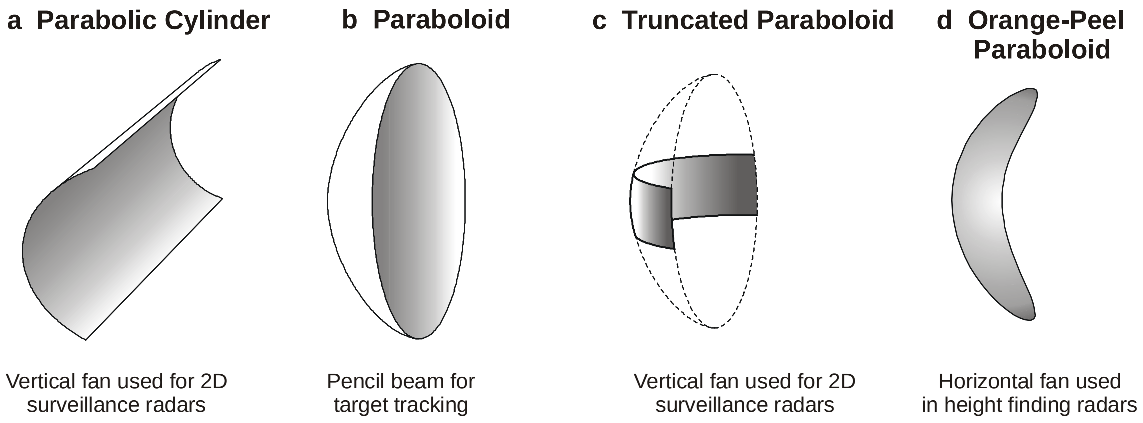
Parabolic Reflectors and Feeds

99. The parabolic reflector is perhaps the most familiar beam-forming device. If an aerial is placed at the focal point, energy radiated towards the reflector is reflected parallel to its axis (Fig 44). The larger the reflector, and the smaller the wavelength, the narrower the beam will be. Scanning or pointing of the beam is accomplished by mechanical movement of the aerial. Different horizontal and vertical beamwidths may be obtained by using different horizontal and vertical dimensions, and the reflector may be parabolic in both dimensions or only in one. Some typical reflector shapes, with their applications, are illustrated in Fig 45. The cosecant^2 shape, in which the top half of the dish has a circular cross-section, and the bottom half a parabolic cross section (Fig 46), is often used in airborne ground mapping radars. It is designed to ensure that all the ground to be mapped is illuminated with uniform energy and, to achieve this, more energy must be directed to long ranges than to short ranges. The required cosec^2 pattern may be also be obtained by placing a small prism of dielectric, or metal plates, in the aperture of a parabolic dish, or by using multiple stacked feeds.

14-20 Fig 44 Parabolic Reflector


F
Reflector sends energy from single point out in constant phase over large area
Parabola
Phase Front

14-20 Fig 45 Some Typical Reflector Shapes and their Applications



a Parabolic Cylinder

Vertical fan used for 2D surveillance radars

b Paraboloid

Pencil beam for target tracking

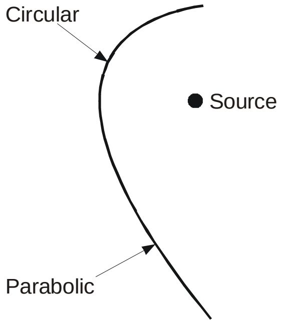
c Truncated Paraboloid

Vertical fan used for 2D surveillance radars

d Orange-Peel Paraboloid

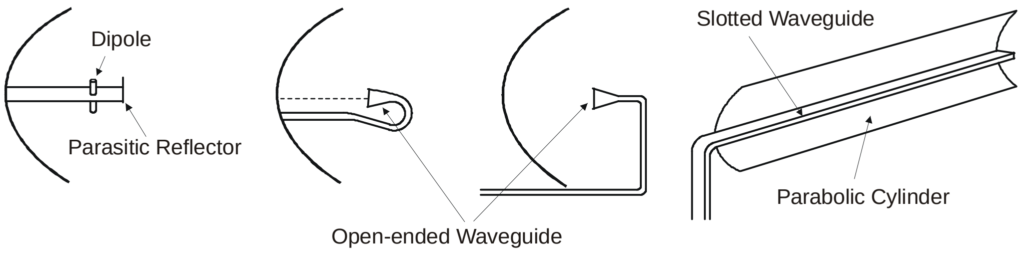
Horizontal fan used in height finding radars

14-20 Fig 46 Cosec^2 Dish



100. **Feeds for Parabolic Aerials**. The feed for a parabolic aerial can only approximate to the theoretical point source. In practice, it may consist of a dipole placed ahead of a parasitic reflector element but, more commonly, an open-ended waveguide is used; some typical arrangements are shown in Fig 47. In the case of the parabolic cylinder, the axial feed usually consists of a length of slotted waveguide. The design of the feed must produce the desired power distribution across the aperture (typically a cosine pattern), and the waveguide plumbing must cause the minimum obscuration of the aperture.

14-20 Fig 47 Feeds for Parabolic Aerials



101. **Beamwidth and Sidelobes**. The beamwidth (θ) of a parabolic reflector may be approximated by:

$$\theta = \frac{70\lambda}{d}$$

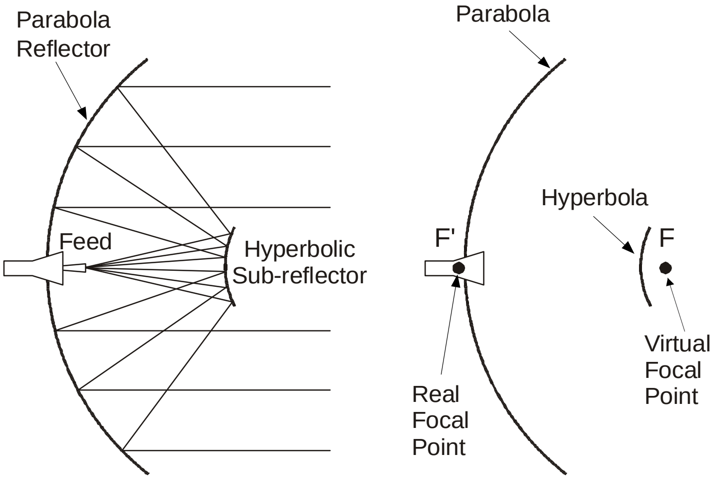
where λ is the wavelength in use and d is the aperture diameter. Sidelobes of the order of 20 dB to 30 dB below the main beam are readily achieved and, with extreme care, it is possible to reduce sidelobes to 40 dB down.

Cassegrain Aerials

102. The feed at the focus of a conventional parabolic reflector tends to cause aperture blocking, leading to larger sidelobes. The Cassegrain aerial system can reduce this problem to an extent (in the twist Cassegrain aerial), while needing shorter waveguide runs, and resulting in a reduction in the overall length of the aerial.

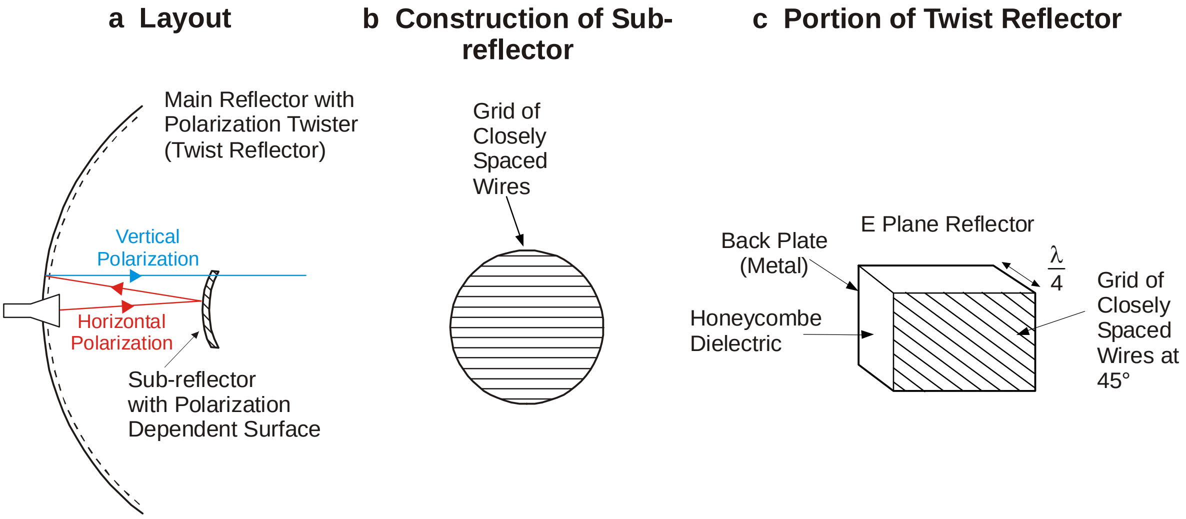
103. The Cassegrain aerial (Fig 48) consists of a combination of a parabolic reflector and a hyperbolic sub-reflector. The focus, F, of the parabola is also one of the conjugate foci of the hyperbola. The other focus, F′, of the hyperbola is coincident with the vertex of the parabola. The geometry of the arrangement is such that all rays parallel to the aerial axis are focused onto the vertex F′ of the parabola, at which point the feed is located. The position of the sub-reflector may be chosen from any one of a family of hyperbole focused on F and F′. The closer it is placed to the feed, the shorter the aerial is, but the greater is the aperture obscuration. It is possible to achieve comparable performance to a conventional parabolic aerial in a Cassegrain of half the overall length.

14-20 Fig 48 Principle of the Cassegrain Aerial



104. **Twist Cassegrain Aerial**. In an active radar system, where it is possible to operate with a single plane of polarization, the blocking of the aperture by the sub-reflector may be reduced by the technique shown in Fig 49 - the twist Cassegrain aerial. The sub-reflector comprises a grid of closely spaced horizontal wires (Fig 49b) which reflects the horizontally polarized energy from the feedback to the main reflector. The main reflector surface is constructed as shown in Fig 49c. The front face is a grid of closely spaced wires at 45º; the metal back plate is separated from the grid, at a distance of λ/4 (where λ is the operating wavelength), by a honeycomb dielectric. The horizontally polarized radiation incident on the main reflector may be considered to have two components, one parallel to the grid of wires at 45º, and one perpendicular to it. The parallel component is reflected at the grid surface, while the perpendicular component passes through the grid to be reflected by the back plate. The path difference between the two components is λ/2, corresponding to a phase difference of 180º. The result is that the emergent radiation has its polarization twisted through 90º, ie it is now vertical and, as such, can pass straight through the sub-reflector. The converse applies when the aerial is being used for reception. Some obscuration does occur because of the finite size of the feed, but this can be made small in comparison to that of a conventional aerial.

14-20 Fig 49 Twist Cassegrain Aerial


a Layout
b Construction of Sub-reflector
c Portion of Twist Reflector
Main Reflector with Polarization Twister (Twist Reflector)
Vertical Polarization
Horizontal Polarization
Sub-reflector with Polarization Dependent Surface
Grid of Closely Spaced Wires
Back Plate (Metal)
E Plane Reflector
λ/4
Grid of Closely Spaced Wires at 45°
Honeycombe Dielectric

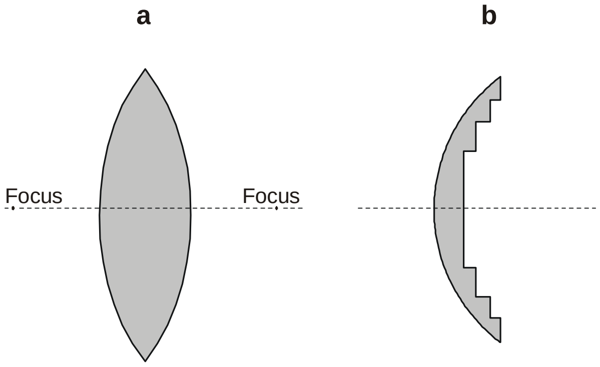
105. **Advantages of a Cassegrain Aerial**. Cassegrain aerials have advantages in monopulse applications, where the complex waveguides for the multiple feeds are normally difficult to engineer, and can cause considerable aperture blocking. The reduced waveguide runs, resulting from placing the feeds behind the reflector, have the further advantage that the required phase matching is more easily achieved. The Cassegrain aerial is specially attractive in infra-red systems because of the minimal waveguide losses between the feed and the amplifier.

Lens Aerials

106. As in the case of light, microwave energy may be focused by means of a lens, instead of by a parabolic reflector. The advantages of lens aerials include the elimination of aperture blocking by the feed (which is situated behind the lens), the lower dimensional tolerances required, and their suitability for wide angle scanning by moving the position of the feed. The development of lens aerials has been slower than for reflector types due initially to the lack of suitable materials possessing low dielectric loss, and the problem of dissipating heat from large lenses which can, in some applications, restrict the power handling capability. The three types of lens aerial applicable to radar are the dielectric lens, the metal plate lens, and lenses with a non-uniform index of refraction.

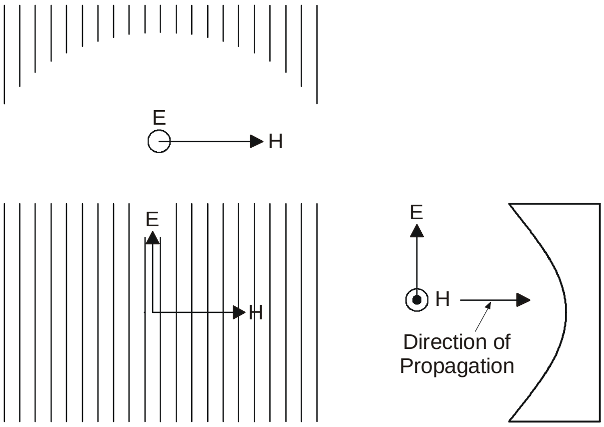
107. **The Dielectric Lens**. The dielectric lens is constructed of solid material such as polyethylene, polystyrene or perspex, and works in exactly the same manner as glass lenses at optical frequencies. The focusing action results from the reduced propagation velocity in the dielectric, the lens being shaped in such a way that rays emanating from a point source of energy at the focus become parallel to the axis after passing through the lens (Fig 50a). The thickness of the lens, which is a disadvantage in terms of bulk, weight, and dielectric loss, may be reduced by employing a stepped or 'zoned' lens as shown in Fig 50b. The depth of the steps, and the points at which they are introduced, are such that the path lengths through the lens in each successive zone differ by one wavelength. Thus, the uniformity of phase in the emergent wavefront is preserved. The penalties resulting from zoning are an increased sidelobe level, and the fact that the lens becomes frequency sensitive.

14-20 Fig 50 Dielectric Lens Aerials



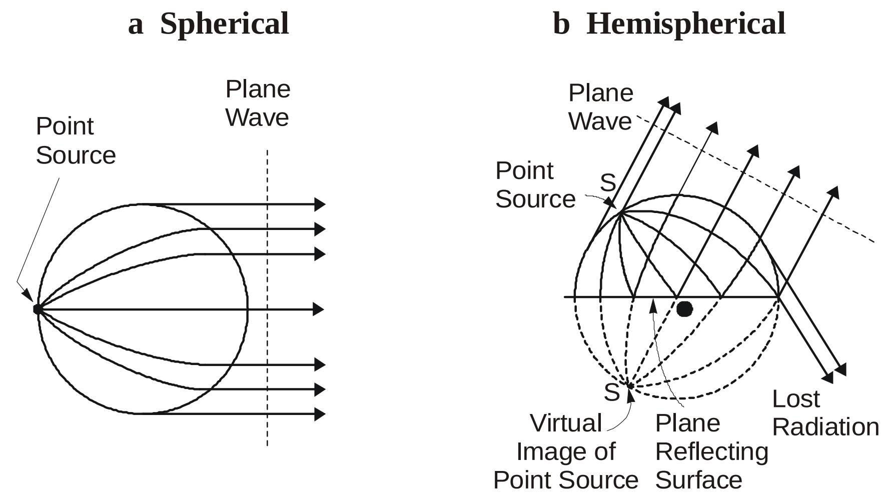
108. **Metal Plate Lenses**. Lens aerials may be formed from a grating of metal plates, spaced at an interval between λ and λ/2, and orientated in the plane of the E field (Fig 51). The plates act as a waveguide through which the waves propagate in the dominant mode with increased phase velocity. The effect is that of a dielectric having a refractive index less than unity, so the lens must be made thinner at the centre. As with dielectric lenses, the thickness of a metal plate lens may be reduced by zoning. Another class of metal plate lens, the 'constrained' or 'path-length' lens, focuses the energy by modifying the path-length through the lens, rather than the phase velocity. The plates may be orientated in either the E or H planes.

14-20 Fig 51 Metal Plate Lens



109. **Luneburg Lens**. Another method of focusing microwave energy is by means of a lens within which the refractive index is varied in some prescribed manner. The most important lens of this type is the spherical Luneburg lens (Fig 52a), in which the refractive index is made to vary in such a way that any plane wave incident on the sphere is focused to a point on the surface diametrically opposite. Conversely, a point source of energy on the surface is converted into a plane wave which can have any orientation. By moving the position of the feed mechanically, or by switching, the Luneburg lens can provide two-dimensional scanning over wide angles, if necessary, with more than one beam. If the required volume of scan is less than a hemisphere, only a part of the lens need be used (Fig 52b), although this would be at the expense of aperture blocking. A spherical Luneburg lens, having a reflecting cap over one hemisphere, is an efficient echo-enhancing device which is less sensitive to aspect than a corner reflector.

14-20 Fig 52 Spherical and Hemispherical Luneburg Lenses

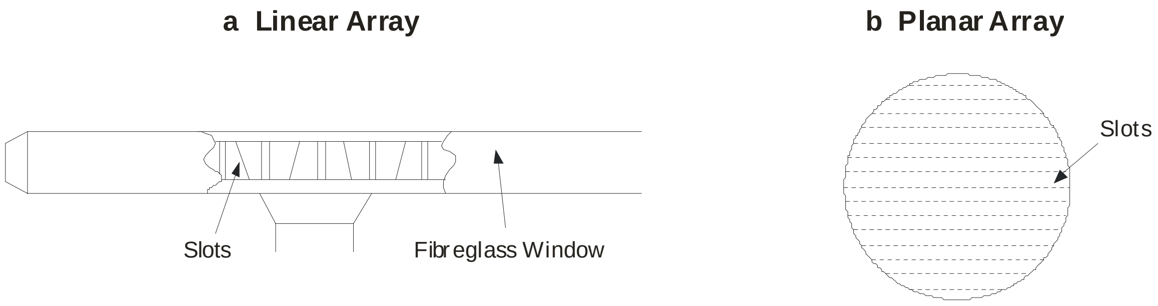


110. **Luneburg Lens – Refractive Index**. The Luneburg lens must be constructed so that its refractive index varies with the radius, from a maximum of $\sqrt{2}$ at the centre, to a minimum of 1 at the periphery. A practical approximation to this condition is made by forming the lens from a series of concentric spherical shells, each of which has a different refractive index. The beamwidth of a Luneburg lens is slightly narrower than that of a parabolic reflector of the same cross-section, but the sidelobe level is greater, typically about 20 to 22 dB below the main beam.

Flat Profile Aerials

111. Instead of a reflector or lens, a beam may be formed by an array of similar radiating slots, eg slots, dipoles, or horns. A linear array (Fig 53a) produces a single fan beam, while a planar array (Fig 53b) may be used to produce either single or multiple beams. Both types may be rotated mechanically to achieve scanning.

14-20 Fig 53 Flat Profile Aerials



Phased Array Aerials

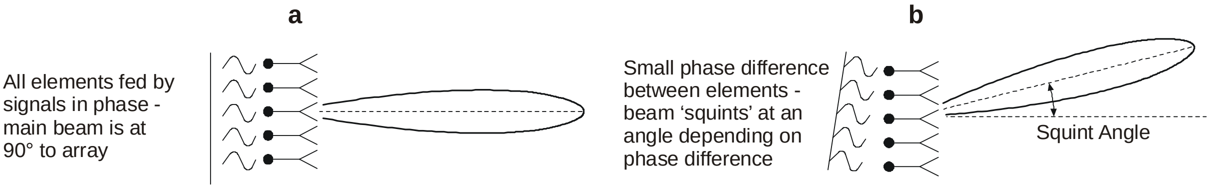
112. A development of the flat profile aerial is the phased array aerial, in which scanning is achieved electronically by varying the phase between the signals applied to the radiating elements, rather than by mechanical rotation.

113. In its simplest form, the phased array aerial consists of a number of radiating elements, equally spaced along a line and fed with in-phase signals of equal or tapered amplitude. This produces a beam at right angles to the array, which is described as a 'broadside linear array' (Fig 54a). If the radiating elements are separated by a distance, s, and a uniform phase difference, $\varnothing$, exists between adjacent elements, the angle of the beam to the broadside direction (θ), the squint angle (Fig 54b), is given by:

$$\theta = \sin^{-1}\frac{\varnothing\lambda}{2\pi s}$$

By interposing variable phase shifters in the feeds, the beam direction may be steered, theoretically up to 90º either side of the broadside direction, but, in practice, only up to 60º to 70º.

14-20 Fig 54 Phased Array Antennae


a
All elements fed by signals in phase - main beam is at 90° to array
b
Small phase difference between elements - beam 'squints' at an angle depending on phase difference
Squint Angle

114. The use of multiple radiating elements allows greater control over aperture power distribution than is possible with other aerial types, and lower sidelobe levels are theoretically possible as a consequence. A further advantage resulting from the use of many elements, is the ability to handle very high peak powers if necessary. The radiating elements may, in theory, be located over a surface of any convenient shape, such as the nose section of a fuselage, but, in practice, current applications use linear, planar, cylindrical, or spherical apertures.

115. **Linear Array**. A linear array produces a fan beam, capable of being scanned in a single dimension. When the beam is at an angle to the broadside direction, it forms the surface of a cone surrounding the array axis. This property has a useful application in Doppler navigation radars owing to the fact that, when the array is aligned with the aircraft ground velocity vector, equal Doppler shifts are produced from all points of intersection of the beam with the ground. In this application, the array is formed from a length of slotted waveguide, and the required phasing is obtained by choosing the correct relationship between the slot spacing and the guide wavelength. Parabolic cylinder aerials are commonly fed by slotted waveguide arrays and may, in some cases, be electronically steered. Two methods of achieving this are by varying the broad dimension of the waveguide mechanically, or by varying the transmission frequency. Both methods achieve their effect by altering the guide wavelength, and hence the relative phase of the radiation from adjacent slots.

116. **Planar Array**. The most popular array for radar applications is the planar array, which is formed from a number of linear arrays. The radiating elements, of which there may be several thousand, may consist of dipoles, open-ended waveguides, or log-periodic spiral elements. A rectangular aperture is capable of producing a fan beam which can be scanned in one dimension without serious deformation, while a square or circular aperture will produce a pencil beam capable of being scanned in two dimensions. A planar array may be steered in two dimensions by means of two control signals, one for azimuth and one for elevation. The phase shifters in any one horizontal or vertical row all receive the same control signal, but the phase shift in successive elements is varied linearly along the row. An array is most efficiently matched when the element spacing is exactly one half a wavelength but, because of its resonant characteristics, an array of this type has a limited bandwidth and it is, therefore, more usual to use a non-resonant spacing between the elements (ie $s < \lambda/2$). On a large array, a process known as 'aperture thinning' allows 5 to 10% of the

elements to be either left out, or to go unserviceable, without reducing the effectiveness significantly. This provides reduced cost, or a measure of redundancy.

117. **Beamwidth of Phased Array Aerials**. The beamwidth of a phased array aerial depends, as with other types of aerial, on the number of wavelengths in the aperture. If there are N elements, spaced at intervals of $\frac{\lambda}{2}$ across the aperture plane, and the power distribution is uniform, the beamwidth (θ_b) in the plane of the aperture is given by:

$$\sin \theta_b = \frac{\lambda}{N} \times \frac{360}{2\pi} \text{ degrees}$$

Thus, for small beamwidths,

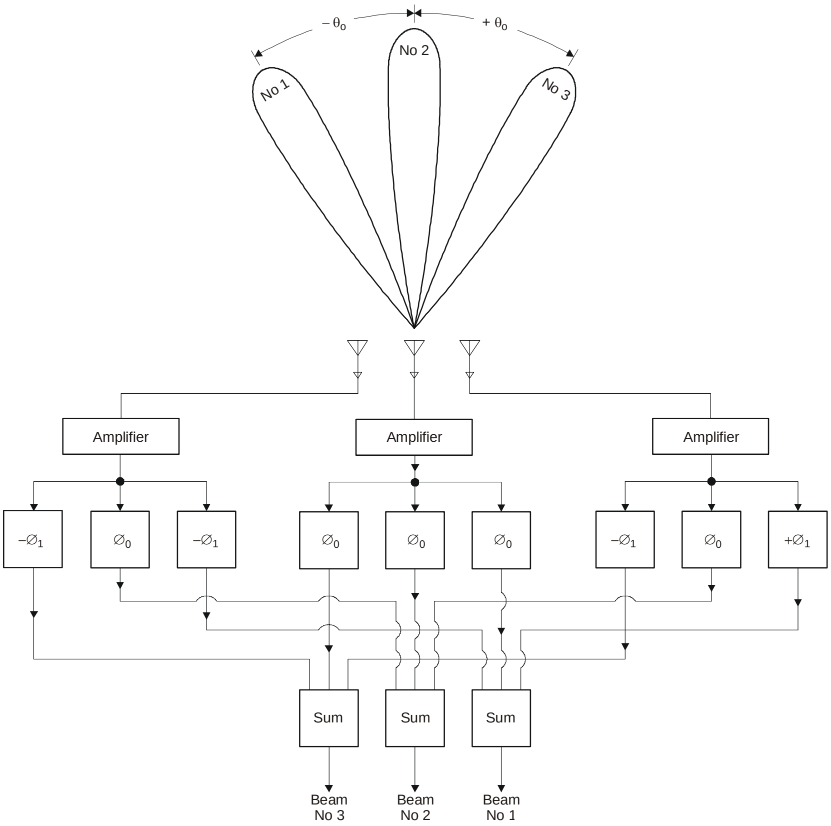
$$\theta_b \simeq \frac{114}{N}$$

A one-degree pencil beam requires a square aperture of comparable size to that of a reflector aerial, and must have about 10,000 elements. The sidelobe level for uniform power distribution is comparable to that of a uniformly illuminated parabolic reflector (about 13 dB below the main beam), and may be reduced by tapering the power towards the edges of the aperture. This may be achieved by reducing the energy radiated by the outer elements (amplitude tapering), or by omitting some of the elements altogether (space tapering). An array in which the latter technique is used is called a 'thinned array' (see para 116). The beamwidth of a broadside array is degraded as the beam is scanned away from the broadside direction and, over the useful angle of scan, increases approximately as the secant of the angle from the normal.

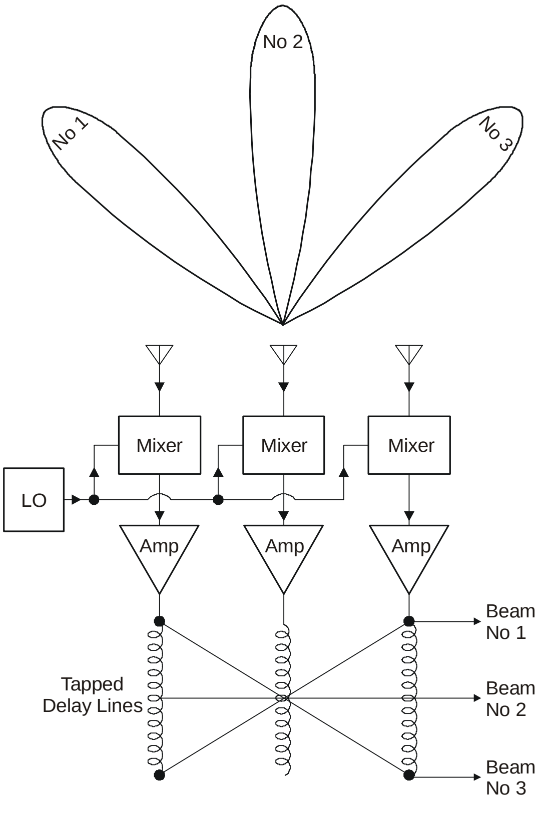
118. **Phase Shifting Devices**. Although, in the past, variable phase shifters have been mechanical devices, the present emphasis is on the development of electronic devices, making use of ferrite and semiconductor diode materials. Electronic phase shifters are classified as reciprocal or non-reciprocal, and may be either analogue or digital. Reciprocal phase shifters have identical characteristics for both transmission and reception of power. Digital phase shifters are generally preferred because of their suitability for direct control by digital computers. Switching speeds are in the region of 1 to 10 microseconds at operating frequencies up to I-band. One of the main problems involved in electronic phase shifters (particularly of the analogue type) is temperature sensitivity, and this necessitates close control of the array temperature.

119. **Multiple Beam Arrays**. The flexibility of array aerials makes it possible to generate a number of beams simultaneously from a single aperture; the array may, therefore, receive from many directions at the same time. The process is generally easier for reception than for transmission. However, the performance of a conventional scanning radar can still be achieved if the transmission is in a broad beam covering the entire field of view, and reception is made simultaneously in a number of stationary narrow beams which together cover the same field of view. For comparable performance, the echoes received in any one beam must be integrated over periods of time corresponding to the scanning period of a conventional radar. Multiple beams may be formed by combining the outputs of the elements in various ways through a system of phase shifters, delay lines, or hybrids. A simple beam-forming network for a three-element array, making use of phase shifters, is shown in Fig 55. Each element has one phase shifter for each beam, and the outputs of these are summed, in the manner shown, to form the three beams. An alternative method, shown in Fig 56, is to obtain the required phasing between the elements by means of tapped delay lines which, in a ground installation, may consist of lengths of waveguide. When the beam forming takes place after IF amplification, the aerial is called a 'post amplification beam-forming array' (PABFA).

14-20 Fig 55 Beam-forming Network Using Additional Phase-shifters

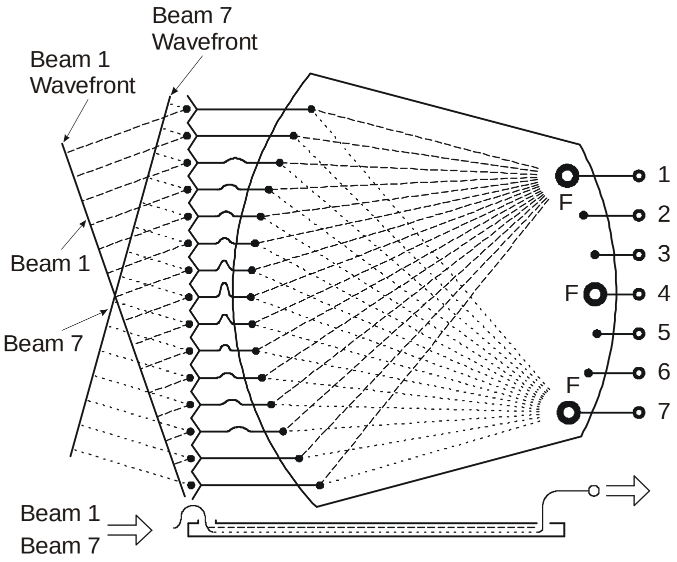


14-20 Fig 56 Beam-forming Network using Tapped Delay Lines



120. **Lens Fed Multiple Beam Arrays**. In addition to the techniques described above, a multiple beam array may be fed by way of a lens system. The Rotman lens is perhaps the most common, and the arrangement is outlined in Fig 57. An area of parallel plates has array port probes along the left side of the lens, and beam port probes on the right side. The array port probes are connected to the array of radiating elements by means of coaxial cables, which have lengths dependent on their position in the array. The lengths of the cables are such as to provide focusing at three points, as shown for beam ports 1, 4, and 7 in the diagram. The focusing is a result of providing equal path lengths from a given focal point out to the corresponding radiated wavefront for each element of the array. The beam formation technique produces fixed beams in space that do not scan with frequency. The number of beam ports corresponds to the number of overlapping lobes, and the number of array ports determines the beamwidth of the individual lobes. Typically, the beam coverage sector is 120º and the system will have a bandwidth of an octave. Although the Rotman lens gives perfect focusing at only three points, the departure from perfect focusing at intermediate points is negligible for all practical purposes. Rotman lenses may be stacked to form a cluster of conical fan beams.

14-20 Fig 57 Rotman Lens



Synthetic Aperture Systems

121. Although it is possible, theoretically, to use a sideways looking aerial array equal to almost the whole length of the aircraft, it is still not possible to achieve the very high azimuth resolution necessary for some applications. For example, a K-band signal ($\lambda = .009$ m) transmitted from a 9 m aerial would have a beamwidth of 0.001 radians, which equates to a linear beamwidth of approximately 15 m at 10 nm. The synthetic aperture technique allows better resolution to be achieved from physically smaller aerials. The underlying principle relies upon using the aircraft movement to create an effectively elongated aerial by synthesizing returns from a target as the aircraft flies past. The echoes returned from each pulse are recorded, and processed when enough returns have been received to create the required aperture. Range resolution is of the same order as azimuth resolution.

122. The system comprises a coherent radar, signal storage and processing equipment, an aerial with a high order of stability, and a display. Due to the requirement for large signal processing equipment, the system is currently restricted to large aircraft.

CHAPTER 21 - RADIO PROPAGATION

Introduction

1. Radio energy is propagated by means of electromagnetic waves which radiate from a transmitting aerial at a velocity which, in free space or air, is independent of the frequency used. In free space, this velocity is approximately 3×10^8 metres per sec (186,000 miles per sec or 162,000 nm per sec). In air it is only slightly less than this, but in other media it may be very much reduced and is frequency dependent.

2. The relationship between the velocity of propagation, wavelength and frequency of a radio wave is given by: $\lambda = \frac{c}{f}$

where λ= wavelength (metres)

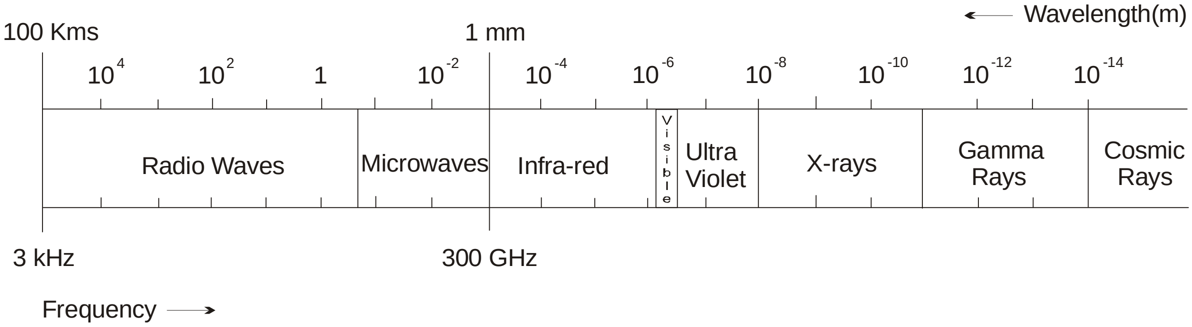
c = velocity of propagation (metres per sec)

f = frequency (cycles per sec or Hertz).

Radio Spectrum

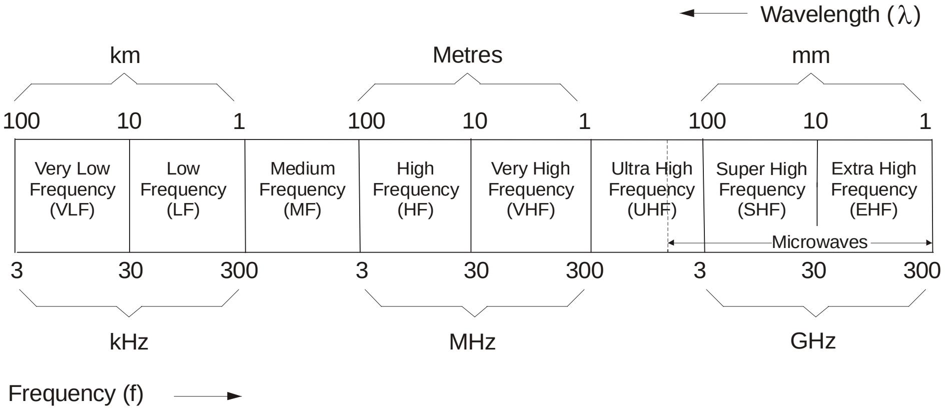
3. The position which radio waves occupy in the electromagnetic spectrum, shown in Fig 1, is not entirely arbitrary. The lower frequency limit is determined by the size and efficiency of the aerials required and the upper frequency limit by the attenuation and absorption of radio waves by the atmosphere.

14-21 Fig 1 The Electromagnetic Spectrum



4. As shown in Fig 2, it is convenient to divide the radio spectrum into frequency bands, each band nominally covering one decade of frequency. Radio waves higher than 1,000 MHz (l GHz) are usually termed microwaves.

14-21 Fig 2 The Radio Spectrum

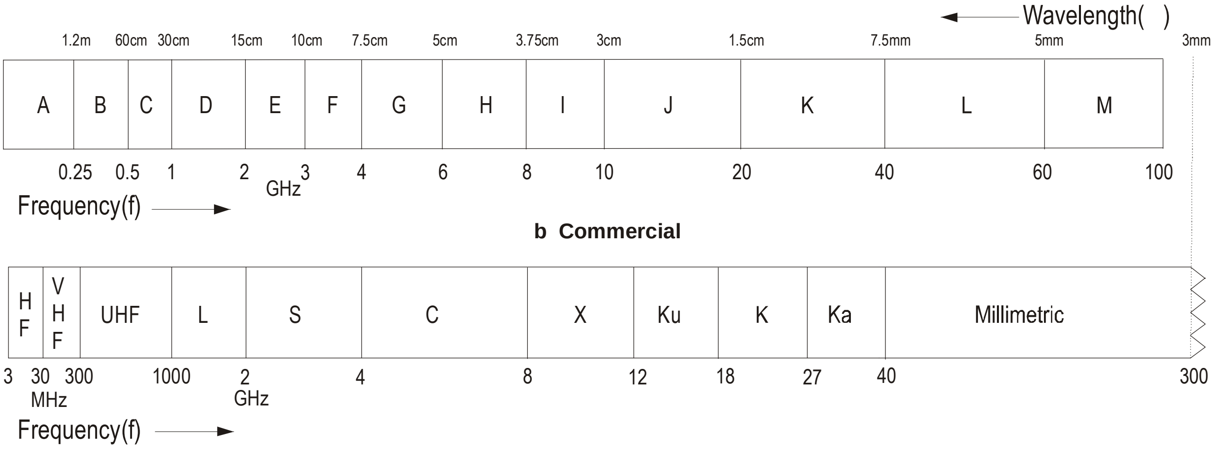


5. Radar frequencies, ranging from VHF to EHF, are usually subdivided into bands designated by letters. Fig 3 depicts the two agreed standards for NATO and commercial use. Each of the NATO bands are further subdivided into 10 channels eg D-1, D-2.........D-10.

14-21 Fig 3 Radar Bands

a NATO

b Commercial



Properties of Radio Waves

6. The radio waves leaving a transmitter exhibit the following fundamental properties:

a. They are transverse electromagnetic waves. That is, they consist of oscillating electric and magnetic fields which are at right angles to each other and at right angles, or transverse, to the direction of propagation.

b. They require no supporting medium.

c. They may be reflected, refracted and diffracted, and are subject to interference and Doppler effects.

d. Unlike light waves, radio waves will pass through many opaque bodies, eg walls, buildings, trees, fog, cloud, rain, although they suffer some attenuation in doing so. The magnitude of the loss is very dependent on the frequency used, being greatest at the highest radio frequencies.

Propagation Around the Earth

7. Fig 4 shows the principal paths which radio waves may follow above the Earth, between a transmitter and a receiver:

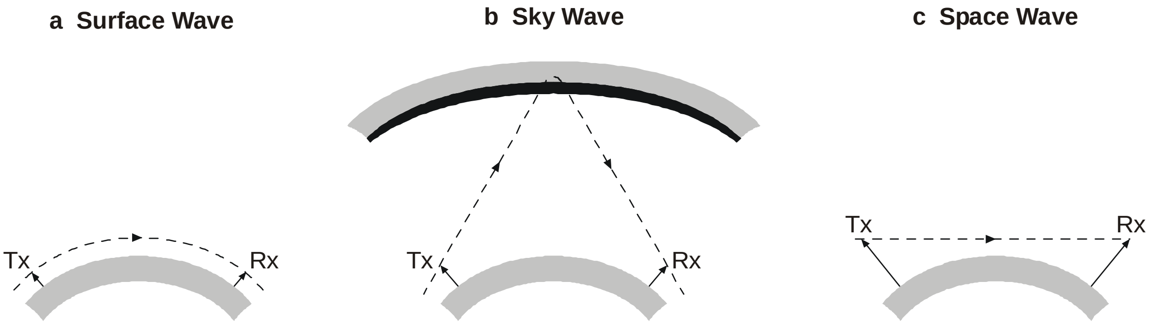
a. The surface wave, which follows the Earth's contour (Fig 4a).

b. The sky wave, which returns after reflection from the ionosphere (Fig 4b).

c. The space wave, which travels in a direct line (Fig 4c).

The ground wave is a combination of the space and surface waves.

14-21 Fig 4 Radio Wave Paths



8. The radio energy reaching a receiver may be made up of components following any one or more of these paths but, depending on the part of the spectrum concerned, one of the three will usually predominate. Very roughly, radio waves from the lower part of the radio spectrum are propagated mainly by surface wave, from the middle of the spectrum by sky wave, and from the upper part of the spectrum by space wave.

Surface Wave Propagation

9. Because of the phenomenon known as diffraction, the surface wave follows the Earth's curvature. This diffraction is assisted by the Earth's attenuation of the radio energy. Therefore, the wave-front in the direction of motion will lag at the surface. Surface wave propagation is of practical importance only in the VLF, LF and, to some extent, in the MF bands.

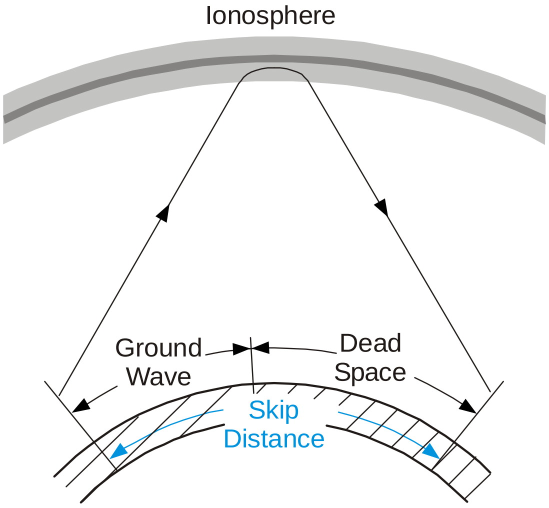
10. **Surface Wave Application**. The extreme stability of low frequency surface wave propagation makes it particularly suited to systems requiring consistency of signal over long distances, such as high-grade communications and certain classes of navigation aid. Surface wave propagation at very low frequencies is also suited to submarine communication because maximum penetration of the surface of the sea is possible. The objections to this type of propagation are the physical size of the aerials, the high cost of transmitting stations and the considerable power required to offset ground attenuation.

Sky Wave Propagation

11. Sky waves are those which ascend into the upper atmosphere and encounter a region containing electrically charged particles (the ionosphere) where they are refracted sufficiently to be returned to Earth. When waves are refracted, they change direction due to a change in velocity. They may then

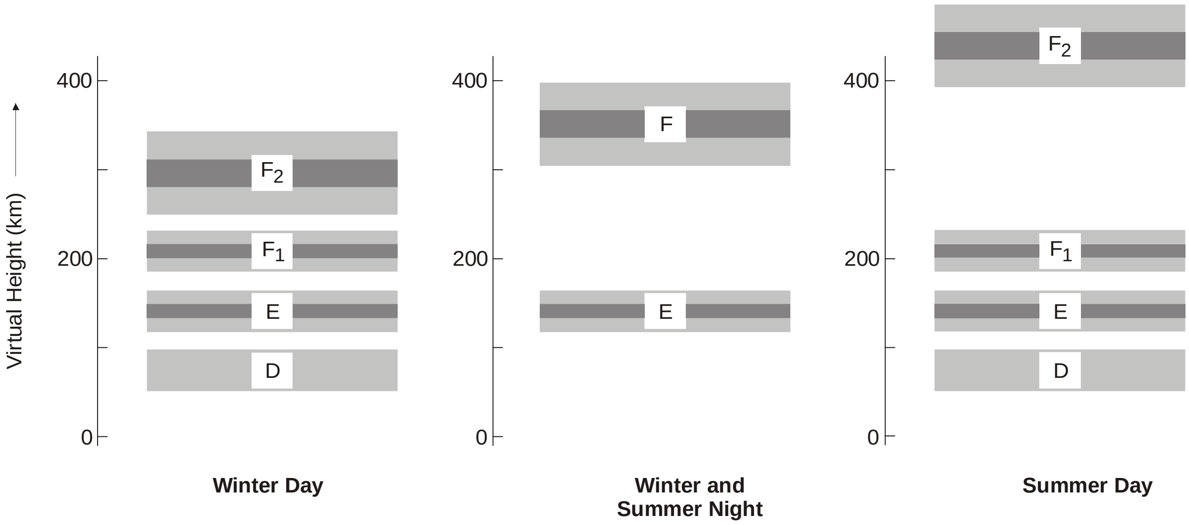
be reflected upward and travel again to the ionosphere. The process may be repeated a number of times, and it is multi-hop transmission of this type which makes it possible to communicate with points on the other side of the globe at HF. As shown in Fig 5, the distance between the transmitter and the point of reception of the first sky wave is known as the 'skip distance'. The area between this point and the limit of the ground wave is known as the 'dead space'.

14-21 Fig 5 Skip Distance and Dead Space



12. **The Ionosphere.** The ionosphere, which consists of a number of conducting layers, generally between heights of about 60 and 400 km, depends for its existence largely on ultra-violet radiations from the sun. For this reason, both the heights and densities of the layers vary according to the time of day and season of year, as shown in Fig 6. There is also a long period cycle of about 11.1 years, connected with sunspot activity. Finally, a number of short-term effects occur in a more or less random fashion, and these result in the ionized layers being in a state of continuous turbulence. Experimental measurements have shown that there are three main layers, and these are designated by the letters D, E and F in order of ascending height. At times, the F layer splits into two separate layers (F_1 and F_2). The D layer is a region of low ionization which only persists by day; the E layer is more marked and remains weakly ionized by night, with little change of height, while the F layer is the most strongly ionized, and has the greatest diurnal variations in height.

14-21 Fig 6 Ionospheric Layer



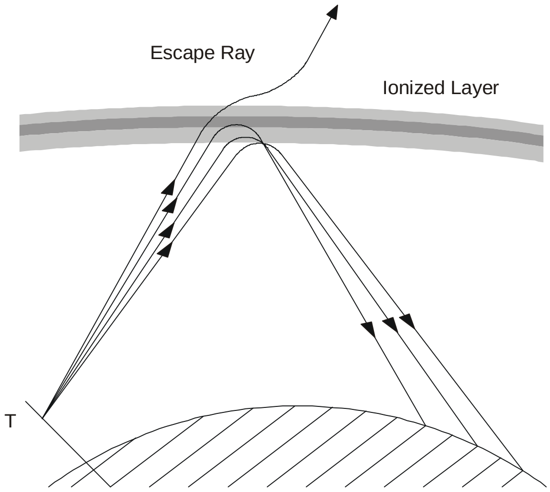
13. **Propagation in the Ionosphere**. The effect of the ionosphere on wave propagation is strongly dependent on frequency:

a. **Low Frequencies**. At frequencies below about 100 kHz, the D layer acts as an almost perfect reflecting surface and, at great distances, the wave propagates as though travelling between two concentric conducting spheres.

b. **Medium and High Frequencies**. At medium and high frequencies, a wave penetrates to the more strongly ionized layers. As it enters one of these layers, it is progressively refracted away from the normal. Whether or not it is sufficiently refracted to be returned to Earth depends on:

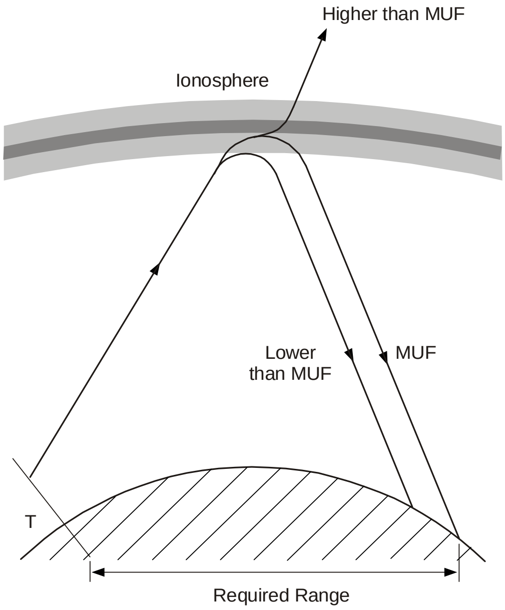
(1) **Angle of Incidence**. Fig 7 shows that if the angle of incidence with the vertical is reduced, the wave penetrates more deeply into the layer before being returned. Finally, an angle is reached which allows the wave to penetrate the layer, and escape. This is known as the critical angle for that particular frequency and that particular layer.

14-21 Fig 7 Effect of Angle of Incidence (Frequency Constant)



(2) **Frequency**. For a given angle of incidence, if the frequency is increased, a sky wave penetrates more deeply into a layer before being returned to Earth. The highest frequency at which reflection can take place for a given angle of incidence is known as the maximum usable frequency (MUF). The optimum condition for communication between two points is achieved by operating close to the MUF. Using a lower frequency results in a loss of signal strength, while using a higher frequency allows the waves to escape without reflection (see Fig 8). In practice, to ensure reliability, a frequency of about 85% of the MUF is chosen. This is called the optimum working frequency (OWF).

14-21 Fig 8 Effect of Frequency (Angle of Incidence Constant)



c. **Very High Frequencies**. Above about 30 MHz, the waves are never sufficiently refracted to be returned to Earth and they escape into outer space.

14. **Sky Wave Applications**. The main application for sky waves is in the HF band, for medium and long range point-to-point communications systems. Compared with low frequency surface wave propagation, the power required is very much lower, but the same degree of signal stability cannot be achieved and, even with frequent changes of frequency, it is not normally possible to maintain continuous communication. Certain types of long range navigation aids also employ sky wave propagation (eg LORAN), but here the requirements of extensive coverage preclude the possibility of operating at the MUF.

Space Wave Propagation

15. Transmissions at VHF and above cannot propagate by either surface or sky wave, and so, at these frequencies, energy can normally only reach a receiver by a direct path plus an Earth-reflected path. The combination of the two components is called the space wave. Propagation by this mechanism is of special importance because of its application to most forms of radar.

16. The main characteristics of space wave propagation are that:

a. The radiation field is confined to the region above the radio horizon.

b. Within this field, there may be marked variations in signal level, due to the interference effects between the direct and Earth-reflected components.

17. **Radio Horizon**. The refractive index of the atmosphere is not constant, but normally decreases uniformly with height towards unity. The effect of this variation on radio waves transmitted at low angles of elevation, is to bend them downwards with the result that, in a standard atmosphere, the distance to the radio horizon is about 15% greater than to the geometric horizon. To facilitate calculations concerned with coverage and performance, it is usual to allow for this effect by substituting a fictitious value for the radius of the Earth (4/3 the actual value in a standard atmosphere) and then assuming that the waves travel in straight lines.

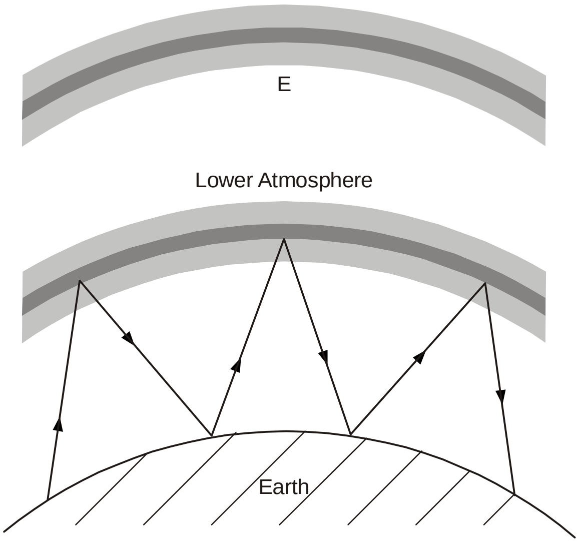
18. **Space Wave Applications**. With two exceptions, to be dealt with below, all propagation at VHF and above is by space wave. This includes short-range communication, television and FM sound broadcasting, radio-relay, radar, navigational and approach aids, and communication by satellites.

Other Propagation Modes

19. Transmissions at VHF and above are normally confined to the region above the radio horizon, but there are two lesser propagation modes by means of which radio energy at these frequencies may be able to penetrate into the region of shadow. The first, known as 'duct propagation', only occurs under abnormal atmospheric conditions, and is therefore of little practical value. The second, known as 'scatter propagation', exists continuously, but can only be exploited by special techniques.

20. **Duct Propagation**. Under certain abnormal climatic conditions, temperature can increase while humidity decreases with height. When this occurs, the refractive index may decrease with height much more rapidly than is normal, and a duct is formed between the Earth and a hundred or so feet above it. Radio waves of lengths which are small compared with the duct height (usually metric and below in temperate latitudes) are then sufficiently refracted to be returned to the surface, and progress in a series of reflections in the manner shown in Fig 9. The trapping of radio energy in this way, results in marked anomalies in performance, and the strong signals which may sometimes be received well beyond the normal radio horizon are often a source of nuisance, particularly in radar. A marked high-level inversion can sometimes produce an elevated duct.

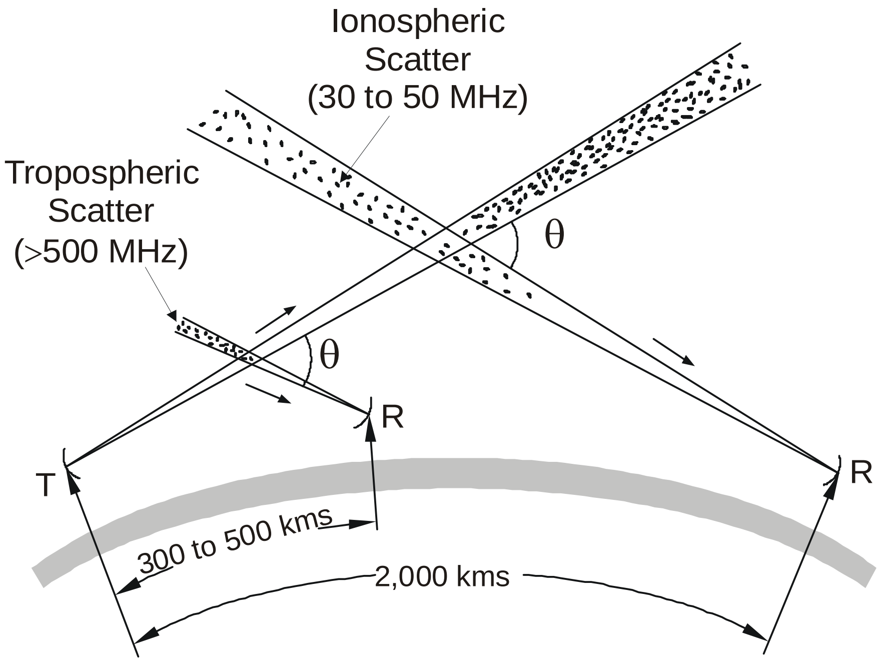
14-21 Fig 9 Duct Propagation



21. **Scatter Propagation**. A radio wave penetrating upwards into the atmosphere encounters regions of turbulence which contain minor local variations in refractive index. These cause a small proportion of the energy to be deflected away from the main wave path. Most of the deflected energy is contained within a conical volume lying along the wave axis with its apex in the main scattering region and, provided the elevation of the wave path is sufficiently low, some of the deflected energy will be returned to Earth. The signal level in the returned wave may be as much as 100 dB below the free space signal for the direct path. To exploit this form of propagation for point-to-point communication, it is necessary to employ highly directive transmitting and receiving aerials with their beams orientated so as to intersect at a small angle in the optimum scattering region. Two such regions exist, the first in the troposphere, where the effect is marked in the UHF band above 500 MHz, and the second in the ionosphere, where the scattering is confined to a relatively narrow band of frequencies in the VHF

band, between about 30 and 50 MHz. As shown in Fig 10, the resultant geometry dictates fairly distinct limits to the usable transmission distances; the tropospheric links normally operating over distances of 300 to 500 km (160 to 270 nm) and ionospheric links out to about 2,000 km (1,100 nm).

14-21 Fig 10 Scatter Propagation



22. **Summary**. Table 1 summarizes the propagation characteristics and typical uses of frequency bands in the radio spectrum.

Table 1 Main Features of the Different Frequency Bands

Frequency Band	Type of Propagation	Typical Uses
VLF (3 to 30 kHz)	Ground wave and sky wave.	Reliable long distance communication. Communication with submerged submarines.
LF (30 to 300 kHz)	Ground wave and sky wave.	Medium and long range navigation aids.
MF (300 to 3,000 kHz)	Ground wave and sky wave. The reception of ground waves is reliable up to medium ranges (200 nm). Sky wave reception, though achieved up to thousands of miles, is unpredictable.	Broadcast stations. Medium and long range navigation aids: eg NDBs.
HF (3 to 30 MHz)	Sky wave: single and multi-hop ranges up to a few thousand miles.	Reliable CW and RT point-to-point communications over long distances.
VHF (30 to 300 MHz)	Space wave: within the radio horizon.	Marker beacons. ILS localizer. Radio telephone. Navigation aids: VOR, PLB.
	Scatter.	Ionospheric scatter communication.
UHF (300 to 3,000 MHz)	Space wave: within the radio horizon.	Radio telephone. ILS glide path. Pulse radio altimeters. TACAN, VORTAC. Search and GCI radars.
	Scatter.	Tropospheric scatter communications. Satellite Communications.
SHF (3 to 30 GHz)	Space wave: within the radio horizon.	Airborne map painting radars. Cloud warning Radar. FMCW radio altimeters. Doppler navigation aids in I and J bands. Weapon control radars Satellite communications.

CHAPTER 22 - TELECOMMUNICATIONS PRINCIPLES

Introduction

1. The purpose of any communications system is to convey information accurately, and usually rapidly, from one place to another. For communications over long distances the only practical method is to use electrical energy, in the form of electromagnetic waves in free space, or directed by transmission lines. Although this chapter will concentrate on radio communication, many of the ideas are equally applicable to other types of communication systems.

Information Types

2. The information to be conveyed by a communication system may originate from a variety of sources and may take one of many forms, e.g. speech, writing, still or moving pictures, digital data. Whatever the form that the information takes, there will be some physical phenomena associated with it which vary with time, for example, speech causes variations in air pressure, and pictures are associated with variations in light intensity. These physical variations must be converted into analogous variations in an electrical signal and the device which performs this transformation is called a transducer. Thus, for example, a microphone is a transducer which converts variations in sound pressure into a varying electrical signal. Table 1 lists a number of types of information, the associated measured phenomenon, and the name of the communication system by which each would normally be conveyed. The portion of the system in which the signal is propagated by electromagnetic means, e.g. in the form of a wave guided by a conductor or unguided in free space, is known as the 'channel', and any communication channel in which information is transmitted by radio is known as radio communications.

3. The information before conversion is usually referred to as the 'message', and after conversion as the 'baseband signal'. When the signal is received at the destination another transducer is needed to convert the baseband signal into a suitable form, which may be the same as, or different from, the original form.

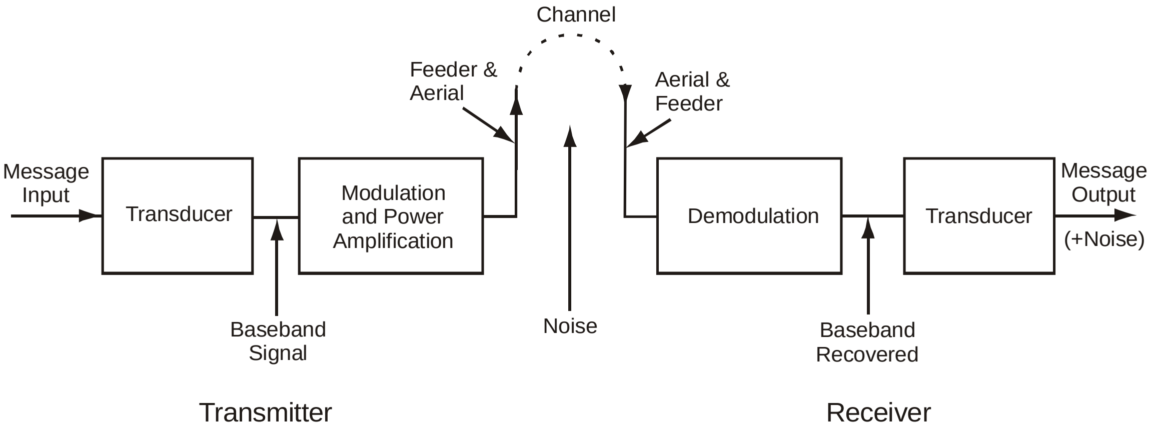
Table 1 Information Sources

Information	System	Measured Phenomenon
Spoken word	Telephony	Variations in air pressure
Coded words/symbols	Telegraphy	Position of holes in punched tape Open or closed switch (Morse key)
Still pictures	Facsimile	Variations in light intensity
Moving pictures	Television	Variations in light intensity
Monitored data e.g.: Acceleration Structural strain Pressure Temperature Control Angles	Telemetry	Mechanical or electrical analogue
Control Data e.g.: Elevator angle Rudder angle Aileron angle Power setting	Telecontrol	Mechanical/electrical analogue of: Height or pitch angle error Heading error Heading of roll angle error Speed error

Basic Radio Communication System

4. The essential processes involved in radio communication are common to all systems and may be described in relation to the basic system shown in Fig 1.

14-22 Fig 1 A Basic Radio System



5. **Input Transducer**. The input transducer is acted upon by the information to be transmitted and produces an electrical equivalent, the baseband signal, consisting of a spectrum of frequencies. In the case of speech this spectrum usually contains components ranging from about 300 Hz to 3 kHz. Music signals occupy a somewhat wider range from about 30 Hz to 15 kHz or more, while television has a baseband width in the region of 5 MHz. There are a number of transducer types to cater for the variety of information sources and they employ a range of techniques such as variation of impedance, voltage, and frequency. Some examples of input transducers, with the varying parameter, are:

a. Microphone - Sound.

b. Photo-electric cell - Light intensity.

c. Potentiometer - Mechanical movement.

d. Strain gauge - Structural strain.

e. Piezo-electric devices - Pressure or force.

f. Thermocouple - Temperature.

6. **Transmitting Equipment**. The spectrum of frequencies contained in the baseband signal is unsuitable for direct radio transmission. The essential process performed by the transmitting equipment is, therefore, the translation of the baseband to a suitable part of the frequency spectrum for transmission. This is achieved by the process known as modulation in which the characteristics of the baseband signal are impressed on to a radio frequency (RF) carrier wave. Before application to the aerial, the modulated wave is then amplified to the required power level.

7. **Feeder**. The feeder conveys the radio frequency energy from the transmitter to the aerial and, ideally, it should neither absorb nor radiate energy. For radio frequencies up to about 100 MHz (3 m wavelength) an open twin-wire feeder may be employed, but above this frequency coaxial cable is used to avoid radiating RF power. Losses from coaxial cable increase as frequency increases and, although coaxial can be used up to about 3 GHz (10 cm wavelength), the normal feeder for frequencies above 1 GHz (30 cm wavelength), which includes the majority of radar systems, is the waveguide.

8. **Transmitting Aerial**. The transmitting aerial should, ideally, radiate the whole of the power conveyed to it without reflecting any back into the transmitter, and it should do this across the entire band of frequencies contained in the RF signal spectrum. In many cases the aerial is required to concentrate the radiated power in a chosen direction with a minimum of wastage in other directions. The physical dimensions of the aerial are related to the transmission wavelength and if the aerial is small in relation to the wavelength, as is the case in the VLF and LF bands, the radiation efficiency will be low.

9. **The Channel**. The radiated signal propagates between the transmitting source and its destination by the propagation phenomena applicable to the part of the spectrum used. In the course of its passage the signal becomes modified by reflection, refraction, attenuation and other effects, and is further contaminated by unwanted radio noise occurring within the bandwidth of the receiving equipment.

10. **Receiving Aerial**. The receiving aerial performs the reciprocal function of the transmitting aerial and is normally arranged to receive most strongly from some favoured direction with a minimum response in other directions.

11. **Receiving Equipment**. The power in the received signal may amount to less than a microwatt and the first function of the receiving equipment is therefore amplification, which it should achieve equally across the range of frequencies contained in the RF signal. Some of the amplification may be carried out directly on the received RF signal, but the greater part is performed after translating the signal to an intermediate (IF) frequency. After amplification, the original baseband signal, now contaminated with noise, is recovered by the process of demodulation, and after further amplification, this is passed to the output transducer. The faithfulness with which the original baseband signal is reproduced and the extent to which it is contaminated with noise is closely related to the bandwidth of the receiving equipment, ie the range of frequencies capable of being amplified in the IF stage.

12. **Output Transducer**. The function of the output transducer is to convert the baseband signal into a form suitable for the recipient. In many cases this form will be the same as the original message, but this is not necessarily the case. Examples of output transducers are:

a. Loudspeaker.

b. Cathode ray tube (CRT).

c. Galvanometer.

d. Electro-mechanical servo.

e. Printer.

Factors Affecting Communication

13. Among the more important characteristics of communication systems to be considered are the channel bandwidth they occupy and the power they must utilize to achieve their purpose. There is a fundamental relationship between the channel bandwidth required and the rate at which information is generated by the source. In addition, allowance has to be made for the type of signal (analogue or digital) used to represent the information and for the method of modulation employed. Operating range, on the other hand, depends on power and on the degree of contamination by noise which can be accepted. The extent to which the signal may be contaminated by noise without affecting the ability to recover the message depends in turn on the type of signal (analogue or digital), on the method of modulation, and on special measures which may be taken to reduce susceptibility to noise.

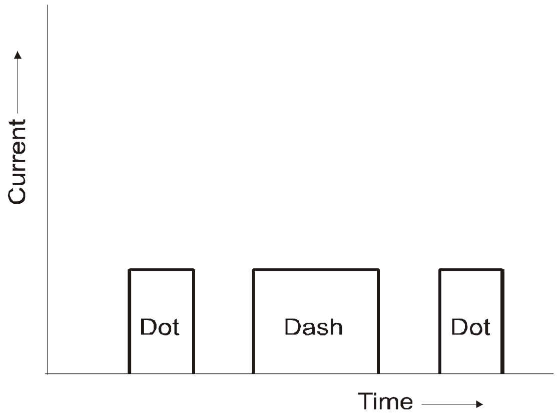
Types of Signal

14. The information to be transmitted may be classified broadly into two types. Some information sources consist of a variable quantity which changes continuously with time, such as speech, and the baseband representation of such a source will be a continually varying voltage. Such a signal is known as an analogue signal. Conversely some information may take a series of discrete permitted values with time and is known as a digital signal. One of the most familiar and simple examples is the Morse code, where each alphanumeric symbol is represented by a series of dots or dashes. Such a code has only two values and may be transmitted by simply switching the transmitter on and off; the amplitude of the transmitted signal does not convey any extra information and may be constant. More complex digital systems have more than just the two (on and off) levels of the Morse code.

Basic Telegraphy Systems

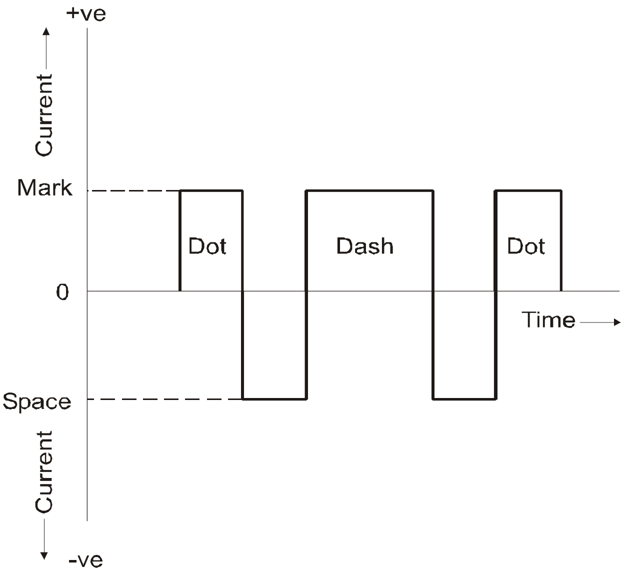
15. A simple code technique in which the transmitter is switched on and off is known as single current working, and is illustrated in Fig 2. The system is said to be 'marking' when the current is flowing, and 'spacing' when it is switched off. An alternative system, known as double current working, may be used where the signal is formed by reversing the direction of the current as shown in Fig 3; the terms 'mark' and 'space' are still used.

14-22 Fig 2 Single Current Working



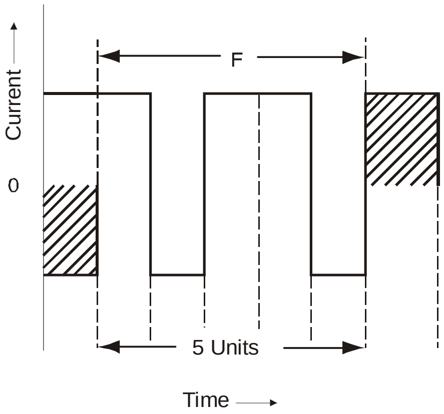
16. The bandwidth of a communications system must be related to the frequencies of the signals to be handled. The square waveforms of Fig 2 and 3 may be analysed into a large number of sinusoidal waves comprising a fundamental frequency and an infinite number of odd harmonics. Thus, theoretically, the telegraph system requires an infinite bandwidth for perfect reproduction. However, since there is a limit to the working bandwidth of the equipments and links, the waveforms must be limited in bandwidth without introducing unacceptable distortion. A reasonable compromise between bandwidth and distortion is obtained by filtering out all harmonics above the third. When this is done it is found that a system bandwidth of 50 Hz suffices for speed of Morse transmission of up to 30 words per minute; the required bandwidth increasing with increasing transmission rate.

14-22 Fig 3 Double Current Working



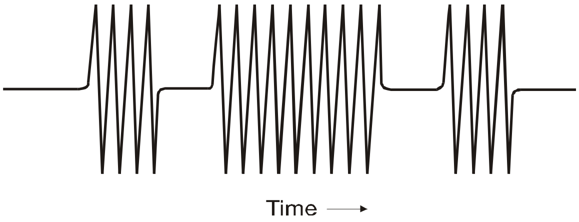
17. The unequal lengths of letters in the Morse code makes it unsuitable for teleprinter operation, and in the RAF a 5-unit code, the Murray code, is used instead. In this code the number of elements is the same for each character and the duration of each element is constant at 20 milliseconds. Each element is in one of two states, either 'mark' or 'space'. Fig 4 shows an example of the Murray code for the letter 'F'. Since the time for each element is 20 milliseconds, the time for two adjacent elements is 40 milliseconds. This is the time period of the signal, which is equivalent to a fundamental frequency of 25 Hz. If up to the third harmonic (75 Hz) is included, a system bandwidth of 80 Hz would suffice.

14-22 Fig 4 Unit 'Murray' Code

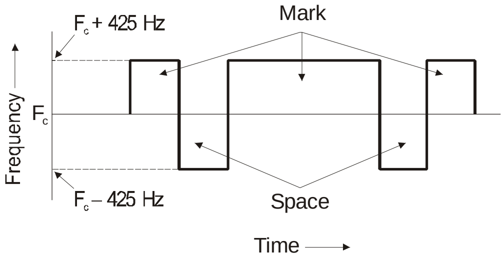


18. In radio telegraphy the coded signal uses a radio link between transmitter and receiver. If the code is transmitted by switching the transmitter on and off in sympathy with the dots and dashes, known as keying, the method of operation is called interrupted continuous wave (ICW) and is shown in Fig 5. This technique has, however, been largely superseded by a method known as frequency shift keying (FSK). Instead of interrupting a continuous train of RF energy, the aerial radiates continuously, but the frequency is dependent upon whether a mark or space is being transmitted. The frequency separation between a mark and space is standardized at 850 Hz in the HF band, ie a mark impulse is radiated on a frequency 425 Hz higher than the nominal frequency (f_c), and a space on a frequency 425 Hz below it (Fig 6). A third method of transmitting the code is to change the phase of the transmitted signal, most commonly by 180^o, whenever there is a change of state between mark and space. This is known as phase shift keying or phase reversal keying (PSK or PRK) and is illustrated in Fig 7.

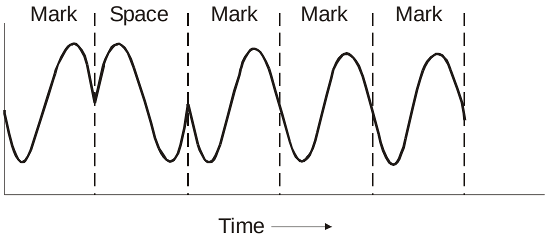
14-22 Fig 5 Interrupted Continuous Wave



14-22 Fig 6 Frequency Shift Keying



14-22 Fig 7 Phase Shift Keying



Modulation

19. In each of the three methods of transmitting a code described, one parameter of the RF signal is changed to convey the information. In ICW the amplitude varies between the normal level and zero, in FSK the frequency varies above and below the nominal frequency, and in PSK the phase alters to indicate a change of state. Each is an example of modulation; amplitude, frequency, and phase modulation respectively. Although phase modulation is confined to the transmission of digital signals as illustrated in Fig 7, amplitude and frequency modulation techniques are also applicable to analogue signals. A further method of transmitting an analogue signal is to convert it into a digital signal, and then use one of the techniques of pulse modulation.

20. The need for modulation in a radio system arises because the range of frequencies in the baseband signal is not the same as the range of frequencies which can be transmitted efficiently in free space. The baseband signal contains frequencies in the audio range whereas radio transmission must operate with frequencies from about 12 kHz upwards.

21. If two or more signals occupy the same frequency band, eg speech from separate sources, transmission by free space is possible if each signal is arranged to modulate a separate high frequency carrier. The separation of signals at the receiver then becomes a matter of separating the carriers, ie tuning.

AMPLITUDE MODULATION (AM)

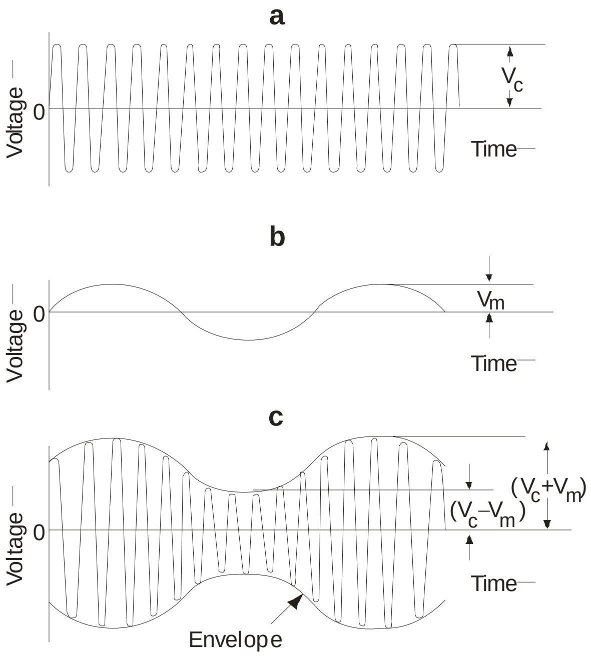
Principle

22. In amplitude modulation the amplitude of the carrier frequency is changed to reflect changes in the baseband signal, the carrier frequency and phase remaining constant. Fig 8 shows the principle:

a. Fig 8a represents the unmodulated continuous wave carrier radiation of constant frequency, f_c, and constant amplitude, V_c.

b. Fig 8b shows the waveform of the baseband signal, an audio frequency modulating voltage of frequency f_m and amplitude V_m. The baseband signal will, in practice, rarely be a simple pure sine wave, but this form serves to illustrate the concept clearly.

14-22 Fig 8 Amplitude-modulated Waveform in Time Domain



c. Fig 8c illustrates the waveform of the RF carrier after amplitude modulation by the audio frequency modulating signal. The outline of the modulated wave (dotted line), known as the modulation envelope, is an exact replica of the modulating signal. In the case of speech, for example, although the modulating waveform is more complex, the modulation envelope must still replicate the waveform if distortion is to be avoided.

23. **Depth of Modulation**. The depth of modulation (m) is defined as the ratio of the amplitude of the modulating signal to that of the carrier wave, usually expressed as a percentage, ie, with reference to Fig 8:

$$m = \frac{V_m}{V_c} \times 100\%$$

At 100% modulation, $V_m = V_c$ and so the amplitude of the modulated wave varies between zero and $2V_c$. This is the maximum depth of modulation which may be used without causing distortion since if V_m were to exceed V_c overmodulation would occur. The modulated carrier would then have zero amplitude for part of the audio cycle, and the modulation envelope would no longer be a replica of the modulating signal, ie distortion would have occurred.

Frequency Analysis

24. It is not possible, from consideration of Fig 8, to determine the bandwidth associated with the waveform. In order to do this it becomes necessary to consider an alternative representation, in the frequency domain rather than in the time domain. This involves a diagrammatic representation of the sinusoidal frequency components which together constitute the amplitude modulated waveform.

25. It can be shown that the amplitude modulated waveform of Fig 8 is equivalent to the sum of the three following radio frequency waves of constant amplitude:

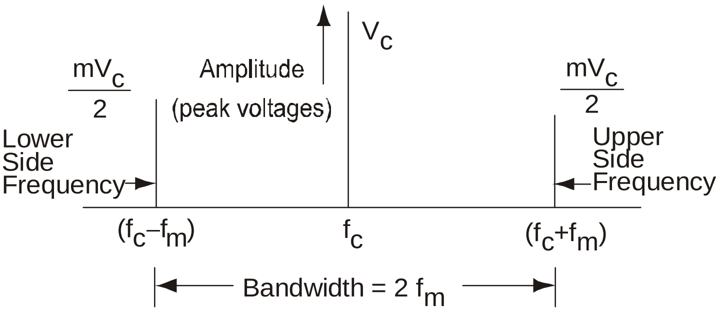
a. The original carrier of frequency f_c and amplitude V_c.

b. An RF wave of frequency ($f_c + f_m$) and amplitude $\frac{mV_c}{2}$ known as the upper side frequency.

c. An RF wave of frequency ($f_c – f_m$) and amplitude $\frac{mV_c}{2}$ known as the lower side frequency.

These radio frequencies can be represented diagrammatically, as in Fig 9, and from this diagram it will be seen that the modulated waveform covers a frequency band from ($f_c – f_m$) to ($f_c + f_m$), ie a bandwidth of twice the modulating frequency.

14-22 Fig 9 Amplitude Modulated Waveform in the Frequency Domain



26. The carrier wave itself contains none of the intelligence of the modulating signal, all of which is contained in the side frequencies. The power in each sinusoidal component is proportional to the square of its amplitude, thus:

$$\text{Power in carrier } \propto V_c$$

$$\text{Power in each sideband } \propto \left(\frac{mV_c}{2}\right)^2$$

The efficiency of the system may therefore be defined by:

$$\text{Efficiency} = \frac{\text{Power in Sidebands}}{\text{Total Power}}$$

$$= \frac{2 \times \left(\frac{mV_c}{2}\right)^2}{2 \times \left(\frac{mV_c}{2}\right)^2 + V_c{}^2}$$

$$= \frac{m^2}{2 + m^2}$$

27. Thus the efficiency depends upon the depth of modulation, and even at 100% modulation (m = 1) is only ⅓, decreasing with decreasing depth of modulation.

28. **Speech Clipping**. Speech waveforms contain a number of isolated peaks whose amplitude is very much in excess of the mean level. If the transmitter is adjusted to give 100% modulation on the peaks the mean level of modulation is extremely low, which means that the transmitter is being operated below maximum efficiency for the major part of the intelligence. Speech clipping is a

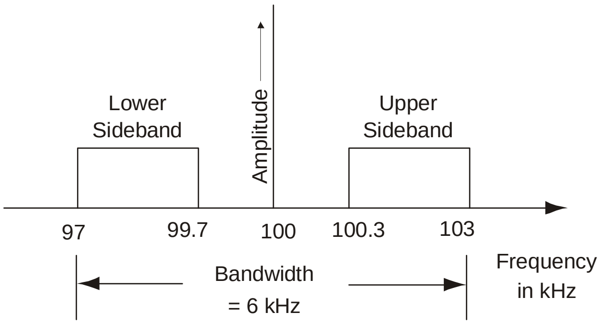
technique to alleviate this problem by limiting the isolated peaks to some suitable value thus allowing the mean level of modulation to be increased.

29. **Sidebands**. In radio telephony transmission the modulating signal is not a single sinusoidal frequency but a complex of many simultaneous audio frequencies, usually ranging from about 300 Hz to about 3 kHz. When an RF carrier is amplitude modulated by such a signal a band of frequencies is produced above and below f_c, a pair of side frequencies being produced for every frequency component of the speech. The band of frequencies below f_c is termed the lower sideband, and that above f_c the upper sideband. Fig 10 illustrates these bands when an RF carrier of 100 kHz is amplitude modulated by speech frequencies in the range 300 Hz to 3 kHz. Since two sidebands are obtained from each baseband signal the modulation process is often termed double sideband amplitude modulation (DSB AM). In the example of Fig 10, for the original speech to be faithfully reproduced at the receiving end of the communication system all parts of the system must be capable of passing equally well all frequencies from 97 kHz to 103 kHz, a bandwidth of 6 kHz. In the general case, the bandwidth will be twice the highest frequency of the modulating signal.

Single Sideband (SSB) Systems

30. The double sideband system is not very efficient since the carrier, which uses power, contains no intelligence, and the same intelligence is present in both the upper and lower sidebands. Even at 100% modulation the proportion of the total transmitted power in the sidebands is only ⅓, and this decreases at lower modulation levels. A single sideband system overcomes these inefficiencies by suppressing the carrier (or reducing it in amplitude) and by transmitting only one of the sidebands. Conventionally the upper sideband is used.

14-22 Fig 10 Sidebands of Amplitude Modulated Wave



31. The envelope of an SSB signal is not a replica of the intelligence and, at the receiver, a locally generated sine wave at the carrier frequency must be added to the received signal before demodulation can be accomplished. For maximum fidelity, this locally generated carrier must be large in amplitude compared with the sideband amplitude and identical to the original transmitter carrier both in frequency and phase. For speech transmission phase shifts have negligible effects, but frequency errors in excess of 50 Hz can destroy intelligibility. The required frequency stability can only be achieved using crystal controlled frequency synthesizers.

32. There are three types of SSB system:

a. **Pilot Carrier System**. In the pilot carrier system a very small carrier signal is transmitted with the selected sideband. Very little power is wasted since the carrier amplitude is only a small fraction of the basic carrier amplitude. In the receiver the carrier and sideband are separated by highly selective filters and amplified separately - the carrier to a high level, the sideband to a lesser degree. The two signals are then recombined in the SSB demodulator.

b. **Controlled Carrier System**. In the controlled carrier system the carrier is transmitted at approximately full amplitude during the brief pauses in speech, or between syllables of speech, and is reduced to a very low level during actual modulation. The level of this controlled carrier is such that the average power output of the transmitter is maintained effectively constant regardless of the presence or absence of modulation. In the receiver, the short bursts of carrier are used to synchronize an oscillator which generates the necessary carrier frequency waveform used in the demodulator.

c. **Suppressed Carrier System**. In the suppressed carrier system no carrier is transmitted and it is therefore the most efficient method since all the transmitted power is contained in one sideband. A highly stable oscillator within the receiver generates a carrier frequency signal which is added to the received sideband in the demodulation process. This method requires the highest possible frequency stability throughout the system and matched crystal oscillators, in thermostatically controlled ovens, are used in both ground and airborne equipments.

33. **Bandwidth**. The bandwidth required for DSB AM transmission is twice the maximum modulating frequency, however an SSB system requires only half of this. It is therefore possible to allocate more channels within a given band of frequencies, or the spacing between channels can be increased to reduce the possibility of adjacent channel interference. Since the receiver bandwidth is half that of a DSB system, noise amplitude is halved.

34. **Fading**. Long range HF communication is usually accomplished using sky wave propagation and the received signal will often have components which have travelled different path lengths. The strength of the received signal is the result of the addition of these various components and, depending on phase differences, there may be variation between complete reinforcement and complete cancellation. With DSB transmission the frequency and phase of any frequency component in the upper sideband should be the same as that of the corresponding component in the lower sideband. Any difference will result in fading and this 'sideband unbalance' can produce so much distortion as to make the signal unintelligible. Although selective fading can still occur with SSB transmission, sideband unbalance does not add to the problem, and any distortion tends to make the voice signals sound odd but not unintelligible.

FREQUENCY MODULATION (FM)

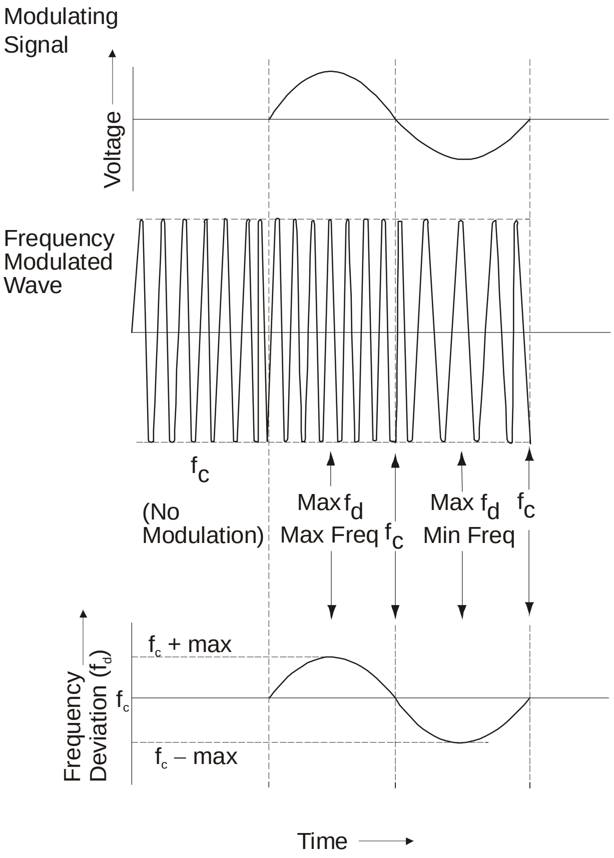
Principle

35. In frequency modulation, the instantaneous frequency of the carrier is caused to vary above and below its nominal value in accordance with variations in the modulating signal. The resulting signal is more complex than the corresponding AM signal and for clarity only the effect of a single sinusoidal waveform-modulating signal will be examined. The amplitude of the resulting FM signal remains constant and this means that the efficiency of the RF power amplifier in the transmitter is constant and may approach the theoretical maximum.

36. Fig 11 shows the effect of modulating a carrier (f_c) by a single frequency signal (f_m), with an initial period of zero modulation. The positive half cycle of the modulation gives a frequency increase and the negative half cycle a frequency decrease. The frequency deviation (f_d) from the nominal carrier frequency at any instant is proportional to the instantaneous voltage of the modulating signal and the

maximum deviation thus corresponds to the maximum amplitude of the modulating signal. The rate at which the carrier frequency is varied depends on the frequency of the modulating signal.

14-22 Fig 11 FM in the Time Domain



37. **Modulation Index**. The level of modulation is indicated by the modulation index (β) where:

$$\beta = \frac{\text{Maximum Frequency Deviation}}{\text{Modulating Frequency}}$$

$$= \frac{f_{max} - f_c}{f_m}$$

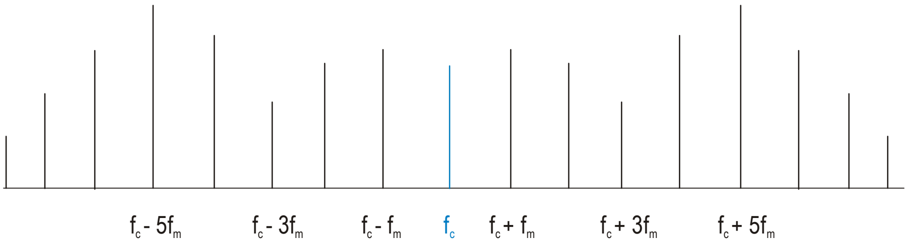
$$= \frac{\max f_d}{f_m}$$

β is usually greater than unity.

FM in the Frequency Domain

38. Mathematical analysis of the modulated waveform will show that it consists of the nominal carrier frequency (f_c) and an infinite number of side frequencies spaced at intervals equal to the modulating frequency (f_m); Fig 12 shows an example.

14-22 Fig 12 FM in the Frequency Domain



39. The carrier component is not necessarily the largest amplitude component and the amplitude of the side frequencies in any particular case depends upon the modulation index. The side frequencies appear in pairs, of equal amplitude, spaced above and below f_c by a multiple of f_m (ie $f_c \pm f_m$, $f_c \pm 2f_m$, $f_c \pm 3f_m$ etc). Relatively close to f_c the amplitude of the side frequencies does not necessarily decrease as the spacing from f_c increases, although this pattern does arise beyond a certain point (eg beyond $f_c \pm 5f_m$ in Fig 12).

Bandwidth

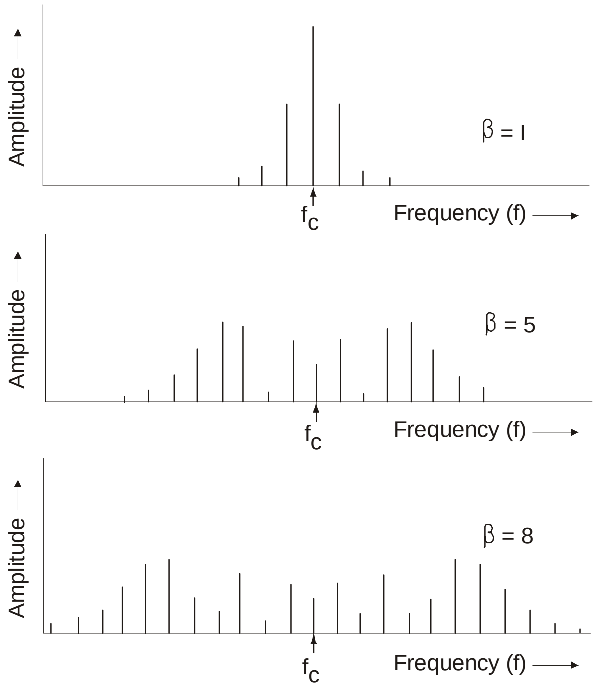
40. Since there are an infinite number of side frequencies, the bandwidth of an FM signal is also theoretically infinite. However, side frequencies well away from f_c gradually decrease in amplitude and there will come a point where the amplitudes are so small that the components can be safely ignored. Therefore in practice the bandwidth can be limited to that of significant side frequencies, which are usually considered to be those whose amplitude is greater than 1% of the amplitude of the unmodulated carrier.

41. The number of significant side frequencies, and therefore the bandwidth, depends on β, and some examples of typical spectra, for a single modulating frequency f_m, and for different values of β, are shown in Fig 13. As an example, VHF sound radio has a maximum deviation set at ±75 kHz and the range of f_m is from 50 Hz to 15 kHz. Therefore:

$$\beta = \frac{75\,\text{kHz}}{15\,\text{kHz}} = 5$$

From Fig 13 it will be seen that there are eight pairs of side frequencies in this case and thus the bandwidth is 15 kHz × 16 = 240 kHz. In practice, the content of frequencies above 10 kHz in normal music is small and, in order to conserve bandwidth, the channel allocation is limited to 200 kHz. The resulting slight distortion of high frequencies is undetectable.

14-22 Fig 13 FM in the Frequency Domain



42. An approximation to the bandwidth can be gained by using the empirical formula:

$$B = 2(\max f_d + \max f_m)$$

where f_d = frequency deviation , and f_m = modulating frequency.

43. Because the bandwidth of FM is wide, the system is normally used at VHF or above, since the bandwidth cannot be accommodated at lower frequencies. The use of very high frequencies restricts FM to line of sight communications.

Noise Suppression and Capture Effect

44. Noise superimposed on an FM signal causes unwanted modulation in both amplitude and phase. However, since the amplitude of an FM signal carries no information, noise due to amplitude variations can be eliminated by designing the receiver so that amplitude fluctuations in the received signal produce no receiver output. Noise due to phase variation is negligible providing the frequency deviation of the wanted signal is large. Noise interference might typically produce a frequency deviation of 30 Hz which is negligible compared with a maximum frequency deviation in the signal of, say, 75 kHz. Thus noise is almost completely suppressed even if it is only slightly weaker than the signal, and the suppression increases with increasing signal frequency deviation.

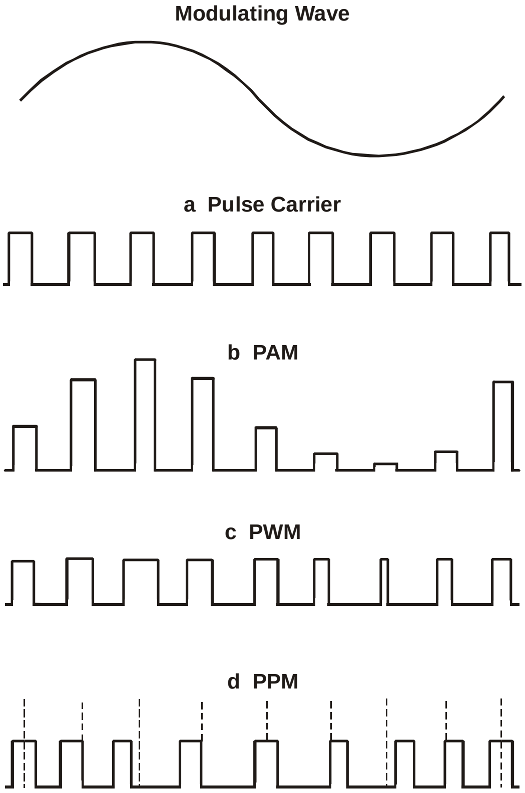
45. When two frequency modulated transmitters located within range of a receiver are operated simultaneously on carrier frequencies which lie within the receiver pass band, the stronger signal tends to suppress the weaker almost entirely and its modulation is prevented from appearing in the receiver output. This 'capture effect' is similar to the noise suppression effect, with the weaker signal equating to the noise.

PULSE MODULATION (PM)

Pulse Analogue Modulation

46. In pulse analogue modulation, a train of pulses is transmitted instead of a continuous waveform, and the modulating signal is used to alter one parameter of the pulses proportional to the amplitude of the signal at the time of the pulse. The three parameters which are used are the pulse amplitude, the pulse width, and the pulse position; the effect of modulating a uniform train of pulses by a simple sine wave signal is shown in Fig 14. Pulse analogue modulation techniques are very efficient and are frequently used for the transmission of data in telemetry and telecontrol. In all pulse modulation methods the pulse frequency must be at least twice the highest frequency contained in the message.

14-22 Fig 14 Methods of Pulse Modulation



47. **Pulse Amplitude Modulation (PAM)**. In pulse amplitude modulation the pulse amplitudes in a uniformly spaced train of pulses are changed by an amount proportional to the modulating signal amplitude at the time of the pulse (Fig 14b). During transmission the pulses may be kept short and pulses of other messages may be interleaved, thus increasing the utilisation of the channel.

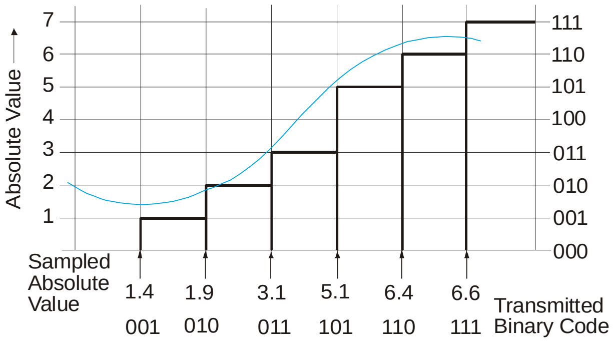
48. **Pulse Width Modulation (PWM)**. Pulse width modulation uses a uniformly spaced train of pulses with the pulse widths changed by an amount proportional to the modulating signal amplitude at the time of the pulse (Fig 14c); the leading edge, trailing edge, or both may be modulated. PWM has a greater resistance to noise than PAM since the pulse amplitude does not carry any information and any modification by noise is therefore immaterial. The noise resistance is improved if the pulse edges are steep which implies a greater bandwidth.

49. **Pulse Position Modulation (PPM)**. A pulse position modulation signal is produced by displacing the pulses in a uniform train by an amount proportional to the modulating signal amplitude (Fig 14d). PPM has similar noise resistance characteristics to PWM.

Pulse Code Modulation (PCM)

50. In the pulse code modulation technique the absolute value of the analogue signal is sampled and this value is transmitted as a binary code (Fig 15). As before, the sampling frequency must be at least twice that of the highest frequency contained in the signal. An analogue signal sampled in this manner is said to be 'quantized'. The accuracy with which the sample is coded is determined by the size of the levels into which the vertical axis is divided and this depends on the number of bits in the code. For 3.3 kHz speech the sampling rate is commonly 8 kHz and an 8-bit code is used. The rounding error is random and may have any size from zero to half the voltage represented by the least significant bit. At the receiver, the reconstructed message samples are equivalent to the correct samples plus an error which appears as noise called quantization noise.

14-22 Fig 15 Pulse Code Modulation



51. In a PCM system, the information is contained in the sequence of ones and zeros; the shape of the pulses is not important other than in determining the bandwidth. As a result, analogue noise can be rejected in the receiver provided that the individual pulses can be accurately recovered. A threshold detector is used and this outputs a clean digital waveform - errors will only occur if the size of a digit has been corrupted by noise to such an extent that it falls on the wrong side of the threshold. The size of the error generated will depend on the significance of the corrupted bit. It is this resistance to noise which makes PCM particularly attractive.

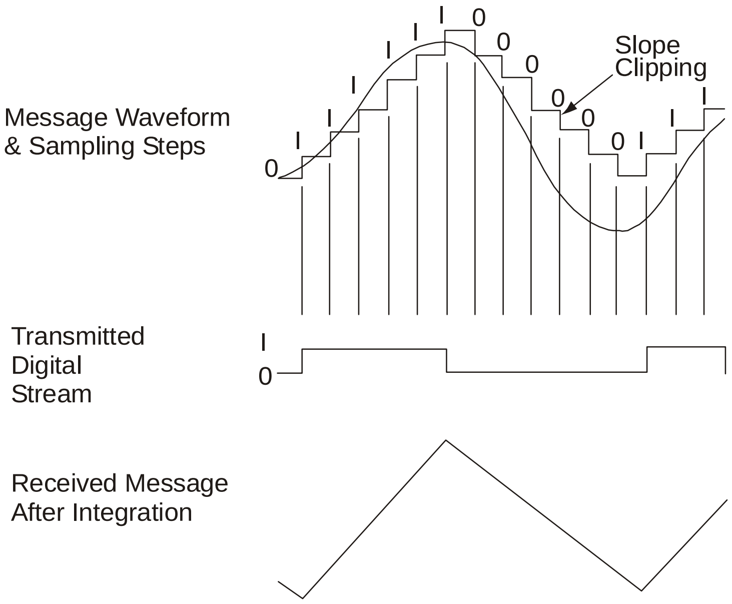
52. **Non-uniform Quantization**. The quantization noise in PCM could be reduced by reducing the spacing between levels. However, if the levels were uniformly spaced, this would require an increase in the number of levels and thus in the number of bits per sample. The signal to noise ratio at any time depends not only on the quantization noise, ie on the spacing between levels, but also upon the size of the signal. By reducing the spacing for small signals and increasing it for large values, the signal to noise ratio may be improved without any increase in the total number of levels. Compared to uniform quantization, non-uniform quantization can be equivalent to a saving of 3 or 4 bits per sample for speech signals.

53. **Differential Encoding**. In the differential encoding technique, instead of the full value of the sample, only the change from the previous value is transmitted. In general, the number of bits necessary to encode the change is less than that needed for the full value. In practice, this potential gain may be offset by selecting a higher sampling rate.

54. **Delta Modulation (ΔM)**. Delta modulation is a very simple variant of differential encoding in which only one bit per sample is produced and therefore the only information transmitted is whether the current sample is greater or smaller than the previous one. At the receiver the change in sample size is determined by circuit values. Fig 16 illustrates the principle showing the sampling steps, the

transmitted binary stream, and the output signal after integration at the receiver. The balance between positive and negative output pulses is a measure of the mean slope of the message. When the message waveform is not varying, or only fluctuating with a small amplitude, the output is a sequence of alternate positive and negative pulses. In order to track small fluctuations and reduce quantization noise a small step size is required; however this prevents steep slopes from being tracked (slope clipping - see Fig 16) unless the sampling rate and transmission bandwidth are increased. A variation of the technique, adaptive Δ-modulation, adjusts the step size to match the slope of the message thereby giving better tracking without increasing the bandwidth.

14-22 Fig 16 Delta Modulation



MULTIPLEXING

Introduction

55. Multiplexing is a technique whereby a number of message channels are combined and treated as a single entity for transmission, and then separated in the receiver. Combining messages allows parts of the system to be common to a number of channels thereby increasing the message handling capacity. There are three methods of multiplexing: frequency division, time division, and code division.

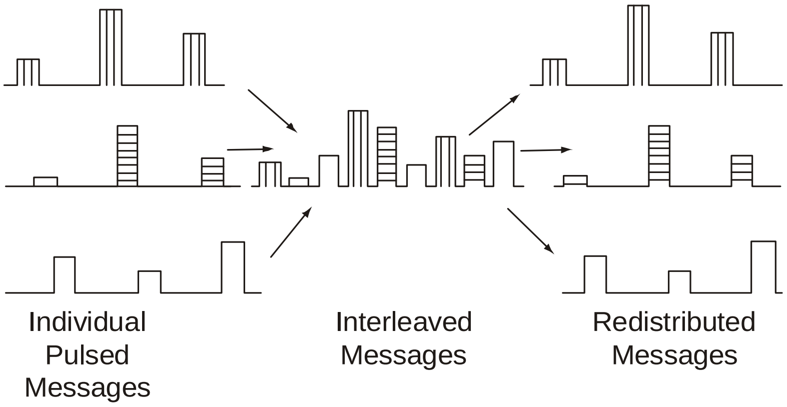
Frequency Division Multiplexing (FDM)

56. In a FDM system each message is frequency shifted and converted to a SSB signal occupying adjacent frequency slots. The combined messages are then treated as a single entity and frequency modulated onto a carrier for transmission. At the receiver the carrier is demodulated and the frequency block is divided up using filters. Each message is recovered to baseband by using an appropriate reference frequency.

Time Division Multiplexing (TDM)

57. A TDM system can only be used with time discrete signals, i.e. digital data, telegraphy, or sampled analogue. Although pulse analogue signals can be used, most systems use constant height, constant width pulses, and therefore PCM or ΔM are preferred. The pulses from a number of messages are interleaved in time, transmitted to a receiver, and then redistributed to their respective channels (Fig 17).

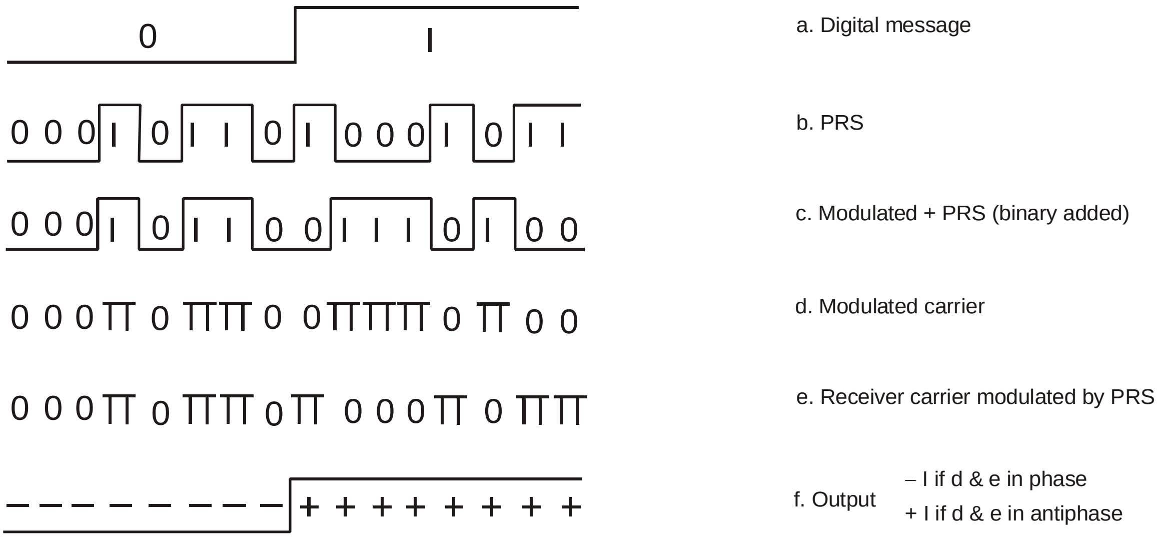
14-22 Fig 17 Time Division Multiplexing



Code Division Multiplexing (CDM)

58. In a CDM system, also known as spread spectrum, digital messages are transmitted in the same frequency band at the same time, separation being achieved by superimposing each message onto a different pseudo-random sequence (PRS). Fig 18 illustrates the principle.

14-22 Fig 18 Code Division Multiplexing


0
I
a. Digital message
0 0 0 I 0 I I 0 I 0 0 0 I 0 I I
b. PRS
0 0 0 I 0 I I 0 0 I I I 0 I 0 0
c. Modulated + PRS (binary added)
d. Modulated carrier
e. Receiver carrier modulated by PRS
f. Output
– I if d & e in phase
+ I if d & e in antiphase

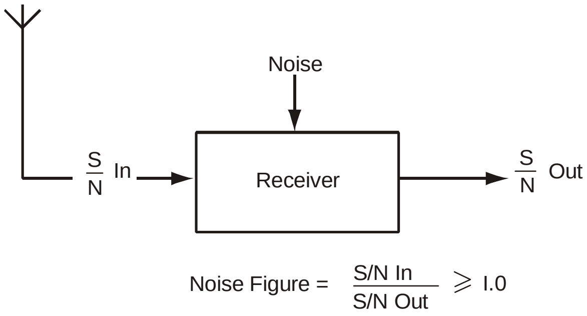
59. Fig 18 line 'a' shows a simple digital message and line 'b' a faster PRS. The message and PRS are binary added to produce a third digital stream (line 'c') which is used to phase modulate a sinusoidal carrier using phase shift keying (see para 18). The phase of the modulated carrier wave, relative to an unmodulated wave, is shown in line 'd'. In the receiver, a similar carrier is phase shift keyed with the same PRS used in the transmitter, and synchronized with the incoming waveform (line 'e'). The two waves are compared and if they are in phase the output is –1 and if in antiphase the output is +1. The output, shown in line 'f', represents the recovered message. It is clearly extremely important that the code in the receiver is synchronized with that received from the transmitter, if not the message output will be another random sequence.

INFLUENCE OF NOISE ON COMMUNICATIONS

General

60. In any communication system there will be unwanted electrical energy present in addition to that of the wanted information signal. This unwanted electrical energy is generally called noise and arises from a number of sources. Some noise occurs naturally and stems from sources such as thunderstorms, ionospheric storms (sunspot activity), and cosmic or galactic activity. Other components are man made and may originate inside or outside of the system under examination. There is a random movement of electrons in all circuit components above absolute zero, and this movement increases with increasing temperature. There is also varying electron velocity in those devices in which electrons are forced to move, eg in transistors and thermionic devices - external man-made noise emanates, for example, from electrical machinery and other communications systems.

14-22 Fig 19 Receiver Noise Figure



61. The ratio of the signal to the noise at any point in the system is usually expressed in decibels and is given by:

$$\text{S/N ratio} = 10 \log \frac{\text{Signal Power}}{\text{Noise Power}}$$

$$\text{or } 20 \log \frac{\text{Signal Voltage}}{\text{Noise Voltage}}$$

The receiver will itself add noise to the signal such that the S/N ratio at the output is always smaller than the S/N ratio at the input. The ratio of these ratios is called the 'noise figure' and is a measure of the quality of a receiver in that it describes the amount by which it degrades the input S/N ratio (Fig 19). For certain applications, devices such as the parametric amplifier and travelling wave tube offer very low noise figures of 5 or less, while the maser can improve on this with a figure approaching unity.

62. The amount of noise power present in a receiver imposes a lower limit to the signal power required for satisfactory operation. Whereas signal power varies directly as the transmitted power at a given range, and inversely as the square of the distance from a transmitter for a given power, the noise power has a constant value determined by the receiver bandwidth and noise figure. With increasing distance the S/N ratio falls until a limiting value is reached below which the quality of the signal becomes unacceptably degraded or the information becomes lost due to errors greater than can be tolerated. The lower the noise figure of the receiver, the greater is the communication range possible for a given transmitter power, or conversely, the lower is the power required to communicate over a given distance.

63. The characteristic value of S/N ratio representing the threshold for satisfactory signal detection varies considerably with different systems. For high quality television a S/N ratio in the order of 10,000 (40 dB) is required, while satisfactory voice communication can be carried out with a S/N ratio as low as 10. Digital signals used in telegraphy and data transmission are highly resistant to noise and can be detected with a S/N ratio in the vicinity of unity. Other than in PAM, the detector has only to detect the presence or otherwise of a pulse; its amplitude carries no information and so the fact that it is contaminated with noise is of no matter. Data corruption is only likely if the noise level is so high that a signal level which should be below the detector threshold is forced above it. PAM signals can often be cleaned of noise to some extent by suitable filters.

CHAPTER 23 - INTRODUCTION TO ELECTRONIC WARFARE (EW)

Introduction

1. Electronic Warfare (EW) encompasses any military action that involves the use, or control of, the Electromagnetic Spectrum (EM) to reduce or prevent hostile use by, or to attack, an enemy.

2. The time around World War Two saw rapid developments in the military applications of radio and radar and it is often considered that the evolution of modern EW stems from this time. Certainly, since 1945, there has been a proliferation of weapons, such as radar guided missiles and radar laid guns, which rely on electronic systems for their operation, and there has been a corresponding expansion of EW techniques in an attempt to reduce the effectiveness of these weapons in the hands of the enemy, while at the same time ensuring friendly use. However, it should not be forgotten that EW embraces the whole of the electromagnetic spectrum, and techniques involving other than radio and radar frequencies, such as the use of camouflage, deception, and smoke screens in the visible band, are much older, and indeed are also extensively encountered in the natural world. Recent years have seen a considerable emphasis in the field of electro-optical weapon systems using both the visible and infra-red parts of the spectrum, and both laser and non-coherent energy. Electro-optical systems can be used by day and by night for the safe and accurate navigation of an aircraft, for target detection, and for the aiming and guidance of weapons, and they are therefore worthwhile targets for the employment of EW techniques. Developments now include Directed Energy (DE) weapons whereby the energy inherent within electromagnetic radiation is used as a damage mechanism; such weapons can be considered as EW constituents, particularly where they are targeted against electronic or electro-optical systems or components.

Electronic Warfare[1]

3. *Electronic Warfare*. *NATO definition: Military action that exploits electromagnetic energy to provide situational awareness and achieve offensive and defensive effects*. EW encompasses any military action that involves the use or control of the EM spectrum to reduce or prevent hostile use or to attack the enemy. It is subdivided into three main mission types or actions:

a. *Electronic Attack (EA)*. *NATO definition: EA is the use of EM energy for offensive purposes*. EA is employed to disrupt, deceive, destroy or deny an adversary's Electro Magnetic Operations (EMO), attack their C2 capabilities and diminish their opportunities to shape or exploit the operational environment. EA has an increasingly important role in joint air/land operations and in enabling destruction of enemy forces by combined EM/kinetic attack. EA includes Directed Energy Weapons (DEW, e.g. EM Pulse and high-power microwaves), when used offensively.

b. *Electronic Defence (ED)*. *NATO definition: ED is the use of EM energy to provide protection and ensure effective friendly use of the EM spectrum*. ED is primarily used to protect individuals and forces, platforms, systems and areas, either alone or in concert with other physical capabilities.

c. *Electronic Surveillance (ES)*. *NATO definition: ES is the use of EM energy to provide to situational awareness and intelligence*. ES is focused on providing immediate shared situational awareness and indicators and warning of operational activity.

[1] References used are the Air Electronic Warfare Course reference book Vn 0.20 dated 1 Nov 2017 and the MoD Joint Doctrine Note 1/18 – Cyber and Electromagnetic Activities.

4. EW actions are further subdivided into 3 distinct EW Measures which may be pertinent to all or some of the definitions above. These measures are as follows:

a. *Electronic Counter Measures (ECM)*. ECM encompasses actions taken to prevent or reduce an enemy's effective use of the EM spectrum through the use of EM energy. There are 3 subdivisions of ECM:

i. *Jamming*. Electronic jamming is the deliberate radiation, re-radiation or reflection of EM energy with the object of impairing the effectiveness of electronic devices, equipment and systems.

ii. *Neutralization*. Electronic neutralization is the deliberate use of EM energy to either temporarily or permanently damage devices that rely exclusively on the EM spectrum.

iii. *Deception*. Electronic deception is the deliberate radiation, re-radiation, alteration, absorption or reflection of EM energy in a manner intended to confuse, distract or seduce an enemy or his electronic systems.

b. *Electronic Protective Measures (EPM)*. EPM comprise those actions to ensure friendly effective use of the EM spectrum despite an adversary's use of the same spectrum. EPM involves the use of active and passive measures to protect personnel, facilities and equipment from the effects of enemy (and occasionally friendly) EW systems that may degrade, neutralise or destroy friendly combat capabilities. EPM prevents the enemy from gaining intelligence and information from friendly transmissions and safeguards command, control and comms systems by imposing both procedural and technical solutions. It is also used to counter the ECM applied by adversary systems by procedures and specialised circuitry. Mutual interference and unintentional jamming can be avoided through the close coordination of intelligence, comms and operations that takes place in the EW Coordination Cell. Spectrum management tools and procedures, that include both frequency management and EW frequency deconfliction, are used for this purpose. There are 2 main types of EPM available to forces namely:

i. *Passive EPM*. Undetectable measures which are meant to ensure friendly effective use of the EM spectrum such as:

(a). Use of reduced power, brevity of transmissions and directional antennas.

(b). Careful positioning of C2 resources to reduce the risk of detection and accuracy of direction finding.

(c). Use of SOPs and Emission Control (EMCON) procedures.

(d). Regular briefings on the latest EW threats.

(e). Regular training for EW staff in the recognition of enemy EW activity and the application of suitable EPM.

ii. *Active EPM*. Measures detectable by an adversary to ensure friendly use of the EM spectrum such as:

(a). Changing frequencies or altering transmitter parameters (wartime modes).

(b). Use of spread spectrum, burst transmissions, frequency hopping, agility or diversity, change of modulations, jittered or staggered pulse repetition frequencies (PRF) and changing or modulating power outputs.

c. *Electronic Support Measures (ESM)*. ESM encompasses those actions taken to search for, intercept and identify EM emissions and to locate their sources for the purpose of immediate threat recognition. It provides a source of information required for immediate decisions involving ECM, EPM and other tactical actions. ESM includes surveillance of the EM spectrum to achieve the following:

i. Immediate threat recognition in support of operations and other tactical actions such as threat avoidance, homing and targeting.

ii. Intelligence collection, deception planning, detection of threat changes, target location, situational awareness and avoidance of fratricide.

Airborne Electronic Warfare

5. Airborne EW is used to enhance the survivability of aircraft and ground assets, and to improve mission effectiveness. ES gives warning that radars are active in the area of operations. Aircraft Defensive Aids Suites (DAS) are designed, and programmed, to identify threats, give warning to the crew, and where appropriate, generate counter-measures. Active jamming or decoys, used in association with tactical manoeuvring by aircraft, may help to disengage from a threat.

6. Avoidance should be the primary method of limiting engagement opportunities when the positions of fixed SAM and AAA sites have been determined from intelligence sources, and aircraft routing can be chosen to minimize exposure. Unavoidable penetration of threat areas may require flight profiles that use terrain masking as a primary means of avoiding, or minimizing, the acquisition by radar-laid SAMs or AAA. Signature reduction technology may be used to reduce the range at which a threat system might be able to detect, track or target an aircraft.

Airborne Support Assets

7. Airborne support assets can be vital contributors to the EW battle and are always considered as critical assets. They take the form of either passive sensors or active jammers.

a. *Passive Sensors*. Passive sensors normally provide SIGINT information but can be used in the Electronic Intelligence (ELINT) role to provide updated and accurate details of EW threats.

b. *Active Jammers*. Active jammers work against radars, communications and navigation aids. Their main targets are early-warning and long-range acquisition radars, C2, navigation and Identification Friend or Foe (IFF) systems.

Summary

8. A knowledge of EW is necessary for aircrew so that they can exploit the electromagnetic spectrum to the best advantage, and perhaps as importantly, can recognize when the enemy is using EW techniques against them so that they can take any necessary actions to negate or reduce the threat. EW is an area of continuous development; any system development by one side will stimulate the development of a counter by the other, which will in turn lead to further developments to circumvent the counter, and so on. There is, therefore, no intention to describe any specific equipment in the

following chapters; rather the nature of ES, EA and EP will be explored, and the underlying general principles of various techniques will be described.

CHAPTER 24 - ELECTRONIC SURVEILLANCE (ES)

Introduction

1. Electronic Surveillance (ES) ES is the use of EM energy to provide to situational awareness and intelligence. ES is focused on providing immediate shared situational awareness and indicators and warning of operational activity and encompass that division of EW involving actions taken to intercept, identify and locate sources of intentional and unintentional radiated EM energy for threat recognition. ES therefore provide a source of information required for immediate decisions involving Electronic Counter Measures (ECM), Electronic Protective Measures (EPM), and other tactical actions such as threat avoidance, targeting and homing.

2. ES should not be confused with the much wider activity of intelligence gathering, which provides information for other than just EW purposes, and is principally a long-term, strategic, activity. Where intelligence gathering is directed towards data on electromagnetic radiations it is known as signals intelligence (SIGINT), which may be divided into communications intelligence (COMINT) and electronic intelligence (ELINT). ELINT in particular has an input into EW in that it provides most of the background knowledge necessary for the effective design and operation of EW systems.

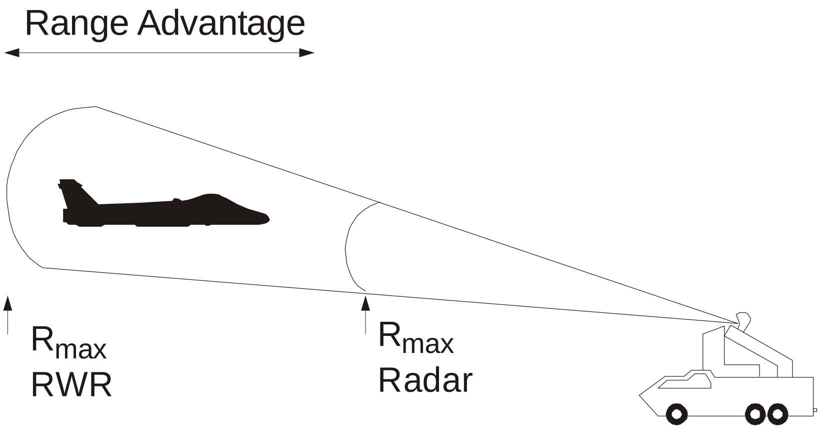
3. Enemy electronic transmissions can be detected and analysed in considerable detail by specialized equipment, but the immediate threat recognition requirement for a combat aircraft is met by the radar warning receiver (RWR).

Radar Warning Receivers

4. The radar warning receiver (RWR) is designed to detect, localize, and identify threat radars in order that appropriate countermeasures can be taken. A RWR may simply provide warning which will require an aircrew to assess the potential threats and decide on a response, or it may be part of an integrated computer controlled EW system which can automatically assign priorities and initiate appropriate ECM.

5. **Range Advantage***.* A RWR should normally be able to intercept a threat radar transmission before the aircraft itself is detected. The margin of this benefit is known as the range advantage (see Fig 1).

14-23 Fig 1 RWR Range Advantage



It occurs because the radar echo to be detected by the radar receiver is subject to a two-way propagation path and the power received back at the radar therefore varies inversely with the fourth power of range, while the radar power received by the RWR has only a one way propagation path and varies inversely with the square of the range. The degree of range advantage depends on a number of factors including:

a. Radar receiver sensitivity.

b. RWR sensitivity.

c. Radar cross-section (RCS) or reflectivity of the target.

d. Local propagation effects.

e. Radar operator efficiency.

The degree of advantage should be a minimum of 10% and can be in excess of 50% depending on the systems under consideration.

6. **Direction Finding**. A RWR can determine the direction of a threat by using either amplitude or phase comparison techniques. In the amplitude comparison method four aerials are aligned with the intercardinal directions relative to the aircraft and the amplitude of the received signal in each aerial is compared. Accuracy is generally about 10° to 15°. The phase comparison method uses a set of aerials in a phased array and the phase difference between the individual aerials for the same incoming radar wave can yield a direction accurate to about 1° or better in many cases.

ES RECEIVERS

The Ideal ES Receiver

7. The ideal ES receiver should:

a. Be sensitive in order to achieve long range intercepts on weak radar signals.

b. Have a 100% intercept probability, ie able to receive, at all times, signals from all directions and all frequencies within its operating capability without having to steer antennae or tune receivers.

c. Cover the full range of frequencies used by threats (say, in the order of 0.1 GHz to 40 GHz), but this is likely to increase.

d. Cover 360° in azimuth and significant elevation angles, say 60°.

e. Provide accurate and immediate direction finding.

f. Measure RF, PRF, PW, pulse modulation, and antenna scan .

g. Cope with pulsed, interrupted continuous wave (ICW), and CW signals.

h. Deal with jittered, staggered, and agile radars.

i. Sort and de-interleave signals when the system receives a large number of different radar emissions simultaneously.

j. Carry out a comparison with the parameter library, and allocate a priority to high threats.

k. Display the results clearly.

8. In practice there will be conflicts between some of these requirements; for example, high sensitivity demands narrow bandwidth receivers with high gain, narrow beamwidth antennae, whereas high intercept probability demands receivers of wide instantaneous bandwidth and antennae covering 360° of azimuth. In addition, the degree of sophistication may be limited by considerations of cost.

9. Simple ES receivers are wide open, i.e. they look at the full beamwidth and bandwidth of the system at all times. This necessitates some sacrifice of sensitivity in order to achieve 100% intercept probability. In more complex systems channelization or scanning techniques are employed. In a channelized system, each channel is a high sensitivity sub-system specializing in a narrow range of directions or frequencies or both. The required overall coverage is achieved by having many such channels working in parallel. A further advantage is that, if the channels are independent, the system can deal with the simultaneous arrival of different signals, carrying out the measurements on each one independently. The scanning technique employs a narrowband system which sweeps its coverage rapidly through the frequency band. This method gives a low intercept probability against an isolated pulse, but provided that the radar keeps transmitting for a reasonable time an overall intercept probability close to 100% can be achieved.

Crystal Video Receiver

10. Early types of RWR use Crystal Video Receivers (CVR) which are 'wide open', relatively cheap, lightweight and compact, but have rather poor frequency discrimination and low sensitivity. The CVR cannot intercept CW signals directly and additional circuits are necessary to impress a modulation on the incoming RF; detection of the modulation (typically a 10 kHz square wave) implies the presence of a CW signal.

11. The simple CVR’s limited frequency discrimination makes it unsuitable for use where the received signals are compared with a stored library of emitter parameters since there will inevitably be ambiguities.

Narrow Band Receivers

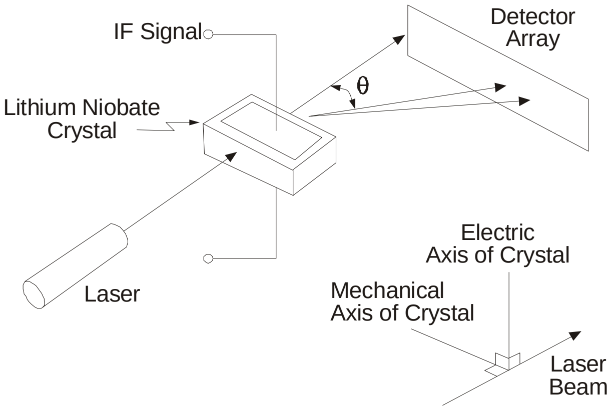
12. In order to reduce the number of ambiguities and unknowns it is necessary to distinguish more accurately between radars which use the same part of the electromagnetic spectrum. To achieve this a receiver with a finer frequency discrimination than the simple CVR is required. Although a CVR could be fitted with a large number of narrowband RF filters, this channelization method tends to be expensive and is normally only used for rather crude filtering, ie allocating the received RF to one of a number of broad bands without accurately measuring the frequency. A more practical alternative is to use a narrowband filter with a pass band which can be retuned rapidly. The filter employed uses a material, Yttrium Iron Garnet (YIG), which has the property that the resonant frequency depends on the strength of any magnetic field in which the crystal is placed. A swept voltage applied to a coil surrounding the crystal causes the crystal resonant frequency to sweep.

13. Instead of a CVR, a superheterodyne (or superhet) receiver is frequently used with either scanning or channelization techniques. Scanning is achieved by sweeping the local oscillator frequency rapidly which causes the intermediate frequency (IF) to sweep at the same rate. A channelized superhet, having a large number of narrowband fixed-tuned superhets operating in parallel and simultaneously, is a very capable arrangement - but is very expensive. For example, in order to cover frequencies from 1 GHz to 18 GHz in 10 MHz channels, requires 1,700 superhets in parallel.

14. The superhet is unable to provide an audio analogue output, so computer generated tones are used. An alternative may be to use a CVR to provide true audio output. Superhet receivers normally only provide a limited range of warning tones to alert the crew to, for example, a radar lock-on or a missile guidance frequency detection.

15*.* **The Bragg Cell Receiver***.* The arrangement of a Bragg Cell receiver is shown in Fig 2. The incoming RF is mixed down to an IF of about 2 MHz which is then applied to a piezo-electric crystal (usually lithium niobate). The piezo-electric action establishes longitudinal waves in the crystal such that there are alternating regions of compression and rarefaction. This pattern acts as a diffraction grating to the laser light causing the laser energy to be deflected. The degree of deflection depends on the separation of the regions in the crystal which in turn varies with the IF. The diffracted laser energy impinges onto an array of detectors, each detector corresponding to a particular IF and therefore to a particular original RF. Frequency resolution depends on the spacing of the detectors, and the system can cope with the simultaneous arrival of signals with different RFs - the laser energy dividing into a number of sub-beams each of which is deflected through the appropriate angle and detected independently.

14-23 Fig 2 Bragg Cell Rx



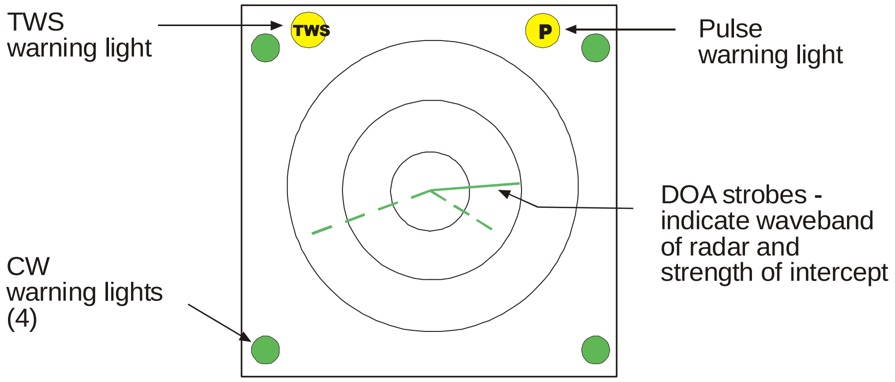
RWR DISPLAYS

General

16. It is important that the information displayed by a RWR should be easily understood and show an appropriate amount of data, without being distracting or confusing. The minimum information given to the crew should be a visual and aural warning, and an indication of the type and direction of the threat.

17. Older RWRs use analogue displays comprised of coloured lights and a small cathode ray tube to display the direction of arrival (DOA) of radar signals in the frequency bands from E to J. A typical display is shown in Fig 3 in which the DOA of CW radar beams is displayed, rather coarsely, by four quadrant lights, while pulse radars are shown as strobes on the CRT, the strobes being coded (solid or broken) to indicate the radar band. The length of the strobe gives a crude indication of the strength of the received signal. Additional warning lights show that a pulse radar, or a mechanical track-while-scan (TWS) system, has been intercepted.

14-23 Fig 3 Typical Analogue RWR Display

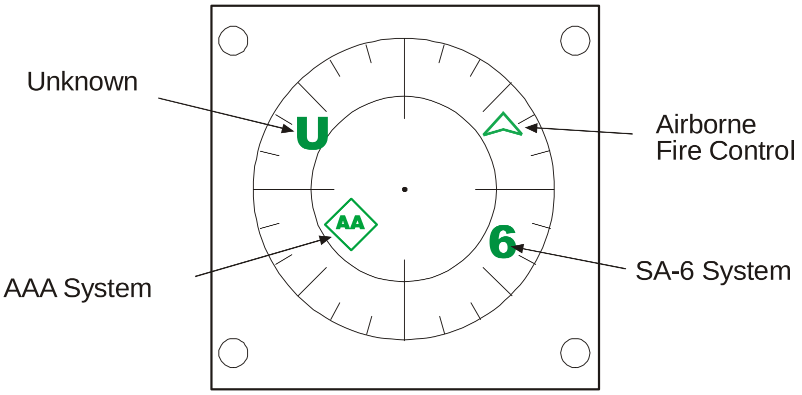

TWS
warning light
TWS
P
Pulse
warning light
DOA strobes - indicate waveband of radar and strength of intercept
CW
warning lights
(4)

18. In addition to the visual display, an audio output is available on the aircraft's intercom system which produces an audio analogue of the PRF of the received radars. The audio is arranged so that long range radars with a low PRF generate a low pitched tone indicative of no immediate threat, while a threatening, short range, high PRF radar generates a more urgent, high pitched tone. Furthermore, early warning and acquisition radars will be distinguished by some seconds of silence between bursts of energy reflecting the radar's relatively long aerial rotation period. Conversely, threatening fire control and tracking radars will be characterized by a more or less continuous tone. Alarm tones for CW radars are synthesized by additional internal circuits.

19. It is possible that, by comparing the radar PRF, band, and modulation or scan type (ie pulse, CW, TWS) to determine the type and possibly the identity of the illuminating radar. However, this analysis requires a considerable degree of skill and imposes a considerable workload during a high stress part of a mission, and can become very difficult in a multi-threat environment. Furthermore, many threat radars now use PRFs well above the audio limits of the intercom system.

20. In order to reduce the degree of skill and time required to interpret an analogue display, an RWR can be fitted with a small digital processor and memory so that the warning lights or CRT display can show the name or function of the identified radar together with its relative position. A typical display is shown schematically in Fig 4; it can be designed to show closer threats either nearer to the centre or nearer to the edge. The audio output from this type of system is identical to that from the analogue display RWR.

14-23 Fig 4 Typical Digital RWR Displays



21. In order to identify a radar, the values of the various radar signal parameters, eg frequency, pulse width, PRF and amplitude, are digitized and the result compared with a set of stored identities in the RWR memory. If and when a match is found the appropriate symbol is displayed. Because the size of

the RWR memory is limited by considerations of, for example, cost and processing time, there will be some instances when the RWR is unable to find any match in its memory and in this case an 'unknown' symbol will be displayed. In order to minimize the number of ambiguities and unknowns the RWR memory should be programmed before flight with a library of emitters reflecting those most liable to be encountered. In addition, the library will have a priority allocation so that if a conflict in processing and display between two intercepted signals arises, the higher threat is portrayed.

CHAPTER 25 - ELECTRONIC ATTACK (EA)

Introduction

1. Electronic Countermeasures (ECM) is part of the division of Electronic Warfare (EW) now called Electronic Attack (EA). EA uses the Electromagnetic Spectrum (EM) or Directed Energy (DE) to attack personnel, facilities or equipment. The UK military forces generally use the EM for defensive purposes, using ECM. ECM involves actions taken to prevent or reduce an enemy's use of the EM, through the use of electromagnetic energy. It includes all the actions taken to deny the use of radar, communications, navigation, and electro-optical systems by jamming, deception, neutralization, decoys, stealth, and directed energy weapons - although this last technique is still largely at the development stage and will not be considered in detail in this chapter. The attack element of EA is a constantly evolving technology and is not considered in detail in this chapter.

RADAR ECM - USE OF ELECTRONICS

General

2. ECM against radars will frequently take the form of on-board electronic devices. It is important that an indication is obtained that an enemy radar is operating, and is a threat, before countermeasures are applied, since their unnecessary early use could alert the enemy to the presence of an otherwise undetected target. The detection of a threat radar will normally be accomplished by the use of ESM equipment, typically a radar warning receiver (RWR), and determination of the threat radar characteristics, e.g. frequency, scan type, CW or pulse, can assist in making the choice of the best ECM technique to be employed.

3. In order to reduce the effectiveness of a radar, the ECM waveforms used must be of a type which will be accepted by the receiver. Thus the ECM signal should match the radar signal in frequency and ideally in polarization, and should have a bandwidth which is at least as wide.

4. The use of electronics as an ECM technique against radars will be considered under the headings of:

a. Electronic Jamming (Noise Jamming).

b. Electronic Deception.

In some modern systems the two techniques are combined in some way so that, for example, deception signals could have noise modulation superimposed, or noise jamming could contain deception signals designed so that the enemy may filter them out, discarding them as real targets.

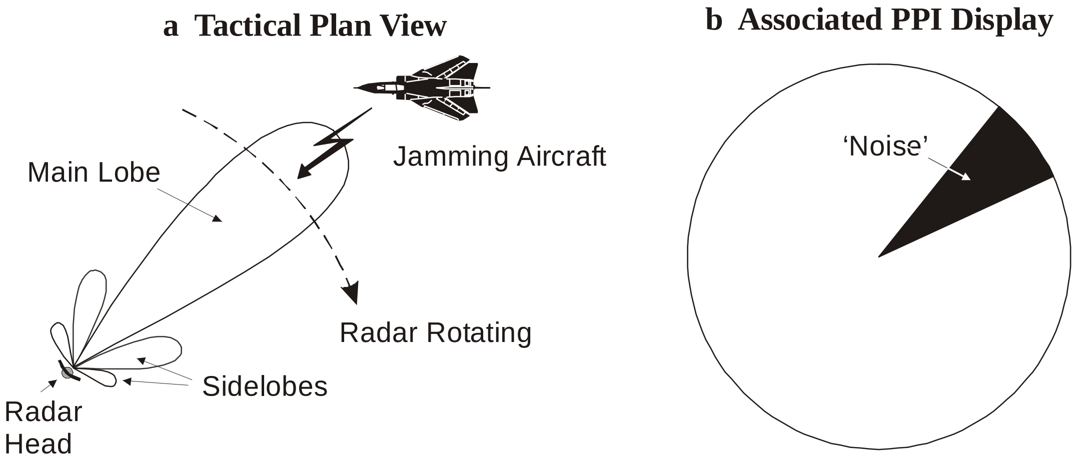
Electronic Jamming (Noise Jamming)

5. Electronic jamming is the deliberate radiation, re-radiation, or reflection of electromagnetic energy, with the object of impairing the effectiveness of electronic devices, equipment or systems being used by the enemy.

6. Radar receivers need to be highly sensitive in order to detect the very weak amplitude of returned radar echoes and this sensitivity makes them vulnerable to jammers which can generate

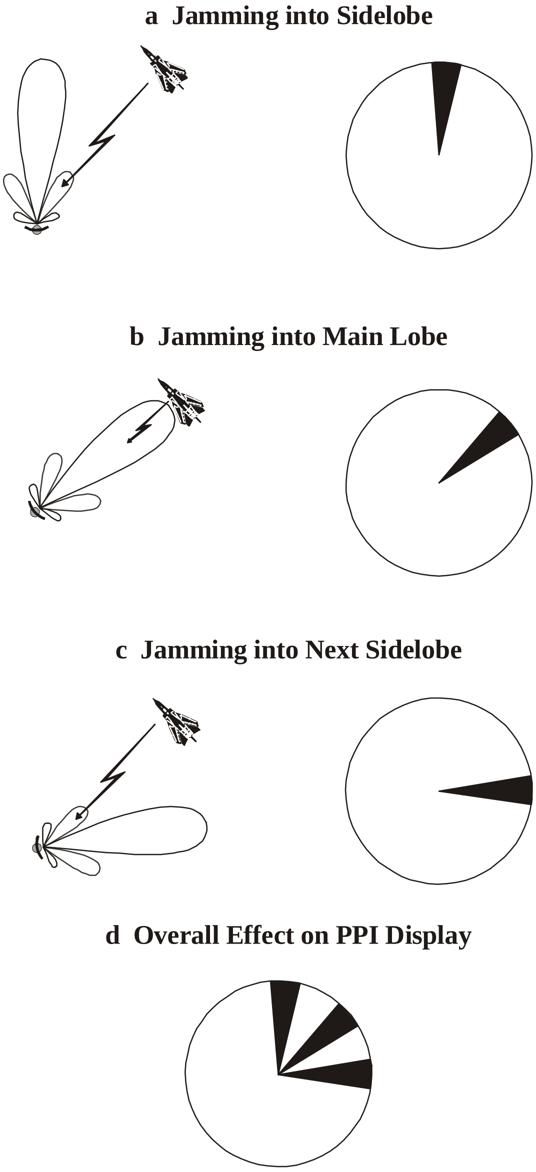
sufficient noise in the receiver to swamp the wanted signal. On a PPI radar display the effect of a noise jammer is to mask the range of the jammer although the azimuth is still apparent (Fig 1).

14-24 Fig 1 Main Lobe Jamming on PPI Display



7. However, if the jammer is sufficiently powerful, energy can be injected into the radar sidelobes, and on a PPI display this noise will be painted on the azimuth of the main lobe. Fig 2 shows the effect as the radar aerial rotates. In Fig 2a the energy enters the sidelobe. As the aerial rotates to the position in Fig 2b the main lobe intercepts the target, and further rotation (Fig 2c) brings the next sidelobe into a receiving position. In each position, a paint is made along the current main lobe azimuth. The display video persistence causes a display similar to that shown in Fig 2d to result.

14-24 Fig 2 Combined Effect of Main Lobe and Sidelobe Jamming



a Jamming into Sidelobe

b Jamming into Main Lobe

c Jamming into Next Sidelobe

d Overall Effect on PPI Display

8. A radar transmission covers a finite band of frequencies, for example, a pulse radar using 0.5 µsec pulses would use a band of about 2 MHz centred on its carrier frequency, and for noise jamming to be successful the noise bandwidth must be at least as wide as that of the victim radar. A jammer has a limited power output and can be designed so that this power is either concentrated over a narrow bandwidth, or spread over a wide bandwidth. There are four major categories of noise jammers reflecting these design options:

a. Spot Jammers.

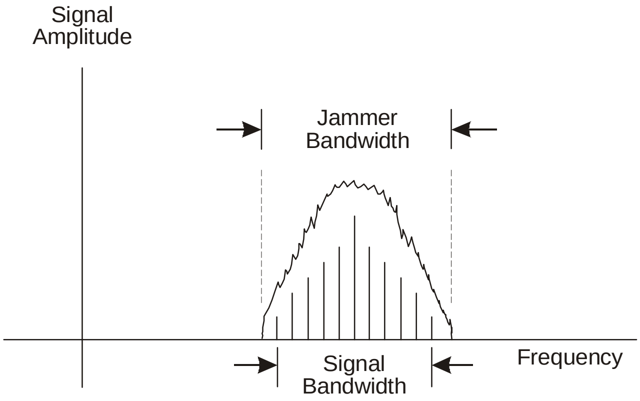
b. Sweep Jammers.

c. Barrage Jammers.

d. Search-lock Jammers.

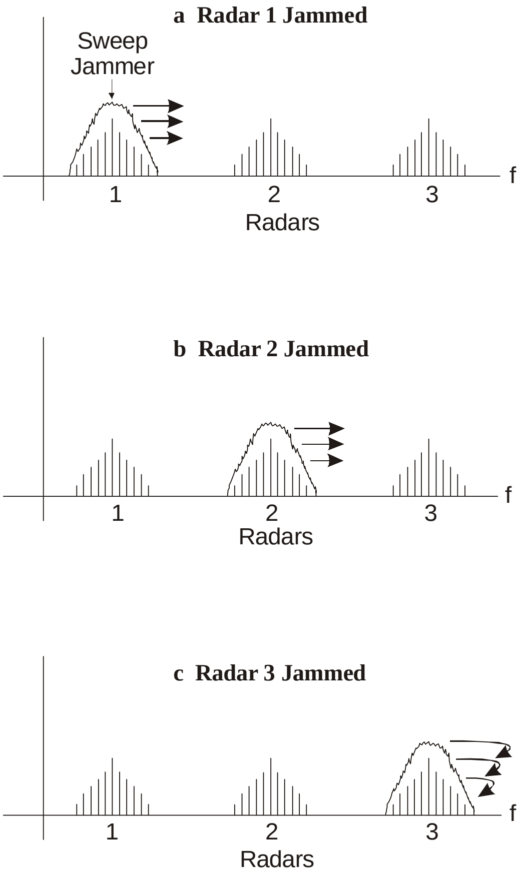
9. **Spot Jammers**. A spot jammer is one which is tuned to, and targeted against, one particular radar. This type of noise jamming provides the maximum jamming power, but the limitation of being able to jam only one radar at a time may be a disadvantage in a complex tactical scenario. In order to use a spot jammer the target radar frequency and bandwidth must be known from intelligence or must be determined using an intercept receiver. The jammer is then tuned manually or automatically to be centred on the target radar's carrier frequency and sufficient bandwidth is allowed to encompass that of the target (Fig 3).

14-24 Fig 3 Spot Jamming



10. **Sweep Jammers**. Sweep jammers are employed when there is a requirement to jam a large number of radars which are operating over a wide frequency band and it would be impractical to employ a similar number of individual spot jammers. The technique can also be employed against frequency agile radars. The jammer is tuned to a narrow bandwidth, sufficient to cover that of any of the target radars, and this narrow frequency band is swept over the wide band containing the target radars (Fig 4). This results in all the radars being affected, but not continuously. Thus for example in Fig 4a, radars 2 and 3 will be able to operate while radar 1 is being jammed. Fast sweeping can sometimes overcome this disadvantage by being able to depress automatic gain control circuits, or cause oscillation in the radar receiver for the time taken by the jammer to return to its frequency.

14-24 Fig 4 Sweep Jamming


a Radar 1 Jammed

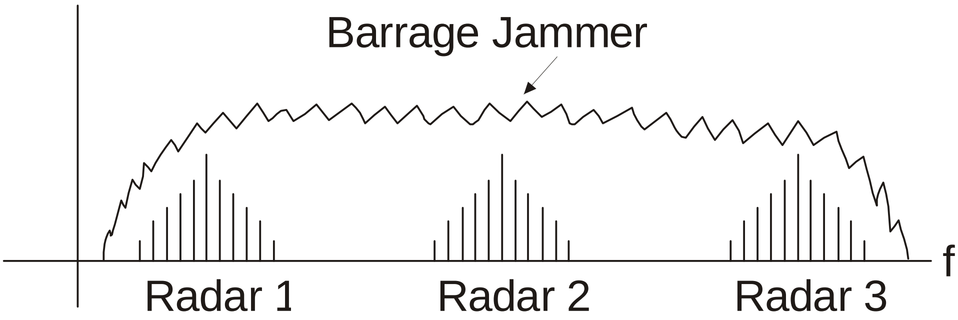
b Radar 2 Jammed

c Radar 3 Jammed

11. **Barrage Jammers**. Barrage jammers are used when it is necessary to jam over a wide bandwidth, enabling a number of radars, including those which are frequency agile, to be jammed simultaneously and continuously (Fig 5). However the jammer power must be spread over this wide band and thus the power at any frequency is reduced in comparison to that available from a spot

jammer with the same available power. In addition, a lot of energy may be wasted in jamming frequencies which are not being used by target radars. In order to overcome these disadvantages, barrage jammers need to be high power devices.

14-24 Fig 5 Barrage Jamming

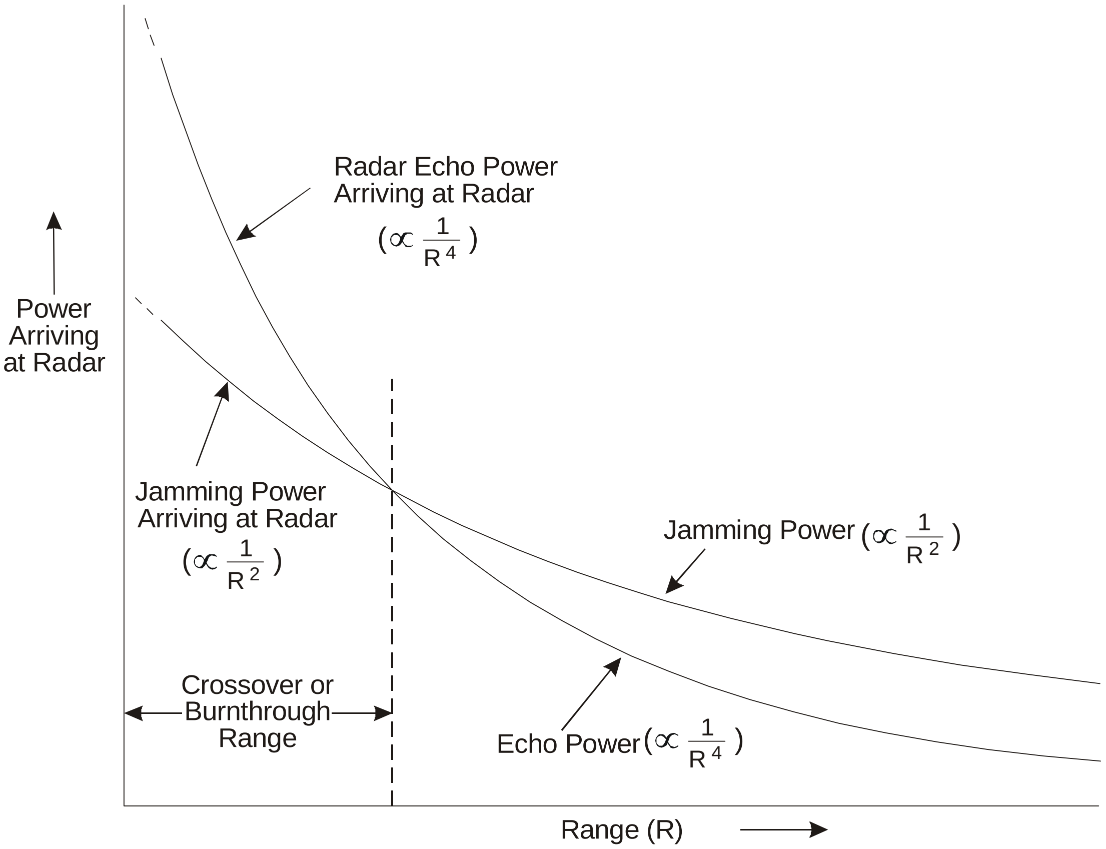


12. **Search-lock Jammer***.* The search-lock jammer is a special version of the spot jammer in which a receiver able to search a large bandwidth is tied in with a spot jammer. Thus the receiver searches for a target radar and initiates the jamming signal on the detected frequency. The system alternates between receiving and transmitting, and so if a victim radar stops transmitting or switches frequency, the receiver notes the lack of signal and the jammer stops transmitting until the receiver re-acquires a signal and retunes the transmitter. The period when the receiver is allowed to look at the victim's radar frequency is known as a 'look-through' period, and some equipments have sufficient isolation between the receiver and transmitter to allow continuous 'look-through'. The 'look-through' capability allows search-lock jammers to operate against frequency agile radars, although if one jammer is working against several frequency agile radars operating in the same frequency band, the effect of the jamming on any individual radar will be significantly reduced. Radars which are frequency agile on a pulse-to-pulse basis are virtually invulnerable to this type of search-lock jammer as it will be constantly searching for the frequency rather than transmitting.

Burnthrough

13. The ECM power arriving at the enemy's radar is always competing with the radar's received target echoes. Since the ECM transmission is only one way its power falls off as the square of the distance from its source. A radar echo, however, travels two ways, to the target and back after reflection and re-radiation. Its power thus falls off as the fourth power of range to the target. A graph showing these two cases of power fall-off against range is shown in Fig 6. It will be seen from the graph that there is a range below which the radar echoes have a greater power than the jamming signals and the jammer will show as a target through the jamming signal. This range is known variously as the Burnthrough Range, the Self-screening Range, or the Crossover Range and its value depends principally upon the radar and jammer power, and on the radar cross-section of the target.

14-24 Fig 6 Comparison of Echo and Jamming Power Arriving at Radar

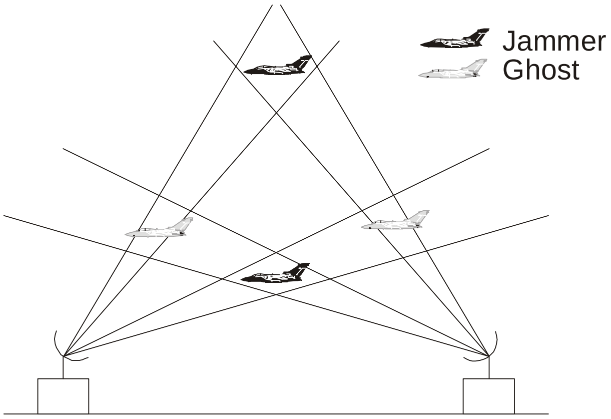


Implications of Noise Jamming

14. Noise jamming is a simple and effective method of screening or denying the range, velocity, and strength of an attack but its implementation must be given careful consideration. Noise jamming advertises the jammer's presence, and if the jamming power is insufficient to jam the radar sidelobes, the azimuth or elevation of the jammer will be available to the defender. If the radar under attack is a surveillance type, and more than one site is available, the jammer's position can be determined by triangulation.

15. When more than one jammer is used against a set of surveillance radars and triangulation is attempted, the defenders will have to deal with ghost targets as shown in Fig 7, the number of possible target positions increasing as the square of the number of jammers.

14-24 Fig 7 Jammer and Ghost Positions



16. In some situations, notably against radar laid guns where lack of range information seriously degrades the weapon system, noise jamming is a viable technique. However, against a missile system noise jamming could be counter-productive since the fire control radar could track the noise strobe and direct a weapon against the target using 3-point or command-to-line-of-sight guidance, making the assumption that the target is within range. In this situation deception jamming is usually the better option.

Electronic Deception

17. Electronic deception is the deliberate radiation, re-radiation, alteration, absorption, or reflection of electromagnetic energy in a manner intended to confuse, distract, or seduce an enemy or his electronic systems.

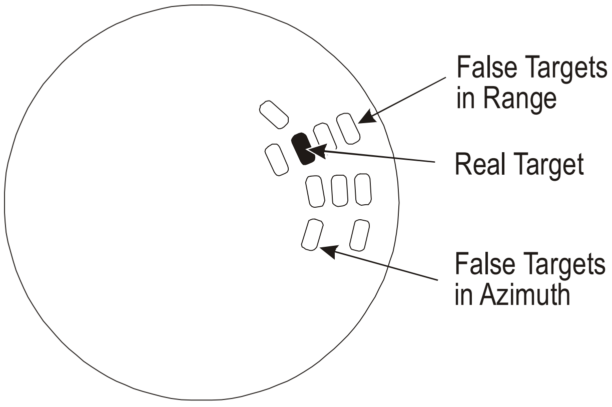
18. Deception ECM (DECM) equipments are generally used in self-protection jamming applications against weapon systems which use fire control radars, and are designed to degrade or deny the range, velocity, or angle tracking capabilities of such a tracking radar. Since, as with noise jamming, it may be possible to fire a weapon against a jamming target even if range or velocity information is denied, angle track breaking has the highest priority.

19. In order for DECM to be successful, the victim radar’s search pattern, radar parameters, and tracking method must be established. Techniques are then available to cause false targets to appear, to pull-off range and velocity tracking gates, or to displace angle tracking. In general, DECM requires less power than noise jamming and in many cases more than one threat can be jammed simultaneously using time sharing techniques. Deception jammers have an additional advantage over noise jammers in that their output signals have waveforms similar to those of the threat radar, and they are therefore less likely to be screened by the radar’s noise reduction circuits.

False Target Generation

20. The aim of false target generation is to produce on the victim radar’s display a set of returns which look like targets, making it difficult for the operator to decide on a correct course of action. The DECM equipment receives the radar transmission and either re-transmits it or uses it to trigger the transmission of a stored replica. This transmission can be made with a delay or in anticipation of a radar pulse to produce returns which vary in range, or it can be injected into the sidelobes at a proportionally higher power resulting in false azimuth returns. Suitable programming of the equipment can cause a matrix of echoes to be built up on the victim radar’s display as shown in Fig 8.

14-24 Fig 8 False Target Matrix on PPI Display



21. Although false targets outside of the target’s range are relatively easy to generate, the production of credible targets elsewhere is rather more difficult. To produce false returns inside the target’s range the deception signal must be transmitted before the radar pulse arrives at the jammer. Whereas this is a relatively straightforward technique if the radar has a stable PRF, it is less so if the radar is PRF agile or

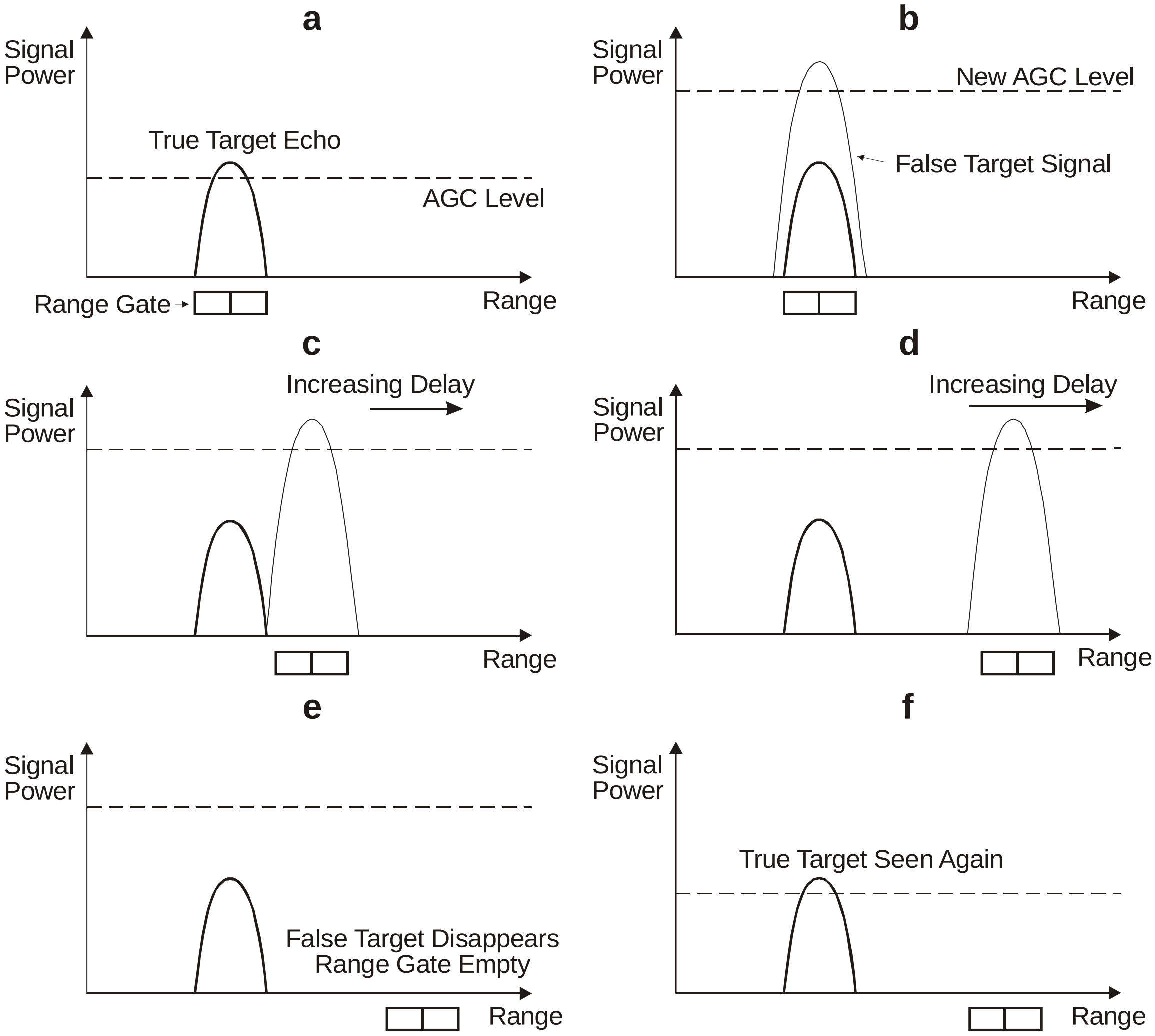
unstable due to design or maintenance problems. In this case the false target range will vary from pulse to pulse and is likely to be recognized as false by the operator. To produce a target on a false azimuth it is necessary to inject the deception signals into the radar sidelobes and ideally at a power level inversely proportional to the ratio of the sidelobe power to the main lobe power. If the signals deviate too widely from this ideal the operator is likely to be able to differentiate between real and false echoes. The successful placement of credible returns over the whole of a PPI will only be possible if a computer is incorporated into the ECM equipment.

Gate Stealing/Angle-lock Breaking

22. Fire control radars require to track the position (angle and range) or velocity of a target in order to guide a weapon successfully; gate stealing or angle-lock breaking DECMs seek to negate this capability.

23. **Range Gate Stealing**. The technique of range gate stealing or range gate pull-off (RGPO) is shown in Fig 9. In a typical pulse radar system a range gate is initially placed in coincidence, either manually or automatically, with the true target return and range gate circuits then maintain the coincidence enabling range data to be continually available. Once the radar has been range-gated, automatic gain control (AGC) circuits set themselves to the appropriate level commensurate with the echo signal strength (Fig 9a). The victim radar's signal is received by the jamming equipment where it is amplified and re-transmitted with minimal delay so as to provide a very strong signal at the radar. This causes the AGC circuits to adjust to a higher level, suppressing the true target, and seducing the tracking gates into following the false target (Fig 9b).

14-24 Fig 9 Range Gate Pull-off (RGPO)



The false signal is successively delayed or advanced away from the true target echo on a pulse-to-pulse basis (Fig 9c and Fig 9d) until after a while the false target generation is stopped and the tracking is lost (Fig 9e). The AGC circuits then adjust until the true target is seen once again (Fig 9f) but if re-lock is attempted the RGPO process starts again. As there is a finite delay between the reception of the radar signal and the transmission of the deception signal, range gate stealers can be defeated by some radars by employing leading edge trackers. This in turn can be overcome by anticipating the pulse arrival at the jammer, but only if the radar has a stable PRF. RGPO can operate against PRF agile radars by using a technique known as cover-pulse jamming in which a relatively long pulse of energy is transmitted such that it is likely to cover the range of PRF agility. The transmission is triggered by the previous radar pulse.

24. **Velocity Gate Stealing**. Velocity gate stealing works on an analogous principle to range gate stealing. Doppler type radars determine and track target velocity using velocity gates and these can be seduced by a jamming signal which shifts in frequency rather than time. The shift can be up or down the frequency spectrum but the rate of change must be limited to that acceptable as a target acceleration by the victim radar.

25. **Angle Deception**. Angle deception should have the highest priority in deception jamming since, even if range or velocity deception techniques are successful, the defender may be able to employ a weapon using 3-point guidance techniques. The appropriate angle breaking technique depends on the type of angle tracking used by the radar:

a. **Mechanical Track-while-scan Radar**. Mechanical track-while-scan systems may be defeated by noise jamming into the main and sidelobes, with proportionally more power into the sidelobes, so that the operator is faced with a matrix of possible target positions and with little chance of employing 3-point guidance.

b. **Overlapping Lobe Radar**. Overlapping lobe radars, such as conical scan or sequential lobing systems, obtain tracking information by correlating angular error signals with lobe position so that servo systems drive the aerial to centralize the target. Inverse square wave jamming equipment receives the scan modulated pulses, inverts the modulation, and re-transmits the signal causing false error signals at the radar which drive the aerial away from the target.

c. **Lobe-on-receive-only Radar**. Lobe-on-receive-only radars are more difficult to deceive, as the scan rate is not readily apparent. However some opportunity may be conceded by the unintentional induction of scan rates onto the transmitted beam by the lobing action of the receiver. Another technique is to slowly sweep the jammer output amplitude modulation through the likely range of scan rates in order to match the actual scan rate and thus induce errors in the tracking.

d. **Monopulse Radar**. Monopulse radars are probably the most difficult to deceive electronically but four techniques are available:

(1) **Cross-polarization Jamming**. All of the beams of a monopulse system will exhibit one orientation of linear polarization. However, if the antenna is of a reflector type it will respond to signals with the orthogonal polarization through the sidelobes. Thus if a jammer radiates powerful signals with this orthogonal polarization into the sidelobes, the radar would assume that the jamming target is in the main beam and angle track would be broken. The technique has the disadvantage that any variance from exact cross-polarization would provide a signal component readily seen by the main beam and the jammer would effectively become a beacon. Flat plate planar array antennae are not susceptible to cross-polarization jamming.

(2) **Cross-eye Jamming**. Cross-eye jamming employs two widely separated coherent transponders on the same airframe emitting signals 180º out of phase. The jamming signal appears as a distorted wavefront at the radar and, as the radar will seek to align itself normal to the received signal, angle track breaking will be achieved.

(3) **Co-operative Modulated Jamming**. Co-operative modulated jamming is a method whereby two or more aircraft employ non-coherent jamming simultaneously against the same victim radar. The technique causes the fire control radar to centre its angle tracking on the centroid of the multiple jamming sources rather than on an individual target. All of the jamming sources must lie within the beamwidth of the radar and there are severe constraints on aircraft manoeuvre while jamming is being attempted. The technique cannot be used against multiple radar sites simultaneously.

(4) **Terrain Bounce Jamming**. Terrain bounce jamming is postulated primarily for use against semi-active missiles. The jammer illuminates the Earth's surface in front of the aircraft so that it bounces up to the missile and causes the missile to home into the ground. The signal must be very strong to overcome losses at the surface and there must be no sidelobes at or above the horizon on to which the missile might home.

RADAR ECM - PHYSICAL AND MECHANICAL METHODS

Introduction

26. On-board electronics do not provide the only ECM techniques, and other methods may be employed either instead of, or in conjunction with, electronic devices. The employment of decoys, chaff, and tactics will be discussed in the following paragraphs; stealth technology, which is a passive technique, will be covered separately.

Decoys

27. An airborne decoy is an expendable or semi-expendable device which is launched from the target aircraft, or from the ground, with the objective of simulating a genuine target. A physically small decoy can be made to be indistinguishable on radar from the aircraft to be protected by the use of corner reflectors which increase and modify the decoy's radar cross section. Decoys may be fitted with noise and deception jammers, and with chaff dispensers, commensurate with the physical constraints of size, payload, and available power. Since the prime aim of decoy craft is to mimic the real aircraft target they should duplicate the speed and height of the launch aircraft and they therefore tend to be expensive. Nevertheless, future fire control radars are likely to be very difficult to jam successfully and decoys may be the only viable option. Current development is toward air-towed decoys.

28. Decoys are also used to simulate potential ground targets in order to deceive attacking aircraft. They may employ deception transmitters or may be mechanical reflectors. Active decoys can be used to seduce anti-radiation missiles away from the main radar head either by acting alone in an expendable mode, or by working in unison with the main radar to shift the overall power centroid. Passive decoys can be used to simulate ground radar features such as roads, bridges, power stations, and airfields in order to confuse radar guided missiles.

Chaff

29. Chaff is a general term covering all elemental passive reflectors, absorbers, or refractors of radar, communication, and other weapon system radiations, which can be floated or suspended in the atmosphere for the purpose of confusion, screening, or otherwise adversely affecting the performance of the victim electronic systems. The most commonly encountered example consists of thin metallic or metallic-coated dielectric strips of various lengths and frequency responses that cause confusion targets, clouds, or corridors on victim radar displays.

30. Most chaff is made of solid aluminium, aluminium coated glass fibres, or silver coated nylon filaments packaged in small units containing hundreds of thousands of elements, and light enough to allow an aircraft to carry large quantities.

31. Chaff is usually cut to match the half-wavelength of the victim's radar frequency to create a half-wave dipole, and as such it re-radiates the incident radar energy very efficiently; a small package of chaff may have a radar cross section similar to, or greater than, that of the dispensing aircraft. Typically a package will contain varying lengths of chaff filaments in order that it will respond to a variety of frequencies. Long lengths of chaff material, known as 'rope', are sometimes used against low frequency, long range, radars, although they are less effective than tuned lengths.

32. Chaff can be employed in large clouds in an attempt to hide attacking aircraft, in a large number of randomly dispensed small packages designed to create false target returns, or as a self-protection measure by an aircraft which has been locked on to by a terminal threat radar. Developments in radar technology have greatly reduced the effectiveness of the first two applications, and chaff is now mainly used as a self-protection aid. It can be quite effective in breaking radar lock when dispensed in small discrete packages, or in a short series of packages, in conjunction with a hard turning manoeuvre.

Tactics

33. Technological counters to an improvement in an enemy's radar or weapon systems are not usually realized quickly, and tactical changes, either to the flying pattern or to the use of existing equipment, will usually be the quickest countermeasures to be developed.

34. Flying at low level makes it more difficult for enemy systems to engage an aircraft due to the limited radar horizon, terrain screening, and the difficulty of achieving a correct weapon fuzing.

35. Tactical countermeasures may also involve evasive manoeuvre, probably in association with electronic or physical measures. A change of heading, altitude, or speed whilst an enemy radar is degraded by ECM is likely to make re-acquisition by the operator more difficult. Long-term degradation or destruction of the enemy system, although desirable, is probably not achievable. However it is only necessary to create confusion or delay for the time taken to escape from the weapon's engagement envelope.

RADAR ECM - STEALTH

Introduction

36. Probably the best ECM is to avoid detection, and the concept of stealth involves the reduction of the detectability of an aircraft, or any other vehicle, with the aim of enhancing its survivability.

37. The fundamental parameter encountered when dealing with radar stealth techniques is the radar cross section (RCS). RCS may be defined as a hypothetical area intercepting that amount of power, which, when scattered equally in all directions, produces an echo at the radar equal to that of the target. RCS therefore provides a measure of radar visibility; an aircraft with a large RCS will be more detectable than one with a small RCS under the same conditions.

38. An aircraft does not have a fixed RCS - the value depends upon the illuminating radar parameters, the aircraft's shape, size, orientation with respect to the radar, and construction material. Also, since it is not possible to take any actions to change an enemy's radar parameters, the range and radar factors will not be explored further.

Effect of Size

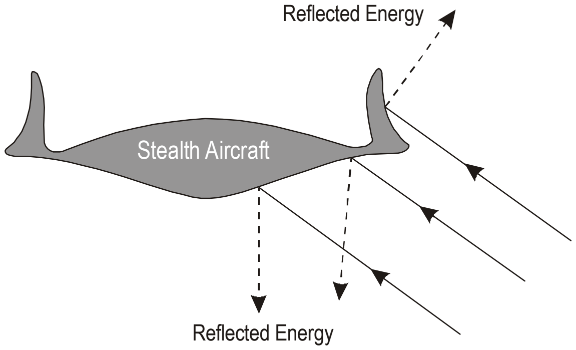
39. In general the RCS of an object increases with increasing size. However resonance effects, which can greatly enhance RCS, can occur when radar waves impinge upon an object which is of a comparable size to the radar wavelength. Whereas, as a whole, an aircraft is larger than any radar wavelength, there are often components, such as cabin conditioning intakes and gun muzzles, which may respond resonantly to, say, an I-band radar. This potential problem can be alleviated to an extent by incorporating the smaller components into a smooth, blended, structure.

Effect of Shape and Aspect

40. Shape and aspect are considered together as the shape of an object will depend upon the aspect presented to the illuminating radar. The RCS of a complex shape, such as an aircraft, depends upon its constituent shapes. In particular a corner reflector has a high RCS compared to its physical size, and one of the principle design aims therefore is to reduce the number of near 90º corners on an aircraft, by, for example, blending the wings into the fuselage. A particularly reflective area is generated by the external carriage of weapons, fuel tanks, and ECM equipment, where the joints between wings, pylons, and stores provide a large number of 90º corners. This problem can be overcome by the internal carriage of stores.

41. The other main sources of reflection are planar surfaces orientated normal, or near normal, to the incident radiation. It will be virtually impossible to eliminate this effect for all viewing aspects, but some reduction in RCS can be achieved by arranging for fins to slope inwards, and by having highly swept leading edges to wings and tailplanes. Ideally an aircraft fuselage cross-section would be one where virtually all of the incident radiation was reflected away from the illuminating radar, as illustrated in Fig 10.

14-24 Fig 10 Stealthy Fuselage - Cross-section



42. Any cavity will efficiently reflect radar - cockpits and engine intakes being good examples. As well as being a resonant cavity, a cockpit normally contains a large number of corner reflectors. One method of reducing the RCS of a cockpit is to coat the canopy with a thin metallic film - gold and indium tin oxide have been used. The RCS of engine intakes is increased by the presence of the LP compressor face, so deep intakes are desirable. The problem could also be reduced by placing the intakes so that they were not visible to the likely threat radars (e.g. above the fuselage to counter ground based radars) but aerodynamic considerations will frequently make this approach unsuitable. More practically, the intakes can be shielded by the use of an intake bullet or lined with radar absorbent material.

Materials

43. One method of reducing the RCS of an aircraft is to reduce the absolute amount of energy being reflected by the use of radar absorbent material.

44. Ideally such material would be of negligible thickness and weight, would be effective over a wide bandwidth and over a wide range of angles of incidence, and could either be moulded to form complex shapes and surfaces, or could be stuck on to them. There is, of course, no one material which completely satisfies all of these requirements but radar absorbent materials suitable for use in aircraft structures have been developed, as have radar absorbent paints.

COUNTERMEASURES TO INFRA-RED HOMING MISSILES

The Infra-red (IR) Homing Missile

45. Historically, most aircraft which have been shot down by guided missiles have been hit by the IR-homing type, so clearly the development of countermeasures to this threat is highly desirable.

46. The IR energy onto which the missile can home can be considered to emanate from three sources:

a. The jet pipe hot metal.

b. The exhaust plume.

c. The kinetically heated leading edges of the aircraft structure.

47. The threat can be countered by using IR stealth measures aimed at preventing the missile from achieving a lock, by using decoys, or by using noise or deception techniques to jam the missile guidance circuits.

Stealth Measures

48. Stealth measures are directed towards reducing the IR visibility of an aircraft, the main areas for consideration being the jet pipe and the exhaust plume, together with a number of miscellaneous 'hot spots'.

49. **Jet Pipe**. Shielding of the hot metal of the jet pipe from the view of an IR missile is especially effective against older IR missiles which use un-cooled sensors and need to home onto the short wavelength energy emitted by the very hot metal. Un-reheated engine exhausts can be shrouded, and helicopter jet pipes are particularly suitable for such treatment.

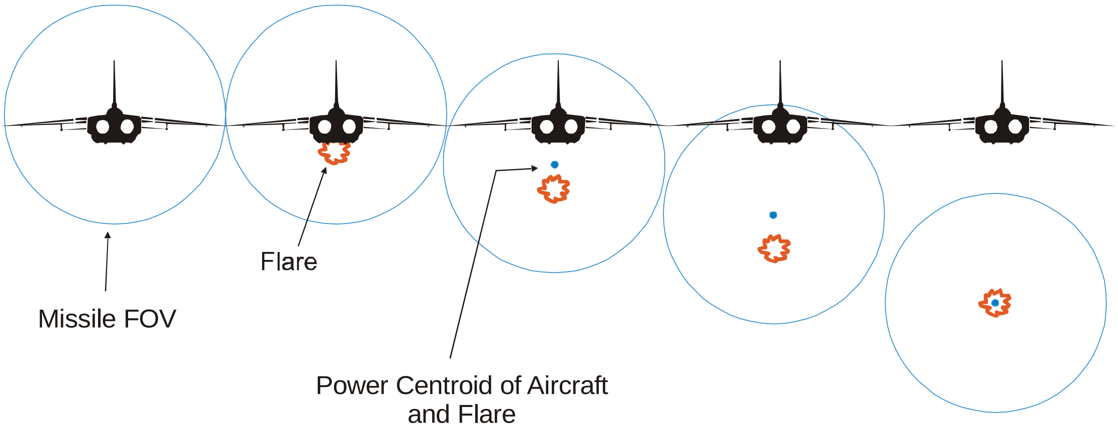
50. **Exhaust Plume**. Missiles with cooled seeker heads are able to home onto the energy emitted by the hot jet exhaust gases. The problem is partly overcome by the use of high bypass ratio turbofan engines where the hot gas is mixed with the cooler bypass gas before expulsion. If afterburners are in use when a missile launch is seen or suspected then it is essential that the engine power is reduced to 'military range' to reduce the IR radiation. Helicopters are often fitted with exhaust deflectors which reduce the plume’s temperature by forcing it into the rotor downwash.

51. **Hot Spots**. Powerful landing lights or searchlights can provide a usable source of IR energy in some circumstances. The most effective measure is to switch off if a missile launch is seen or suspected, but if this action is unacceptable then it will be necessary to resort to decoys or jammers. In some circumstances sun glint from glossy surfaces can be sufficient for a missile to acquire lock - this can be overcome by the use of matt or semi-matt finishes.

Infra-red Decoys

52. The aim of infra-red decoys is to present a target to the missile which is more compelling than the aircraft. Decoys are in the form of flares which emit energy at all appropriate IR frequencies, and the missile guidance transfers lock either to the flare or to the power centroid between the aircraft and the flare. As the flare falls away, the missile seeker follows the movement until the aircraft is no longer within the field of view (FOV) as shown in Fig 11. In order to be successful the flare must ignite quickly so that it is sufficiently bright while still in the missile’s FOV, and then burn for long enough so that the aircraft can escape from the FOV.

14-24 Fig 11 Flare Decoying of IR Homing Missile



53. Decoys are not without drawbacks. The initial problem is that only a few aircraft, as yet, are fitted with missile approach warners, and so there is still a reliance upon the aircrew seeing the launch flash, being told of the launch by other aircraft in a formation, or simply releasing decoys in a predetermined pattern as a precautionary measure during a period of maximum susceptibility. Most aircraft will not be able to carry enough flares for them to be used indiscriminately.

54. If a series of flares is initiated manually there is a danger, if the time gap between flares is too short, that the missile will transfer lock up the series until it arrives back at the aircraft - the 'stepladder' effect.

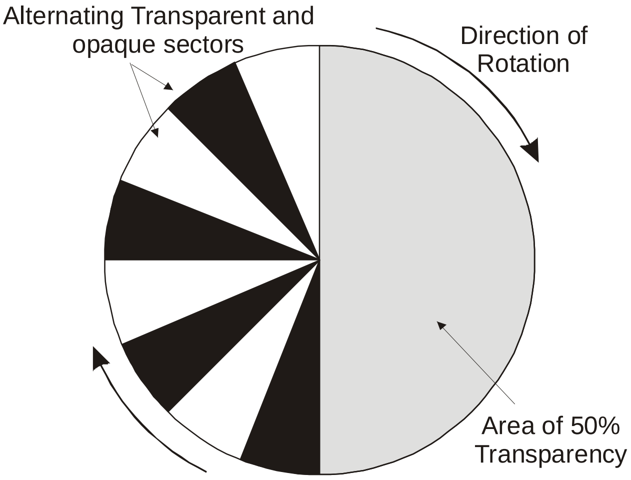
55. In addition to artificial flares, many missiles can be decoyed by flying so that the sun, or a strong solar reflection from a cloud edge, enters the missile FOV. Similarly, IR missiles may be confused if presented with strong IR sources on the ground, eg fires or explosions.

Infra-red Jamming

56. The tracking system of an IR missile employs a device which modulates the incoming infra-red energy, either in amplitude or frequency, in response to the direction of the infra-red source with respect to the missile boresight. The aim of an infra-red jammer is to disrupt this modulated signal so as to cause errors in the missile guidance.

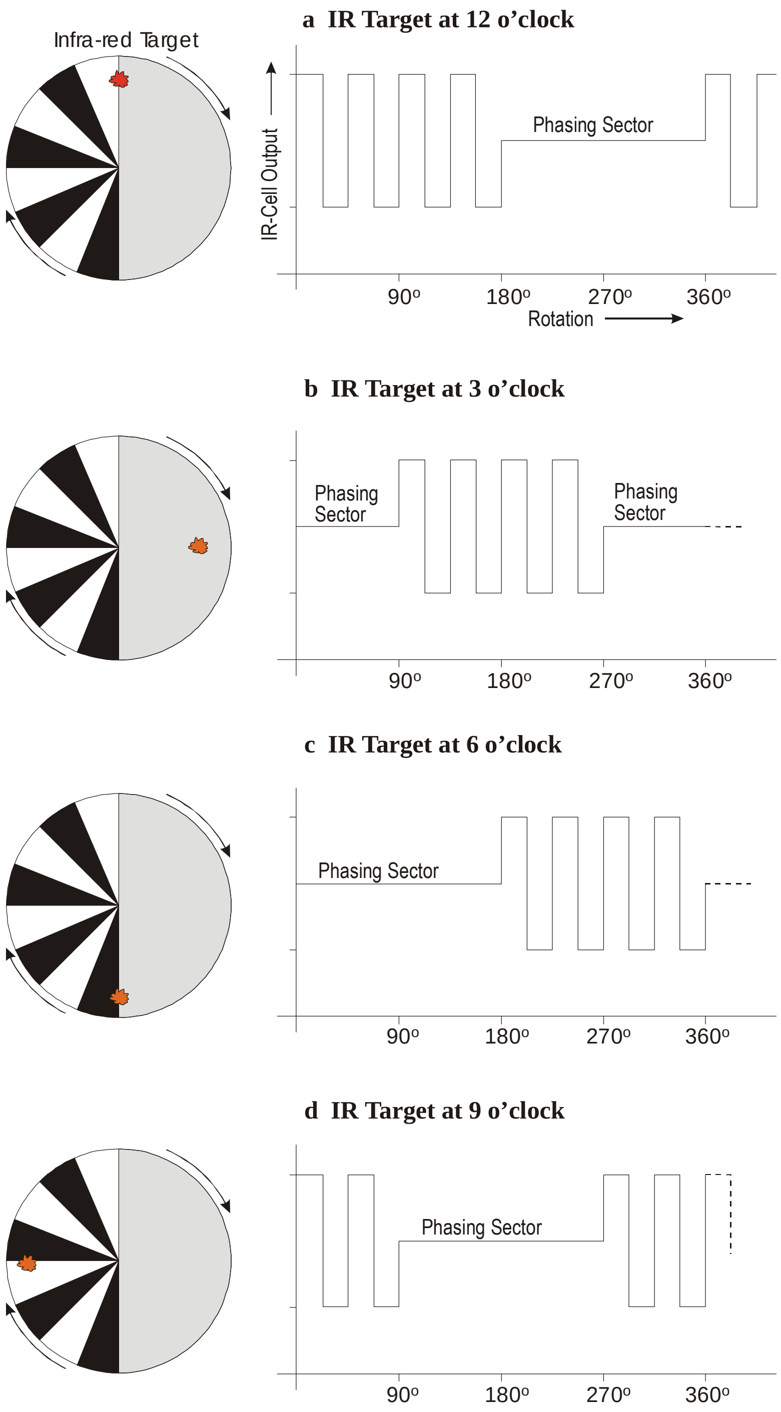
57. In the missile the incoming infra-red energy is focused onto an infra-red sensitive cell by a set of optics. In order to produce an output signal which is modulated in response to the source direction, one of two systems is generally used. Each system incorporates a reticule which has a pattern of opaque and transparent sectors. In the spin scan seeker head the reticule rotates, while in the conical scan seeker head moving mirrors in the optical system move the image of the infra-red source over a fixed reticule. The pattern of this 'chopper' disc varies from missile to missile, but a simple type of spin scan reticule, known as a 'rising sun' reticule, is shown in Fig 12 and its mode of operation will serve to illustrate the principle of IR jammers.

14-24 Fig 12 A 'Rising Sun' Reticule



58. Half of the disc has alternating transparent and opaque sectors; the other half is an area of 50% transparency. Fig 13a shows the output from the sensitive cell for a target appearing near the 12 o’clock position on the reticule. As the disc rotates the IR energy will be alternately passed and stopped during the first 180º of rotation, and then 50% of the energy will be allowed through for the second 180º. This 50% transmission interval is known as the phasing sector. Figs 13 b, c and d show the output signals for targets at 3, 6, and 9 o’clock respectively. This output signal is used to modulate a carrier wave which, by comparison to a reference signal, allows the position of the phasing sector to be determined and thereby a tracking error signal to be generated. If the target is on boresight there is no modulation, and therefore no error signal.

14-24 Fig 13 IR Cell Output for Various Target Positions



a IR Target at 12 o'clock

b IR Target at 3 o'clock

c IR Target at 6 o'clock

d IR Target at 9 o'clock

59. **Infra-red Jammers**. The IR source in a jammer is typically either a doped graphite rod, an alkali metal vapour lamp or a LASER. The graphite rod is contained in a sealed case of sapphire and is heated to 1,800 K or 2,000 K. Graphite sources are usually limited to helicopters and light aircraft. The element of an alkali vapour lamp is commonly caesium or rubidium and is heated to between 3,800 K and 4,000 K. The metal then forms a plasma between the lamp's two electrodes so that when a current is passed a large IR pulse is emitted. The output of the alkali vapour lamp is significantly greater than that of the graphite source and so this type of jammer is commonly found on large helicopters and transport aircraft. Lamp based IR jammers are commonly used pre-emptively, whilst the Large Aircraft Infra-red Countermeasures (LAIRCM) are used as a reactive system to protect large aircraft and helicopters from IR missiles. The LAIRCM is a turret mounted, directional, LASER based system that is usually positioned below the host platform to give a greater field of view.

60. **Noise Jamming**. For noise jamming applications the IR output is modulated to produce a constant stream of pulses, ideally at or close to the chopping rate of the missile seeker head. The effect is to add pulses to that part of the seeker head output signal which was originally the phasing sector, and to alter the amplitude of the originally pulsed part of the signal. The overall result is to reduce the size of the tracking error signal which reduces the missile's acquisition range, reduces the seeker's tracking sensitivity, and prevents the seeker from following a highly manoeuvring target.

61. **Deception Jamming**. For deception jamming purposes the output of the jammer is modulated to produce a series of pulses similar to that produced by the reticule when it is tracking an off-boresight target. With a suitable jam/signal power ratio, the jammer pulses will tend to dominate and, if out of phase with the reticule output, will put pulses into the phasing sector. The carrier wave output will have an amplitude and phase altered from that due to the target alone and false tracking signals will result. When the missile guidance circuits receive the false phasing information from the reticule system, the missile will attempt to fly to bring the false target onto boresight and, in doing so, lose the real target from its FOV. Deception jammers are able to induce much larger miss distances than noise jammers, and although more complex, can require less power if an electrically modulated lamp source is used.

62. **Effect**. Jammers are less effective against conical scan seeker heads than they are against the spin scan system described. In general they will fail to break lock completely but will induce a small tracking offset which will result in a relatively small miss distance.

COUNTERMEASURES TO ELECTRO-OPTICAL SYSTEMS

Techniques

63. The use of passive viewing systems by an enemy is extremely difficult to detect and, if a system is detected, and a countermeasure is used, it is very difficult to assess the effectiveness. Against a purely optical system it will be necessary to attack the operator, but the techniques of contrast reduction, deception, obscuration, jamming, and tactics can be employed to defeat an electro-optical system.

64. **Contrast Reduction**. Contrast reduction, or camouflage, is a well-established technique. Aircraft can be painted to blend with their likely background and these paint schemes may need to be altered as theatres of operation change. Camouflage nets are commonly used to conceal ground targets, although to be fully effective they should be responsive at both visible and infra-red wavelengths.

65. **Deception**. It is possible to produce decoys to closely resemble many types of target regardless of the type of sensor employed, for example, a decoy aircraft or tank could be finished with IR reflecting material, and equipped with radar reflectors.

66. **Obscuration**. All electro-optical systems are degraded to some extent in the presence of cloud, haze, fog, or smoke. Thus the introduction of such a scattering medium into the field of view can be an effective countermeasure, although consideration must be given to both visible and infra-red wavelengths. Terrain or vegetative masking might also be classified as an obscuration countermeasure.

67. **Jamming**. Lasers or flares can be used to overload an indirect viewing system, or to distract the operator or damage his eyes. The flashing of very bright lights at certain rates can also be employed to disable the operator.

68. **Tactics**. The employment of tactics to defeat electro-optical systems include planning attacks so as to maximize terrain screening and ideally such that the sun is behind the aircraft. Manoeuvres can be employed to escape from the system's FOV or to fly into cloud to achieve obscuration.

LASER COUNTERMEASURES

Laser Uses and Countermeasures

69. Lasers are currently used in target designators, rangefinders, and communications links. However laser damage weapons may become operational in the future, either in the form of eye damage weapons, or as high-energy lasers capable of inflicting a 'hard kill' against equipment. Military lasers operate in the visible or infra-red bands and the radiation is subject to the same attenuation and scattering effects as non-coherent radiation in these bands. Laser countermeasures generally use one or more of the following techniques:

a. Absorption.

b. Reflection.

c. Ablation.

d. Jamming.

70. **Absorption**. Rangefinders, designators, and communications links can be disrupted by the use of smoke, the particles of which are able to absorb and scatter the beam. However, it seems unlikely that active smoke dispensers could be employed by high speed aircraft, although some protection may be afforded by clouds and perhaps by the smoke of battle. Land targets on the other hand may well be able to use smoke screens as a viable countermeasure. Eye damage weapons could in theory be countered by the use of narrow band filters built into the aircrew visor or into the aircraft cockpit. However, the wide range of laser frequencies available would entail the use of a large number of filters, and the total light attenuation would be unacceptable. Devices which darken when light falls on them have been proposed but at present their reaction time is too slow. The most promising current line of development is to decouple the eyes from the outside world by using some form of indirect viewing system. Such a system could be made more tolerant to high laser energy levels than the eye.

71. **Reflection**. A possible counter to laser damage weapons is to use highly reflective aircraft finishes, although these are not effective against designators and trackers. Unfortunately these reflective finishes are difficult to maintain in practice, particularly in a combat situation, and of course defeat any attempt at visible camouflage.

72. **Ablation**. Ablative coatings are a possible counter to 'hard kill' laser damage weapons. These coatings absorb the incident energy and either melt or flake away and thereby protect the underlying structure. The technique has the disadvantage of imposing a significant weight penalty.

73. **Jamming**. Rangefinders, designators, and communications links employ amplitude or frequency modulation and are therefore theoretically susceptible to jamming. Unlike radars, however, laser systems have no sidelobes and the jammer must therefore be placed in the narrow main beam. If this can be achieved, and the laser wavelength is known, noise jamming is relatively easy. Deception jamming is difficult since the laser uses a very high frequency and very wide bandwidth and therefore very complex coding.

COMMUNICATIONS COUNTERMEASURES

General Considerations

74. 'Communications' in these paragraphs refers to the transmission and reception of messages by means of radio. The messages may be in speech or data form and the system used may be teleprinter, telegraphy, telephony, data link or a visual system.

75. **Bandwidth Considerations**. On average, human hearing covers a frequency range from about 20 Hz to about 16 kHz but for acceptable clarity of speech it is only necessary for the ear to receive up to about 4 kHz. The sound waves are converted to electrical signals which modulate an electromagnetic carrier wave and the bandwidth around the carrier frequency extends over a few kilohertz. In a typical aircraft communications system the carrier frequency may be about 200 MHz, although in general communications carrier frequencies may lie between 100 kHz and several gigahertz. Messages may be sent in coded form rather than as speech. As with radar, when pulses are used to modulate a carrier, the bandwidth necessary for adequate reproduction is much higher than that adequate for simple speech modulation. As an example, for a pulse width of 0.25 μsec, a bandwidth of about 4 MHz would be required.

76. **Propagation Considerations**. Radio communications may be affected by using either stationary or mobile transmitters and receivers. Although it may be easier to determine the requirements of a jammer if it is to be used against a static site, it may in practice be more difficult to position the jammer satisfactorily due to geographical considerations.

77. **Antennae**. A communications aerial does not scan like a search radar antenna, but either transmits omnidirectionally, or is directional - in a microwave link the beam is highly directional and very narrow.

78. **Reception**. In most cases a receiver is concerned only with one-way propagation direct from the transmitter and thus the signal power at the receiver is essentially inversely proportional to the square of the distance from the transmitter, although this simple relationship will often be modified by the effects of the local terrain. Some systems do not rely on direct transmission but make use of ionospheric refraction or forward scatter to achieve long distance communication.

Electronic Jamming

79. The principle drawback to the noise jamming of enemy communications is that the jamming is likely to create similar havoc to friendly communications and intelligence gathering, both on the jammed frequency and on close frequencies within the spread of the noise spectrum. Intentional communications jamming has therefore to be carefully controlled and authorized. It is possible nevertheless to plant short-range jammers, perhaps dropped by parachute, adjacent to enemy communications centres so that the likelihood of interference to friendly systems is reduced.

80. Communications noise jamming can take the same forms as radar jamming, ie spot jamming, sweep jamming, barrage jamming, or some hybrid form. Ordinary communications transmitters can be used as jammers if set to the victim frequency and modulated with voice, music, tones, or noise. In addition, deceptive messages can be included in the hope that when filtered from the noise the enemy will believe them to be genuine.

81. The sidelobes of communications aerials are considerably down on power compared to the main lobe and thus it is necessary for jamming to be injected into the main lobe. This can be difficult in, for example, microwave links which have very narrow highly directional beams; the jamming would have to originate somewhere close to the line joining the two aerials.

82. Satellite communication systems present a rather different situation. The satellite is usually communicating with a number of ground stations and its radiation pattern is consequently broad. Conversely, the ground station directs a very narrow beam continuously at the satellite. As a result, jamming normally has to be concentrated on the up-link to exploit the satellite receiver's broad radiation pattern. Since these systems use short wavelength radiation, the jammer aerials can be designed to have narrow beams which reduce the probability of interfering with friendly communications.

NAVIGATION AIDS COUNTERMEASURES

Types of Radio Navigation Aids and Countermeasures

83. Radio navigation aids can be broadly classified into three categories; those giving bearing, those giving range, and hyperbolic systems generating a fix. Although it is quite feasible to jam most radio navigation aids, it is debatable whether an enemy would consider it worthwhile to expend energy and resources in doing so, especially as most military aircraft can operate satisfactorily without using external aids.

84. **Bearing Aids***.* Radio aids which give a bearing from a ground station may operate either by determining the direction of arrival of the wavefront using a directional aerial (ADF), or by phase comparison techniques (eg VOR or TACAN). ADF systems are easily deceived by broadcasting on the same frequency using a CW transmitter. This creates a composite signal and the directional aerial will indicate a direction towards the mean of the two components. VOR and TACAN can be countered by broadcasting a carrier signal on the same frequency as the beacon, and modulated in the same way, which would cause errors in phase measurement. In order to attack the receivers on low flying aircraft it is likely that the jammers would need to be airborne.

85. **Range Aids***.* The range element of TACAN and DME operate identically and a lock can be broken by constant transmission of noise-generated pulses of sufficient amplitude. The jammer must be line of sight with the ground beacon and this constraint implies using airborne equipment.

86. **Hyperbolic Systems**. Noise jamming is not likely to be simple or effective against hyperbolic systems, due to the narrow bandwidth of the systems. The most appropriate techniques involve copying and re-broadcasting the signals leading to phase measurement errors in the equipment. Because of the large power and aerial size requirements any jammers would have to be ground based. Navstar is very resistant to ECM and it is unlikely that the system could be degraded electronically over a wide area.

CHAPTER 26 - ELECTRONIC DEFENCE (ED)

Introduction

1. Electronic Defence (ED) involves all actions taken to protect personnel, facilities and equipment from any effects of friendly, or enemy, employment of Electronic Warfare (EW) that degrade, neutralize or destroy friendly combat capability.

2. A resolute enemy could put any electronic system out of useful action provided that he was prepared to allocate the necessary resources. One of the principal aims of ED is to make the cost of carrying out such a campaign unacceptable.

3. Much can be done to effect ED before any equipment is introduced into service by careful overall planning, and by designing equipment with EW in mind so that susceptibility to Electronic Attack (EA) is minimized. Once an equipment is in use much reliance will be placed on appropriate tactics and on operator skills.

Overall Planning

4. **Frequency Allocation**. One of the main considerations at the planning stage is the allocation of operating frequencies within the usable parts of the spectrum. The allocation will depend upon knowledge of both the military and technical requirements of the numerous radar and communications systems, and the problems caused both by electronic warfare and by interference effects between friendly systems. In a war situation an ill-considered frequency allocation policy may well make the enemy's EA task easier, and possibly unnecessary if mutual interference causes significant problems. Alternative communications frequencies should be allocated so that a reversion can be made if jamming becomes intolerable, and if intelligence or ESM has made enemy frequencies known, these can be used as a last resort.

5. **Operating Procedures**. The establishment of standard operating and communications authentication procedures will make it harder for the enemy to enter a network by posing as a friendly station and thereby introducing deception messages.

6. **Parallel Operation**. The enemy will inevitably attempt to exploit any weaknesses in a radar or communications system. Any potential weak points should therefore be identified at the planning stage and protection measures implemented. One remedy is the use of parallel operation in which two or more devices are available to do the same task. As an example, more than one radar can be used to supply the same information so that if one is jammed the other may still be able to operate successfully, particularly if the two sets use widely separated frequency bands (frequency diversity). This method increases the resources needed by the enemy to jam or destroy the facility but of course this desirability must be weighed against the cost of providing the additional installations; much will depend on how vital the radar (or other) information is considered to be.

Radar Equipment Design

7. The design of any radar equipment should take due account of ED features. One such feature is frequency agility, another technique is that of diplexing in which a common aerial is fed from two separate transmitters /receivers operating on different frequencies within the same band. When one receiver is jammed the other is still able to operate and when neither channel is being jammed the two signal powers may be added to enhance the signal-to-noise ratio.

8. In many cases the inherent characteristics of a radar which are included to enable it to perform a particular task will in themselves contribute to ED; such features include MTI, pulse-width compression, and sidelobe cancellation or blanking.

9. It is desirable for ED purposes that a radar should be designed with the maximum permissible transmitter power and with a high aerial gain. This forces the enemy into using a high power jammer in order to jam the radar successfully. This implies more cost and, if the jammer is to be airborne, there may be weight and volume penalties.

10. Several anti-clutter devices have been devised and, although these are primarily intended to reduce the effect of unwanted 'natural' noise, they also provide ED to varying degrees. These techniques include swept gain, logarithmic amplifiers, fast time constant, Dicke-fix, instantaneous automatic gain control, pulselength discrimination, and pulse interference suppression.

Communications Equipment Design

11. As with radar equipment, communications equipment for military use should be designed with EW in mind and, where possible, circuits should be selected for the ED features they exhibit. There are, for example, circuits and techniques which can increase the signal-to-noise ratio, and some pulse transmission systems which will only handle pulses of a designated width, thus excluding many jamming signals.

12. The range of a transmission should be restricted to that necessary to accomplish the task as this reduces the chances of enemy interception; this implies that the equipment should be designed to use the highest practicable frequency, as this will limit the range of propagation.

13. Although some applications will need to use omni-directional aerials, the employment of directional transmission aerials will reduce the probability of signal interception, and directional reception aerials will be less susceptible to jamming signals.

14. Further ED advantages can be gained by reducing transmission time, thereby reducing the chances of both signal interception and of the enemy determining the transmitter location by D/F methods. Much can be achieved in this way by planning, procedures, and control, but equally high-speed transmission devices should be made use of where possible.

15. **Frequency Diversity and Agility**. Frequency diversity allows jamming to be avoided by switching to an alternative frequency band. Frequency agility can keep spot jamming at bay by rapidly changing the operating frequency - the enemy is continually trying to determine the current frequency rather than jamming. Frequency agility will not protect against barrage jamming, but of course this results in either jamming on any individual frequency being less intense, or it imposes a high power and therefore cost penalty on the enemy. The penalty of frequency agility is that, because of the broad band of frequencies used, the number of available communications channels will be reduced.

16. **Spread Spectrum**. The spread spectrum technique is an efficient measure in which a pseudo-random noise signal is added to the normal transmission modulation. The receiver employs an identical, synchronized, pseudo-random noise generator to remove the noise code and reveal the original voice or data signal. Since the spread spectrum signal occupies a wide bandwidth an enemy will have to resort to broadband jamming to defeat the signal. Furthermore, it will be extremely difficult for the enemy to demodulate the noise and reveal any information. The system has the additional advantage of being able to transmit several signals simultaneously provided that each transmitter and receiver pair has its own individual noise code.

Training and Tactics

17. Despite the many technological protective measures available, a large part of ED will be vested in the effective use of resources and in the skill of operators.

18. One of the most effective ways of avoiding jamming is to refrain from using the equipment. Although this is not entirely acceptable, a considerable amount of protection can be gained by restricting transmissions, since doing so reduces the enemy's chances of detection, location and jamming. Regular routines should be avoided, rather operating schedules should, as far as is possible, be pseudo-random and opportunities should be taken to vary frequencies. The location of a jammer may often be determined by triangulation or homing techniques and consideration might be given to implementing the ultimate ED measure - the destruction of the jammer.

19. Operators will often be the best ED and much reliance will be placed on their skill and initiative; this skill can only be developed through effective theoretical and practical training. It is important that operators are exposed to a variety of jamming scenarios so that they can learn to recognize and minimize the impact of these effects, and how to implement the appropriate anti-jamming techniques.