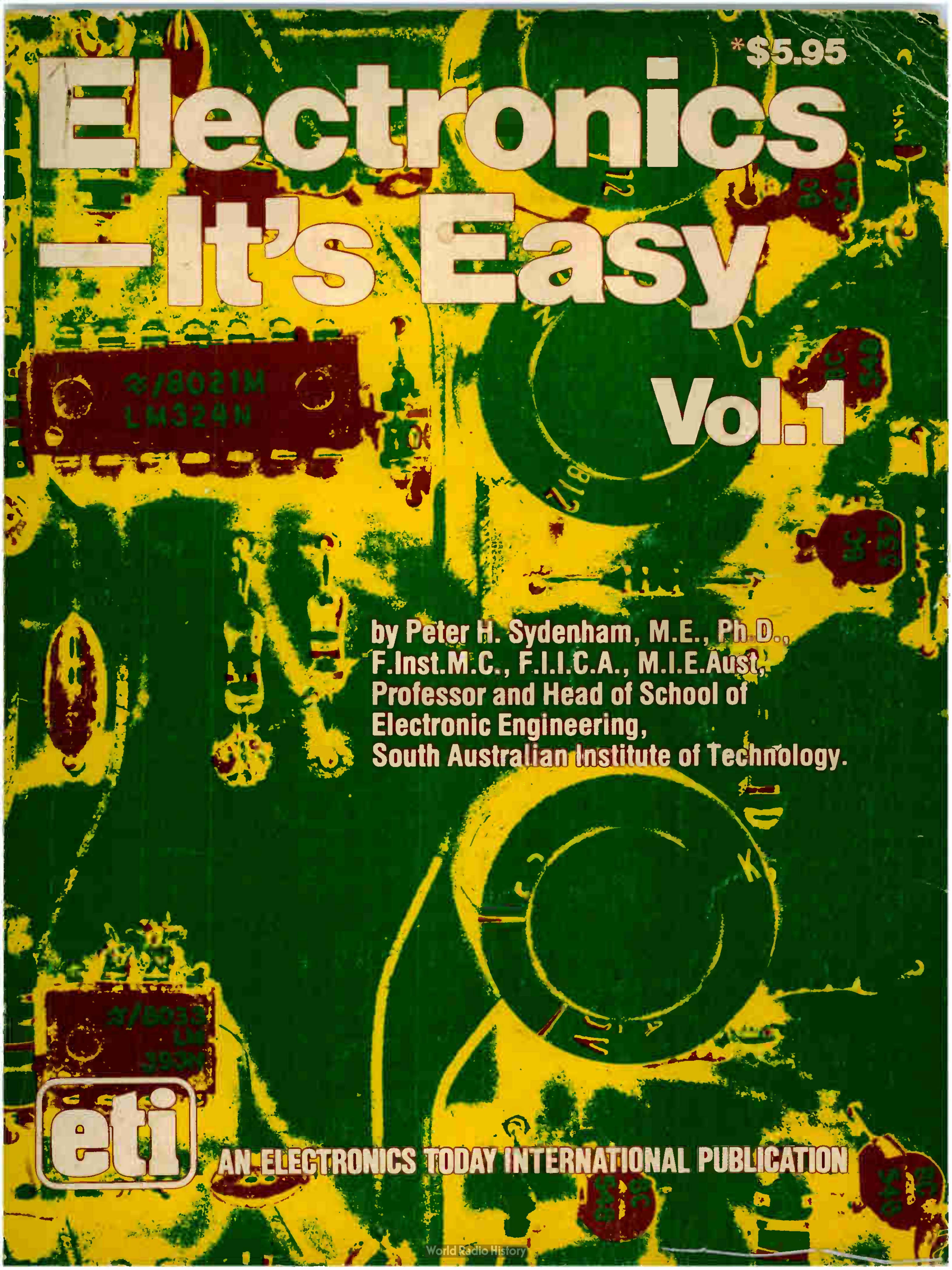


*$5.95

Electronics — It's Easy

Vol. 1

by Peter H. Sydenham, M.E., Ph.D.,
F.Inst.M.C., F.I.I.C.A., M.I.E.Aust.,
Professor and Head of School of
Electronic Engineering,
South Australian Institute of Technology.

eti
AN ELECTRONICS TODAY INTERNATIONAL PUBLICATION

Electronics —It's Easy Vol.1

by Peter H. Sydenham, M.E., Ph.D., F.Inst.M.C., F.I.I.C.A., M.I.E.Aust. Professor and Head of School of Electronic Engineering, South Australian Institute of Technology.

Special publications editor: Jan Vernon
Layout: Bill Crump
Managing editor: Collyn Rivers

Published by the Scientific Division of Modern Magazines (Holdings) Ltd, 15 Boundary Street, Rushcutters Bay, NSW 2011 in May 1981.

Printed by Offset Alpine Printing Pty Ltd, Silverwater, NSW.
Distributed by Gordon and Gotch.
 *Recommended retail price only.

Introduction

IN THE FIRST half of the nineteenth century the great Faraday coined the term 'electron' to help describe the actions occurring in the electrolysis of solutions.

His work, plus that of the Ancient Greeks, who knew about charged substances through the effects of static electricity, laid the foundations of the modern discipline of electronics.

Throughout the nineteenth century, scientists researched gaseous discharges in evacuated and gas-filled devices. In the last decades of the 1800s Edison observed the rectifying effect of the thermionic diode during his mammoth effort to find a way of making a light globe that would give a useful operating lifetime. His observation, called the Edison effect, was not however to find electronic application through his direct effort. He missed a great opportunity, for at that time there existed a need to rectify alternating current signals and also to amplify them. Edison's work was noted but not used to fulfil these needs by him. Those uses came later.

There is a popular belief that radio came into being as the result of the availability of rudimentary electronic devices. This belief is incorrect. It was the existence of working radio telegraphy that provided the circumstances needing the invention and development of thermionic tubes rather than vice versa. Marconi and many others had given the everyday world practical wireless telegraphy from 1896 onward. By 1904 lack of efficient methods for detecting and amplifying weak radio signals had become major obstacles in the further development of radio.

In 1904 Fleming made thermionic devices that could rectify alternating current signals. In 1907 Lee de Forest added a third plate to Fleming's diode device — thereby inventing an amplifying valve.

These two inventions, plus similar devices made by other people who are not popularly credited in so clear a manner, laid the way for the discipline of electronics to emerge and grow. By the 1920s electronic valves were commonplace in radio and were beginning to make an impact on other areas, such as laboratory instrumentation and industrial control.

By the early 1950s electronics was a well-established and broadly based discipline. It had by then been proved time and again that electronic methods were usually superior to the formerly used mechanical procedures for computation and control. A gigantic swing to electronics continued as designers (great and small) attempted to do everything electronically. Digital computers were devised that calculated at speeds and with power never before even considered possible.

This aspect played a dominant role in the birth of semiconductor devices that have since all but replaced valves. These led in turn to integrated circuits, which in turn allowed even more powerful signal processing to be achieved at remarkably low cost.

Electronic method moved firmly into the field of instrumentation and into the lifestyle of consumers. The introduction of integrated circuit methods led to some degree of standardisation in the overall design attitude of people using them. Thus electronics moved into an area wherein it is practicable to assemble a vast range of devices from a relatively few basic building blocks.

Today electronics plays an essential role in virtually every field of human endeavour. Wherever signals containing information are being processed and where power flow needs to be controlled you will usually find some electronics.

The discipline needs to be and can be understood by intelligent people who have not had any formal training in electrical and electronic method. This is the purpose of this book.

1 The systems approach

TO the unenlightened even the simplest process can be a mystery. Yet with training the mystery vanishes when it is seen how simple logical techniques are combined to fulfil a task.

As we try to further our knowledge of the world about us, we collect facts gained from practical research and a great deal of thinking. Then by further mental effort we construct fundamental concepts that describe the observed facts.

These concepts, are more readily understood and remembered than an enormously long list of individual facts; and, coupled with imagination, provide the power to create and understand processes of great intricacy. Furthermore, by getting down to basics it is possible to build out again in quite new directions.

It is a human habit to try and make all experience black or white, classifying it into distinct compartments having a generally accepted name — medicine, engineering, farming, etc. It seems so tidy and assists information retrieval, but there is an ever-growing awareness that life is not like this, and processes are only understood properly by a multi-disciplinary approach.

Electronics, although seen by many in the past as a self-contained subject, should, more correctly, be regarded as a universal discipline necessary to a remarkably wide range of endeavour. It is vital to communications, archaeology, medicine, language teaching, banking, education, farming — a complete list would be never ending.

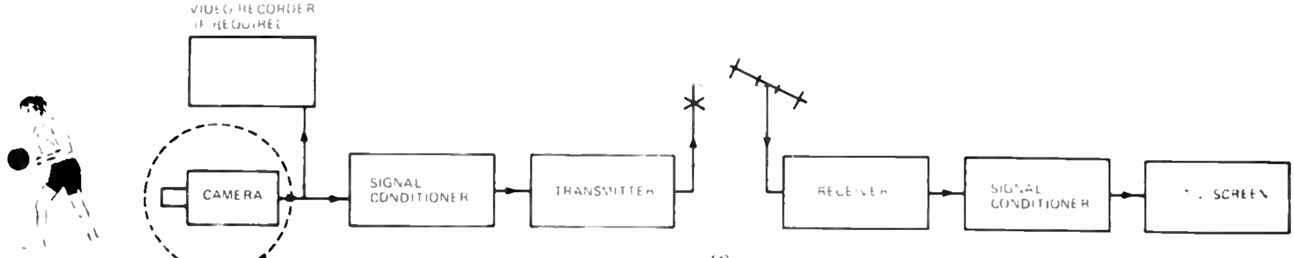
Because of this, throughout this course, our treatment of electronics will be a general basic approach — known loosely nowadays as a systems study. It will contrast with the more traditional approach, given in books and courses, by placing more emphasis on where a concept fits, rather than on how a manufacturer makes components or how the fundamental particles involved behave.

THE BLACK-BOX APPROACH

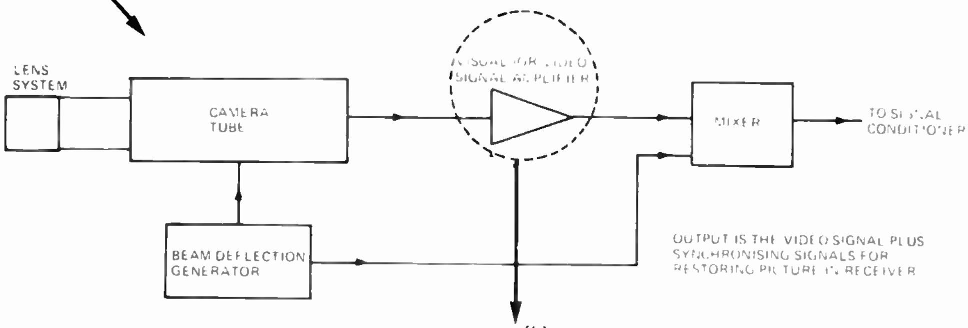
The physical world consists of numerous processes that interact with each other to form a reasonably well-balanced mammoth process. The nature of the individual processes vary enormously. In the natural world they involve such phenomena as biological behaviour and electro-chemical reactions. Man has added processes of his own that function by means of optical and

Fig. 1. As both tape-recorders perform basically similar tasks, their 'black box' representation may be the same.

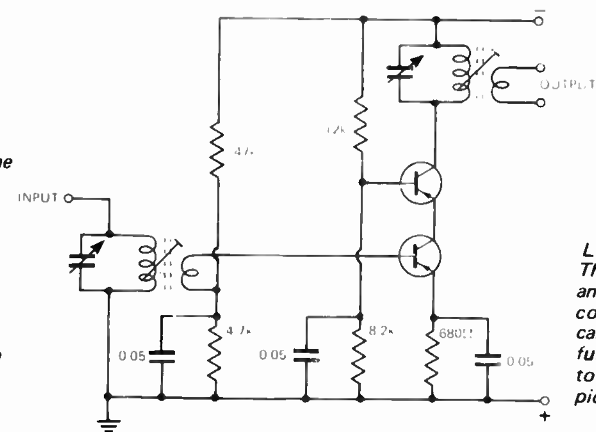
BLACK—BOX REPRESENTATION



(a)



(b)



(c)

TOP: Fig 2a). Complete TV transmitter and receiver — this 'black box' representation shows the rudiments of the system. Each individual 'box' may take one of many forms — yet the system as a whole can always be portrayed as shown here.
CENTRE: (Fig. 2b). Black box representation of the internal workings of the TV camera (shown ringed in drawing above). Here, an elementary knowledge of television techniques would be needed in order to understand the individual 'black boxes' into which the camera has now been broken down.

LEFT: (Fig. 2c). Final 'breakdown'. This drawing shows the components and their interconnections that collectively form the part of the TV camera in Fig 2b. (As will be explained further in this course, symbols are used to represent components — rather than pictures of the components themsleves).

electronic hardware put together to create the machinery needed to make life easier.

To gain an understanding of the overall function of a system of any type, we need to break it down into recognisable basic blocks that each behave in a way that is comprehensible to us. This approach also enables one to realise what else the total system might do if the circumstances were a little different. Alternatively, it should tell how to modify a block or two to obtain a different behaviour.

Such blocks in electronics are commonly called "black boxes". The behaviour of a given kind of "black box" is always the same (by definition), but the internal mechanism used to achieve the given performance could be quite different (as shown in Fig. 1).

At a systems level of study it does not matter what is *inside* the box, its role is to provide characteristics of a certain kind. Understanding the behaviour of the system needs little knowledge of the inside of such "black boxes". Similarly when designing a new system, it is first realised as a string of "black boxes" picked from one's catalogue of feasible concepts. (There is a catch, however, for technology is changing so rapidly that there is an ever-increasing and apparently never-ending supply of new functions coming into being. A compromise must, therefore, be drawn between being right up-to-date and actually getting on with building a working system).

When a system fails to operate, the faulty "black box" can be isolated for repair; this is achieved by applying carefully thought-out tests to the system to diagnose the fault, or in the case of small systems, by simply replacing "black boxes" one by one until the system works again. Designing and repairing "black boxes" needs a knowledge of more basic electronic design — we will be mainly concerned with this level in the early stages of the course.

At a stage more basic again, are scientists, research engineers and circuit designers who each have a specialized knowledge of the many individual facets of the basic components — it is they who invent and develop new devices.

To illustrate this hierarchy, consider the system used to transmit visual information to other places — television. In Fig. 2a a television system is depicted as a number of interconnected black boxes. The names in this form of portrayal (called a schematic) tell even the untrained the purpose of each box. The next stage of complexity is another schematic (still drawn as boxes, for we are not yet at the component level) that uses commonly available functions. The camera only, of Fig. 2a is drawn in Fig. 2b to illustrate this — the complete system diagram would need a great deal of space. If we wished actually to construct or fault-find the camera, we would need to know about the inside of each box,

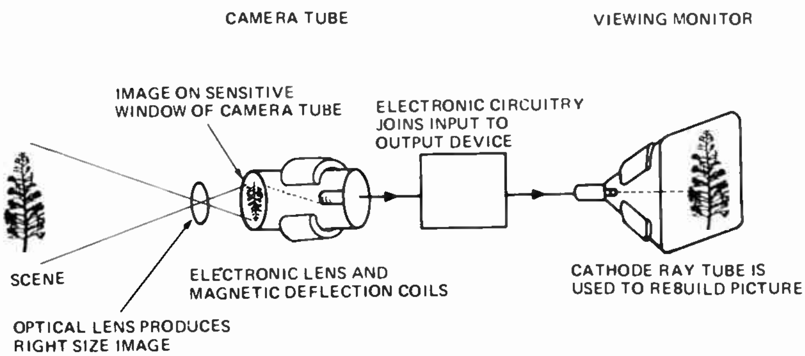


Fig. 3. Electronic circuits invariably need to be used in conjunction with 'bits' from other disciplines. Television, for example, utilises a number of mechanical and optical components — as our drawing shows.

and here we use circuit diagrams that show the actual connections made between components.

In electronic circuit diagrams, symbols are used to represent the various components — thus a battery is shown as —|||— regardless of its actual size or shape, a resistor is usually shown as —ʌʌʌ— , and a capacitor as —||— . Circuit diagrams are in fact a shorthand way of showing components and their interconnections.

The circuit shown in Fig 2c is that of the video signal amplifier (Fig. 2b). Given a circuit diagram and a little basic knowledge, it is relatively easy to assemble the circuit to form a more complicated black box.

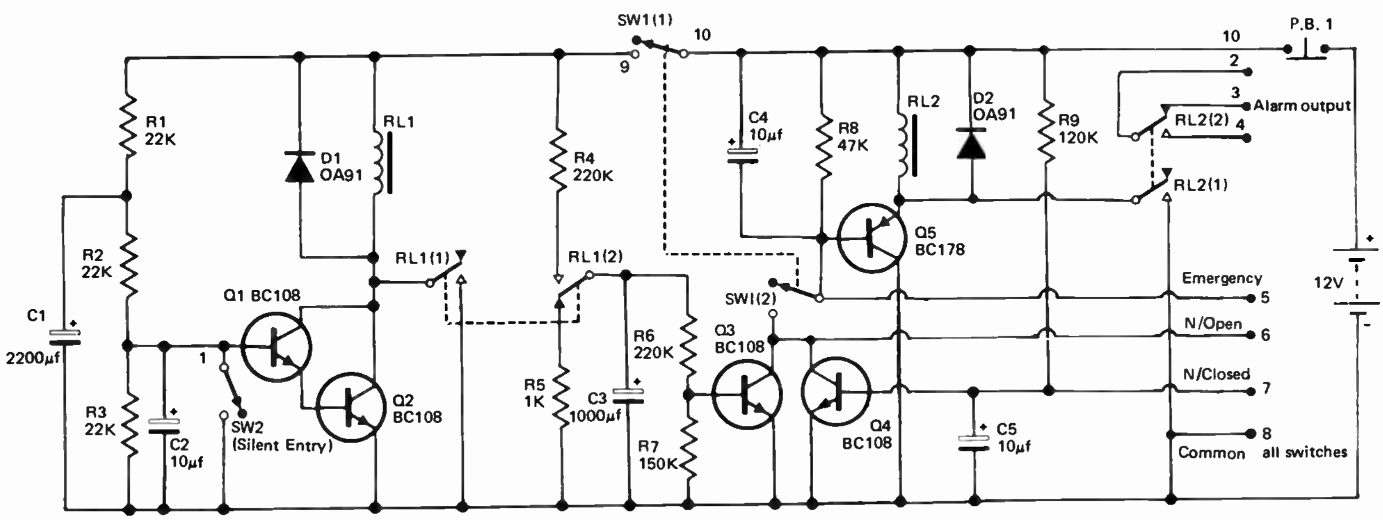
Occasionally, especially when designing new circuits, it helps to have more fundamental details of the operation, manufacturing process and material properties of components, but that stage is not entirely essential if the need is only to make designs already detailed by a designer, in application notes, or an electronic magazine. (A point to remember when studying schematics is that the supply of power necessary to operate the circuit is often omitted to simplify the drawing).

The systems approach to a problem is not restricted to use in electronics alone. It is just as useable in the study and design of mechanical and optical systems, as well as a host of non-physical processes. The electronic worker cannot avoid becoming involved with other disciplines — in the study of television for example, he or she would need to know something of optical techniques, photography and acoustics (as shown in Fig. 3).

When systems are studied as boxes at the various levels described in this brief introductory article, a seemingly incomprehensible device (like that shown in Fig. 4) crumbles, slowly perhaps, but assuredly, to a stage where it is almost obvious — the mystery has vanished. With training and experience, that this course will provide, it will become possible to recognise the individual blocks in an intricate circuit diagram and thus realise its behaviour.

In electronics that which was regarded as a complex system component a decade or so ago, might now be merely a sub-system of another larger system. Twenty years ago it was a major project to design and build a stable amplifier for precision applications. (This is a device commonly used in many branches of electronics. Its function is to enlarge signals, and will be studied later). Today, they are of fingernail size, consume only a minute amount of power, perform equally as well as the best of yesteryear, yet sell at a price that enables them to be used with little regard for their cost. The earlier units used a thousand times more power, cost a hundred times as much, and were at least the size of a shoe box. The old and the new forms are contrasted in Fig. 5.

This trend towards the sale of complete inexpensive sub-systems as the most basic building block enables even the learner of today to build



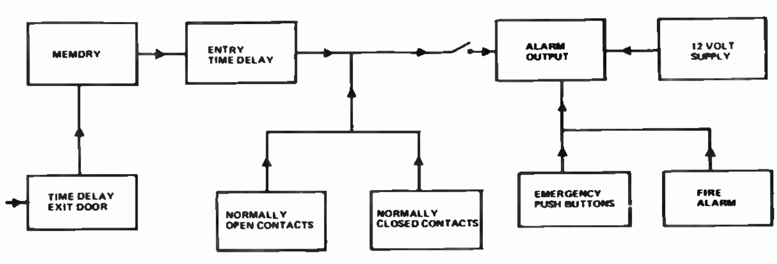


Fig.4. Complex devices may appear baffling until broken down to show the actual functions: a/ Circuit diagram of a complex burglar alarm. b/ Its operation is much clearer as shown in the lower diagram.

sophisticated devices speedily and at reasonable cost. It is because of this development that this particular course is different from most others on electronics.

So much for the way in which we treat a system to gain an understanding of its operations. Let us now concentrate on the fundamental nature of black boxes.

Fig. 5. Dramatically illustrating the rapidly changing nature of electronic techniques, the tiny module has virtually identical performance to the massive unit shown above it. Both are high power audio amplifiers.

Power flow in electronic black-boxes

For a system to operate, it usually must have an energy or power supply. The law of conservation of energy says there must be an energy balance (Fig. 6) — energy given out by a system as useful output, plus the energy losses of components, must equal the input energy from the supply.

Black boxes, therefore, have inputs and outputs of energy. For example, there might be an input of power to drive it and an input to operate the output. The relationship between the input and the output is called the transfer function of the box, for it expresses how the input is effectively transferred to the output. In boxes consuming small powers we refer to the input and output energies as signals. The input signal to a black box invariably controls the power flow to the output — like a tap controls water flow. There are a few electronic systems that derive their power from the input signal, but they are not common. The old-fashioned crystal radio set of our grandfather's day was an example of this. The energy used to drive the headphones was actually derived from the signal transmitted by the broadcasting station, and no battery or other form of power supply was required.

Black boxes connected to the power supply will be constructed from two classes of component. They can either dissipate (or waste) the energy as losses (for example, the heating of a resistor) or they can store energy giving it back later. An example of this — explained in detail later in this course — is a coil of wire forming an inductor. This can store electrical energy by virtue of a magnetic field. Another example is that of two close, but not touching, metal plates (known as a capacitor) that can store energy as an electric charge.

Resistors (often abbreviated when written, to (R)), inductors (L) and capacitors (C) are the basic elements of electronics. In practice, each has some degree of unwanted power-loss or storage and this may be important, as will be seen later. Basic resistors, inductors, or capacitors have a wire

Fig. 6. The total amount of energy flowing into any system will always be the same as the total amount of energy flowing out of that system.

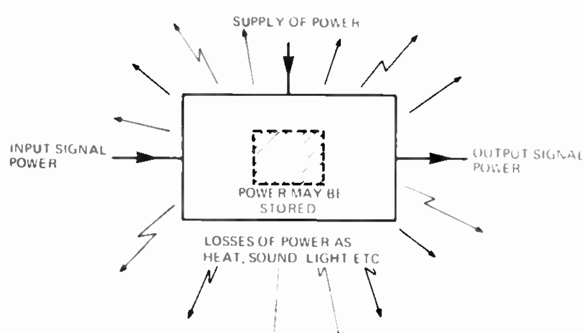


Fig. 7. Signals from many processes may not be in a form suitable for electronic processing. Transducers, are used to make the necessary transformations. Here, information from many different parts of this mammoth steel rolling mill is transduced into electrical form and displayed on this control desk.

going into them, and another leading from them. (They are 'passive elements' in that they are unable to increase (amplify) the power level of input signals transferred to the output. They can be used to set the flow-rate of power but cannot produce a higher power at their output than that at their input.

THE AMPLIFIER

Another class of basic element, the amplifier, by contrast, has three terminals (at least) — input signal, power input and output signal, and with these an output signal can be made much larger than the input signal.

An amplifier does not increase power in a mysterious way. It merely acts as a device whereby a small input power can control a large output power by allowing it to flow (under control) from the power supply — just as a small hydraulic tap is operated to control the lift of a car-hoist in a service station. Such devices are known as 'active elements'. Individual passive elements are often combined to produce a passive circuit; these circuits can then be combined with active elements to form larger circuits.

THE NEED FOR TRANSDUCER BLACK BOXES

Some black boxes serve the purpose of interfacing an electronic system with the physical world, and vice versa. They change (or in electronic parlance, 'transduce') physical variables, such as sound, brightness and length into an electrical equivalent signal that is compatible with electronic techniques.

These are the 'sensors' of man-made systems, acting much as eyes, ears, etc do in other ways in humans.

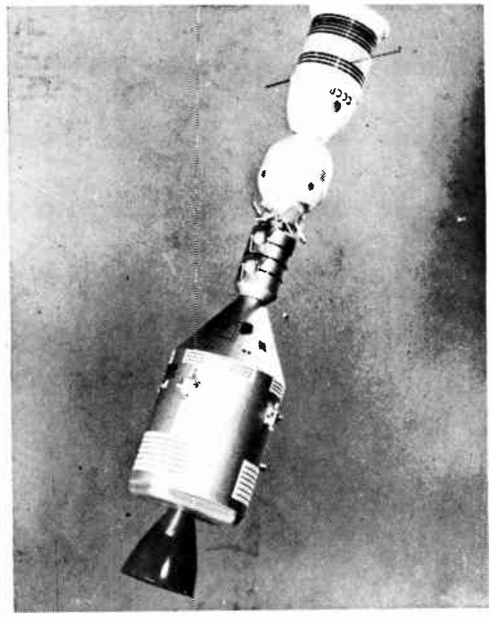
The television camera, for example, changes visual images into electronic signals suitable for broadcasting. Once the electronic signals are processed, it will eventually be necessary to change them back to a non-electronic form (which may, for example, be the output of a record player).

The loudspeaker is one such output transducer, for it converts electric currents to the motion of a diaphragm, thus producing sound pressure waves in the air that we recognise as music or speech.

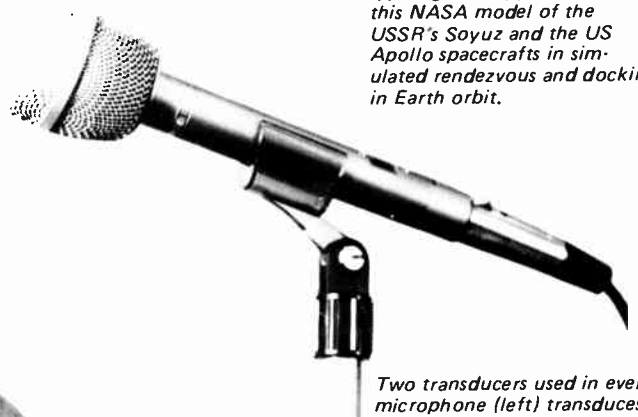
Electronic systems then, are built up from passive elements (resistors, inductors and capacitors in the main) that can either store or lose energy, and active elements (amplifiers) that enable energy flow to be regulated from a main supply. A proper understanding of these basic differences greatly assists comprehension of the operations of circuits that are encountered as we proceed.

As it is too early to start practical work, find out if you really understand the systems approach by sketching the black-box diagrams of common processes around you — they need not be electronic.

Examples worth trying are a motor-car, or bicycle system with a rider, the movie film process from scene to screen, automatic street lighting where the sun is used to switch the lamp off during the day, and traffic lights controlling vehicles at an intersection. Remember to identify where the power is coming from and going to, and which are the active and passive elements of each system.

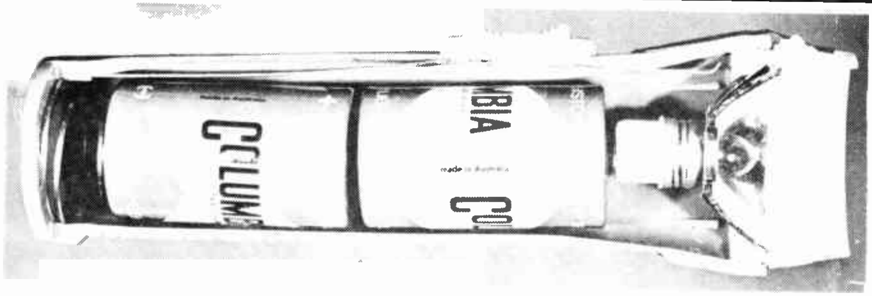


Perhaps the most complex electric systems yet devised are those used in space — typifing such applications is this NASA model of the USSR's Soyuz and the US Apollo spacecrafts in simulated rendezvous and docking in Earth orbit.



Two transducers used in everyday life. The microphone (left) transduces sound energy into electrical energy, conversely the loudspeakers (below) transduces electrical energy back into sound.

2 Basic concepts



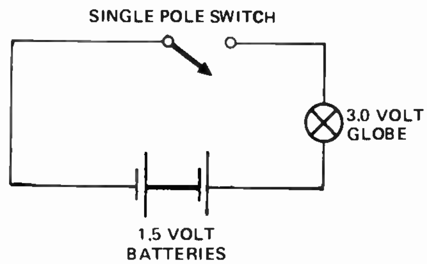


Fig. 1(a) Cutaway view of the common hand torch.

Fig. (b). The same torch in its schematic symbol form.

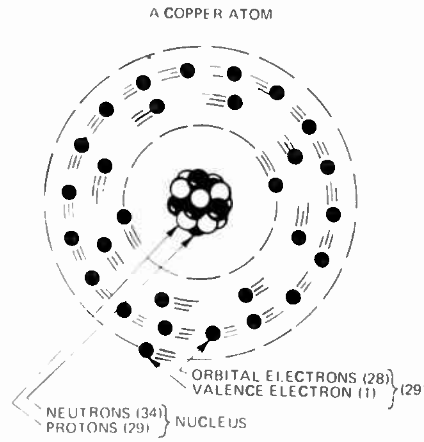


Fig. 2. Representation of a copper atom. Electrons in orbital shells surround a tightly packed central nucleus of neutrons and protons. Neutrons are neutral particles which really are a combination of proton and electron. Together they exhibit neutral charge. The valence electron in the outer shell is the electron which forms molecular bonds and is also the one which may easily be stripped off and become part of an electrical current flow.

TO UNDERSTAND the operation of the basic components used to build system blocks, and to see how they work together to produce many different functions, we must look at electricity at its most basic level. This lays the foundation for understanding how things operate — and for the design of simple devices.

We are going to start by considering a simple torch — basically a light globe energised by power supplied from a couple of batteries.

Figure 1a shows how the torch is constructed.

It is obviously not very convenient to use actual photographs or detailed drawings of all electronic components — hence a system of representative symbols is used instead. Each symbol represents a component or 'black box'.

Figure 1b shows how the torch would be drawn using these symbols. This method of representation is called a circuit drawing and is almost invariably used to depict electronic circuits.

Looking at the torch as a 'system', the batteries supply electrical energy to the torch globe where it is converted to visible light energy. The purpose of the connecting wires and switch is to control *when* and *where* this energy conversion process takes place.

The paragraph above explains the purpose and operational requirements of the torch, but does not explain why and how it operates. To understand this we must look at the internal action of the components themselves — getting right at the structure of matter.

THE STRUCTURE OF MATERIAL

Our physical world is made entirely of chemical elements. There are over a hundred different kinds, but each has a basic similarity.

Each element — no matter what it is — is composed of tiny entities called atoms. These in turn consist of even smaller particles called protons and electrons. It is the number of such protons and electrons, and their orientation with respect to each other, that varies from element to element. (There are also a number of other sub-atomic particles making up the structure of the atom. These include neutrons, mesons, etc. These particles play no part in electronic theory and for this reason will not be discussed).

Each atom has a central, very dense part (called the nucleus) that is made up of one or more protons held together as a single unit. Around this, at great speed, whirl one or more electrons, at a radius considerably larger than that of the nucleus. The mass of the electrons is negligible compared to that of the nucleus. Thus our concept of an atom is one of shells of electrons surrounding a tiny nucleus. Normally there are as many protons as there are electrons — but not always, as we shall see later.

As the electrons whirl — at enormous speed — around the central protons, outward forces are generated that, unless balanced in some way, must inevitably cause the electrons to be hurled from their orbits.

A fundamental property of protons and electrons is that each is physically attracted to the other (whilst electrons and protons each repel their own kind). It is this attractive force between the protons and the whirling electrons that (normally) balances the outward force — thus maintaining a stable situation.

This attractive or repulsive effect is known as 'charge'. By convention, the charge on an electron is called 'negative', and that on a proton is called 'positive'.

The simplest of the elements is hydrogen. This has just one proton in its nucleus, and one electron in orbit. But just where the electron *is* at any time, is impossible to define, for the orbit changes direction continually.

Moving up the periodic table (the classification chart listing the chemical elements in order of number of protons) the combinations become increasingly more complex as the number of protons and electrons increase.

Electrons also exist in more than one shell — following certain basic physical laws. An element having many shells is shown in Fig. 2. Normally the charges balance, giving neutral overall charge, but if as can be done, an electron is *removed*, the atom then has a surplus of positive charge and is called a *positive ion*. If an electron is *gained* it is known as a *negative ion*.

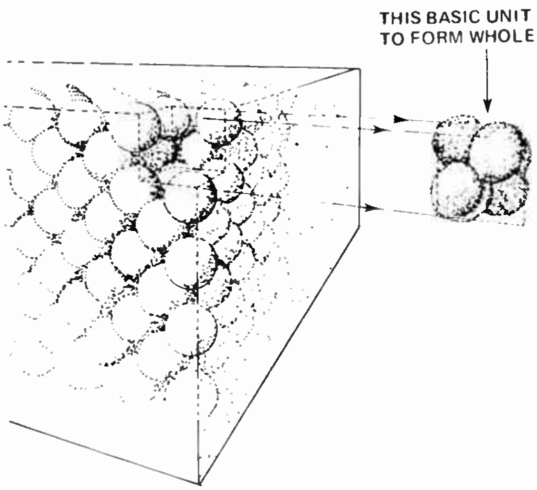


Fig. 3. How atoms join together to form matter.

Like people, atoms rarely exist solely by themselves. They like to form associations with others of the same kind (or other kinds). These combinations of atoms are called molecules and it is large assemblies of molecules, held together by molecular forces, that form the physical matter of the universe. Water, for example, is formed of molecules each consisting of two hydrogen atoms and one oxygen atom.

CONDUCTORS, RESISTORS AND SEMICONDUCTORS

All matter then is made of atoms arranged in a more or less uniform matrix — as shown in Fig. 3. In some materials a few of the electrons, in the outer shells, are not rigidly attached to any particular nucleus. They make what is called a "sea" of electrons, formed by the free electrons, as depicted in Fig. 4.

Materials in which this occurs to a marked extent are called *conductors* of electricity — for the free electrons can be made to flow around a loop of material if a charge unbalance is produced in some way. The wires in the torch, the filament in the globe and the switch connections are *conductors*. In these, electrons flow easily, although, as we see later, less easily in the filament.

In other materials the electrons are tightly bound to the nucleus and there are no free electrons to form the "sea". It is not possible to produce a flow of electrons. These materials are called insulators.

Insulators enable us to isolate electron flow, thus allowing it to occur only when we wish. The coating on the wires and the case of the torch are both insulators. The insulating coating ensures that battery power does not leak away - but is only used to energize the lamp when we want it to.

Air is an excellent insulator, when our torch is switched off, for example, the action of the switch is to separate a pair of contacts. The air gap thus introduced effectively blocks electron flow.

There are no sharp dividing lines between conductors and insulators. At one extreme there are exceptionally good conductors such as silver, copper, and gold. Then there are reasonably good conductors such as steel and brass. At the opposite extreme there are very poor conductors such as rubber, dry wood, plastics, phenolic boards and ceramics. Poor conductors such as these are generally known as insulators.

The filament in the torch globe is a poor conductor of electrons by comparison with the connecting wires and switch contacts. It resists the flow of electrical energy. In so doing, heat is generated — to the extent that if the battery can supply the necessary energy the filament will glow white hot, thus providing light.

There are, as said before, a range of materials whose properties lie somewhere between that of conductors and that of insulators. Some of these materials have other specialised properties — these will be described later in this series.

CURRENT

If two pieces of material, one with an excess of electrons, and one with a deficiency of electrons, are joined by a conductor, electrons will flow from the material with an excess of electrons to the material with a deficiency of electrons until the charge on the two pieces of material has been equalised (Fig. 5).

Material with an excess of electrons is known as 'negatively charged' and conversely, material with a deficiency of electrons is known as 'positively charged'. Thus electron flow is from negative to positive.

Early experimenters in electricity knew nothing of atomic theory and, unfortunately for us, arbitarily agreed to accept a direction of current flow that is in fact precisely opposite to that which takes place. This concept is called conventional flow, whilst the later (and correct) concept is called electron flow.

The flow of electrons along a conductor is called an electric current. It is measured in units called amperes, (amps for short), rather than in actual quantities of electrons, because the number of electrons flowing is enormous — even our humble torch would have well over 10^{18} electrons flowing through its components each second! (One ampere = 6.24×10^{18} electrons/second).

Currents in electronic circuits are usually much smaller than in power

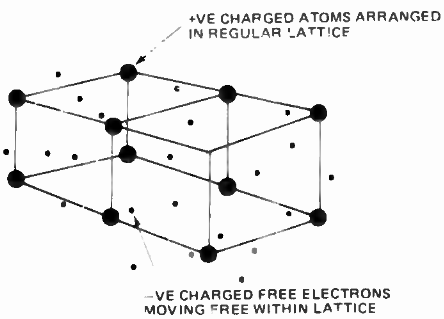


Fig. 4. In a conducting material, numerous electrons are free to move around as portrayed in this simplified picture of a piece of conductor.

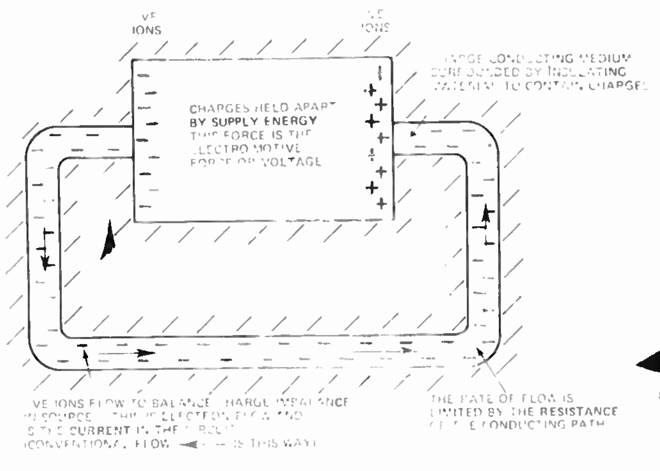


Fig. 5. Current is a flow of charges.

Fig. 7. Typical resistors – the ones on the left are adjustable in value.

equipment. For example the current flowing in a pocket transistor radio is only a few hundredths of an amp. In an electric heating radiator it is several amps. The current picked up by the aerial of a radio receiver is only a few millionths of an amp.

So that they may be expressed more easily, measurement units are, where necessary, prefixed to indicate a larger or smaller value of the base unit. These standard prefixes are shown in Fig. 6.

Suffixes 'k', 'm', 'M' etc *after* component values indicate a numerical multiplier or divider of the basic unit of electrical value.

For example –

Multipliers
k = X 1000
M = X 1000 000
G = X 1000 000 000

Dividers
μ = ÷ 1000 000
n = ÷ 1000 000 000
p = ÷ 1000 000 000 000

Where the numerical value includes a decimal point, the traditional way of showing it was, for example, 4.7 k. Experience showed that printing errors occurred due to accidental marks being mistaken for decimal points. We have used the traditional method throughout this work. However around 1976, there was a strong movement towards an improved method. The Standard now calls for the ex-suffix to be used in place of the decimal point. Thus a 4.7 k resistor is now usually shown as 4k7. A 2.2 μF capacitor is now shown as 2μ2 etc.

Some confusion still exists with capacitor markings. Capacitors used to be marked with multiples or sub-multiples of microfarads – thus 0.001 μF, 470 μF etc. Markings are now generally in sub-multiples of a Farad, using 'thousands'.

Thus –
1 microfarad (1μ) = 1×10^{-6} F
1 nanofarad (1n) = 1×10^{-9} F
1 picofarad (1p) = 1×10^{-12} F

VOLTAGE

Current flow is caused by an imbalance of positive and negative charges. This imbalance may be called 'electron pressure'. The greater the different between the positive and negative charge the greater this electron pressure will be.

The amount of imbalance is called the 'voltage' – the unit of electron pressure being the 'volt'.

Voltage, being akin to pressure, determines the amount of current flow.

Like the unit for current, the volt is also given suitable prefixes to cover the wide variations in magnitude that can occur in electrical phenomena. The voltage level in a radio receiving aerial will be a few microvolts. Lightning strikes may be in megavolts.

In our torch, each battery provides 1.5 volts. When the two batteries are connected together their voltages add to provide 3.0 volts.

PREFIXES

A prefix is attached to a unit to indicate a multiple or sub-multiple of the unit –
milli-ohm is one thousandth of an ohm
kilo-ohm is one thousand ohms

The following table shows the factor by which the prefix multiplies the unit to which it is attached.

PREFIX	SYMBOL	MEANING	FACTOR BY WHICH UNIT IS MULTIPLIED		
*tera	T	one million million	10^{12}	=	1 000 000 000 000
giga	G	one thousand million	10^{9}	=	1 000 000 000
mega	M	one million	10^{6}	=	1 000 000
kilo	k	one thousand	10^{3}	=	1 000
*hecto	h	one hundred	10^{2}	=	100
*deca	da	ten	10	=	10
*deci	d	one tenth	10^{-1}	=	0.1
*centi	c	one hundredth	10^{-2}	=	0.01
milli	m	one thousandth	10^{-3}	=	0.001
micro	μ	one millionth	10^{-6}	=	0.000 001
*nano	n	one thousand millionth	10^{-9}	=	0.000 000 001
*pico	p	one million millionth	10^{-12}	=	0.000 000 000 001

*not generally encountered in electronics

Fig. 6. Prefixes of units used to denote small and large numbers.

An older term — electro-motive force — is still sometimes used for voltage. Yet another term is voltage potential — or just potential.

RESISTANCE

Electrons inevitably collide with atoms as current flows through any material capable of electrical conduction. These collisions impede the flow of current and cause the material to heat up.

Current flow is also affected by the cross-sectional area of the material, for just like water, more current can flow through a large conductor than a small one.

The combination of these effects — that cause a material to impede the flow of current — is known as resistance. Resistance is measured in ohms, a unit often represented by the Greek letter Omega (Ω).

In many instances resistance is an undesired effect, for by impeding the flow of current, and thus causing heat to be generated, energy is wasted. The leads connecting a car battery to the starter motor are, for example, made of large sectional area and kept as short as possible in order to reduce this wastage of energy.

But in other applications — especially in the field of electronics — resistance is deliberately used to control current, voltage — or both.

Resistors, manufactured for use in electronic circuits, have either a fixed or an adjustable value. Their values of resistance may vary from fractions of an ohm to billions of ohms. A number of typical resistors are illustrated in Fig. 7. (The ohmic value of resistors is usually shown in the form of concentric bands of colour — the relevant code is shown in Fig. 8).

OHMS LAW

We have seen that an increase in voltage will cause an increase in current flow, and that increased resistance will cause a decrease in current flow. Ohm, in the 19th century discovered that there is an exact relationship between voltage, current and resistance.

This relationship, which has become known as Ohms Law is perhaps the most basic and certainly one of the most used laws in the whole of electronics and electrical engineering.

From it the designer can determine just how much resistance is needed to limit current to desired values at various voltages, to establish what voltage is needed to perform certain functions, even in the case of our torch, to calculate how much resistance the globe filament must have to be able to produce the desired light.

Ohms law may be expressed as

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

where I = current in amps
V = voltage in volts
R = resistance in ohms

That is, the current flowing in a circuit is linearly related to voltage, and also linearly related to the inverse of the resistance.

Thus if we know any two quantities we may calculate the third, eg, we may find the voltage if current and resistance are known by using simple algebra on the formula. This then becomes:—

$$V = IR$$

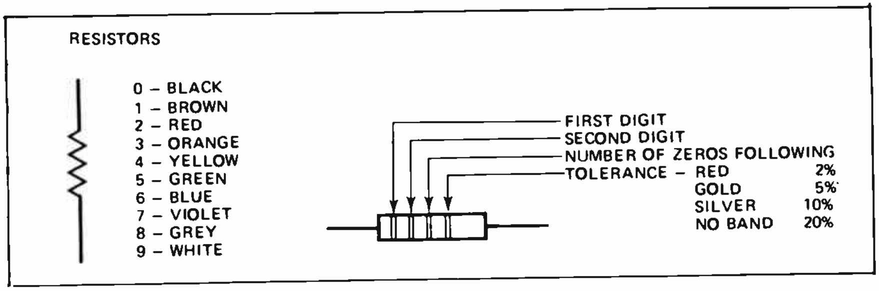
Similarly, resistance may be found by using the form:—

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

For example, the total resistance of a circuit, in which one volt causes one amp to flow, is one ohm. Or if 10 volts is applied across a resistor of 10 ohms — then a current of one amp will flow.

Conductors and insulators are really so-called because of their largely different values of resistance (for materials of the same size and cross-sectional area). There are also certain materials called semiconductors that do not conform to Ohms Law — more about these later.

FIG. 8. KNOWING RESISTOR VALUES



READING THE RESISTOR CODE

Resistors are coded with coloured bands to ease the problem of marking such small components.

The numbers corresponding to the ten colours used and the values per position are shown above.

For example, 180 000 ohms is coded with the first digit brown, then grey and finally yellow. The fourth band indicates the tolerance that the value has with respect to the stated value. For example, silver indicates 10% tolerance meaning the 180 000 ohms could vary between 180 000 ±18 000 i.e. 162 000 to 198 000.

These tolerance may seem to reflect poor manufacture but in most circuits they are, in fact, quite satisfactory. Relaxing the tolerance enables the maker to sell them more cheaply.

PREFERRED VALUES

If the maker tried to produce and sell every value of resistance that exists there would be chaos and the costs would be greatly increased. The actual values made, therefore, are limited to a range called the preferred values. These are listed in the table at right.

The values may seem illogical at first sight, but this is not so. They stem from the fact that the tolerance extremes of a value reach the extremes of adjacent values, thereby covering the whole range without overlap. Values normally available stop in the megohm decade.

Tolerance

±5%	±10%	±20%
1.0	1.0	1.0
1.1		
1.2	1.2	
1.5	1.5	1.5
1.6		
1.8	1.8	
2.0		
2.2	2.2	2.2
2.4		
2.7	2.7	
3.0		
3.3	3.3	3.3
3.6		
3.9	3.9	
4.3		
4.7	4.7	4.7
5.1		
5.6	5.6	
6.2		
6.8	6.8	6.8
7.5		
8.2		
9.1		

COMBINATIONS OF RESISTORS

We have seen that if we know the total resistance in a circuit, together with the applied voltage – then we can calculate the current flowing. Often however we will find that there are a number of resistances connected end-to-end (series), or across each other (parallel) – or even mixtures of the two. In such cases it is necessary to work out the 'effective' resistance of the whole circuit.

Resistors in series have a total resistance equal to the sum of each. For example, in Fig. 9 the total resistance is 2050 ohms. Obviously the total must exceed the value of the largest of the resistors – a good check that a decimal place has not been lost.

In our torch, the total resistance is that of the globe plus interconnecting wires and switch contacts. But here the resistance of the wires and contacts is so small (tiny fractions of an ohm) that they do not significantly affect the total and therefore may be ignored.

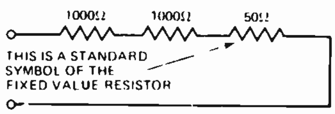


Fig. 9. Resistors connected in series, total resistance is 2050 ohms.

Resistors in parallel are rather more tricky. The rule here is

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

where R total = total resistance where R_1, R_2 etc = individual resistors.

Thus the circuit shown in Fig. 10 has a total resistance of 45.45 ohms. (Note that with resistors in parallel the total must *always* be smaller than the value of the smallest resistor).

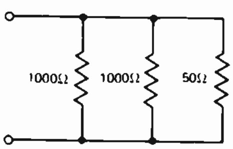


Fig. 10. Resistors connected in parallel. Total resistance is 45.45 ohms.

When a circuit has a combination of series and parallel resistors (Fig. 11 for example), the total resistance can be determined by reducing individual sets in turn. The two parallel 10 ohm resistors reduce to an effective value of 5 ohms. Thus we now have two five ohm resistors in series – equivalent to 10 ohms. Finally we have this effective 10 ohms in parallel with the remaining single five ohms resulting in a final effective value of 3.33 ohms.

POWER RATING OF RESISTORS

When discussing the energy relationships of black boxes we said that some dissipate power – a resistor does this of course, producing heat energy that is lost. It is essential to know just how much heat is dissipated, for overheating could lead to incorrect operation – or even failure if that resistor was not designed to withstand the heat generated.

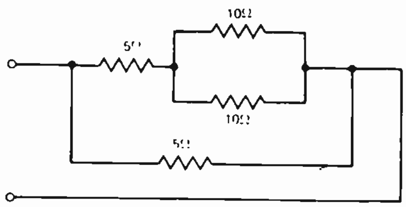
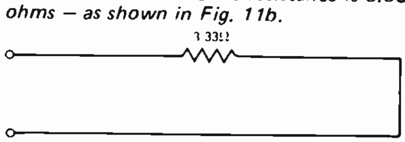


Fig. 11. Combination of series and parallel resistors – their effective resistance is 3.33 ohms – as shown in Fig. 11b.



Power is measured in units called watts. The method of calculating power dissipated in a resistor is W = VI where W equals power in watts
I equals current in amps
V equals voltage in volts

By substituting the appropriate Ohms Law equation for V or I we obtain two other equations. Thus substituting V = IR we get:–

$W = (IR)\ I = I^2 R.$

Similarly substituting $I = \frac{V}{R}$ we get:–

$W = V.V/R = V^2/R.$

Hence if we know any two of the three Ohms Law quantities we can calculate the power dissipated in a resistor as well.

For example, a resistor of 1000 ohms connected to a supply voltage of 10 volts has
$W = 10^2/1000 = 0.1W$ (or 100 mW).

Resistors come in various wattage ratings, either as fixed values or variable for use where the value needs to be adjusted. Ratings range from 1/8W to 25 W and more. The majority of circuitry uses 1/4 and 1/2 W resistors. A point worth remembering is that the rating quoted is the maximum that the manufacturer recommends. A resistor run at that rating gets quite hot. It is good design to use them to only half the rated value.

POTENTIOMETERS

Variable resistors, which are called potentiometers are made in many different forms for a vast number of applications. Factors which affect design are, for example, wattage and resistance range, whether the control is to be continually or intermittently adjustable, the 'law' of resistance and the variety of mechanical arrangements required.

The maximum power which may be dissipated is specified for the condition where a voltage is applied across the end terminals continuously. However under certain conditions, power dissipations much lower than this can cause damage. This is because the resistive element also has a maximum *current* limitation. Thus a potentiometer set to its lowest resistance could be damaged if excessive current were drawn via the slider terminal.

Potentiometers are constructed with various relationships of resistance versus rotation. Those most commonly used are:–

A linear
B logarithmic
C reversed logarithmic
D tapped
E balance

The type of relationship used is indicated by stamping the appropriate letter symbol on the body of the potentiometer, eg. 10 kC is a 10 k ohm reversed logarithmic potentiometer.

Logarithmic potentiometers are used so that an apparent increase in loudness from, say, a volume control occurs with a proportional rotation of the spindle.

ELECTRONICS - in practice

The best way to learn practical skills is to be actually involved with the hardware. Now is the time to start building circuits in order to learn how to solder properly and to become familiar with components.

If you do not already have them you should purchase a small set of the essential handtools. You will need a lightweight soldering iron, a pair of side cutters, long nose pliers, screw-drivers and preferably an electric or hand drill complete with drill bits.

In this first exercise we will use resistors in conjunction with a device called a relay. The complete circuit is given in Fig. 12 and a list of parts needed for the exercises is given to assist you in purchasing the necessary components.

As this is the first encounter with practical electronics, the procedure will be carefully detailed.

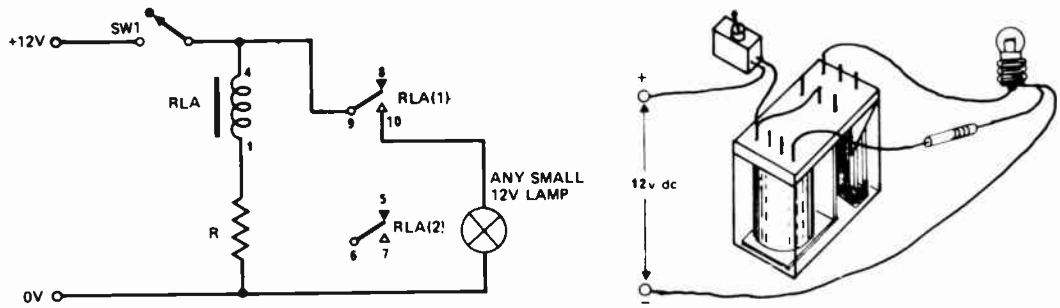


Fig. 12. On the left is the basic relay circuit. The sketch on the right shows the actual component connections for the circuit.

The 12 V supply can be obtained by using eight 1½ volt torch-cells placed positive to negative to obtain 12 V. Alternatively, a 12 V car or motorcycle battery, or a model train supply, will do in this instance.

The relay is a switching device in which the current flowing in the coil magnetizes a soft iron core which pulls down an armature. The armature in turn mechanically actuates a set of contacts. In the relay specified, the operation of the armature and contacts can be seen through the plastic cover. The relay has two sets of contacts. We will use this component to build many simple and interesting devices.

A switch between the circuit and the battery supply is desirable but not essential for operation.

Having now determined the characteristics of our relay we can use another interesting device, the light dependent resistor (LDR), to turn the relay on whenever light falling on the LDR exceeds a certain level.

This device may be wired in place of 'R' in Fig. 12.

Unlike a normal resistor the LDR changes its resistance in accordance with the light intensity falling on its grid-like structure (see Fig. 13a). The relationship between light level and resistance is non-linear and is best expressed by means of a graph, as in Fig. 13b. Such a graph, as it tells us the characteristics of a particular device, is called a *Characteristic Curve*.

We previously determined the resistance necessary to just operate the relay from a 12 volt supply. By finding this value of resistance on the graph we can look across and down, to find the corresponding light intensity necessary to the relay when using an LDR type ORP 12.

HOW TO SOLDER

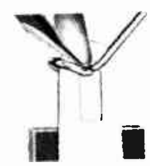
Good soldering is most important — many of the problems that beginners have with their first projects are due to poor joints. The following hints will aid you to become adept at soldering.

1. Purchase a good quality iron with a wattage rating between 15 and 25 watts.
2. Use only resin-cored solder (60/40 tin-lead content). Do not use acid flux.
3. A new, or worn, iron will need tinning. To do this let the iron get quite hot and file the tip smooth to expose fresh clean copper. Quickly, before the copper has time to discolour, apply resin — cored solder — it should flow all over the tip forming a shiny coating.
4. Keep your soldering iron clean. Wipe it frequently with a damp cloth or sponge.
5. Make sure the connection to be soldered is clean. Wax, frayed insulation, and other foreign substances will result in inferior joints.
6. With older components, or copper wire, it will be necessary to clean and tin the individual components before soldering them together (see 3 above).
7. Attach the wires to be soldered. Do not make more than a half turn in a lead to be soldered — twisting makes subsequent removal difficult.
8. Heat the connection with the iron and apply the solder to the connection. Do not melt solder on the iron and carry it to the joint.
9. Keep the iron on the joint until the solder just commences to flow on the connection. Too little heat results in a high-resistance joint (known as a dry joint) too much causes component damage and evaporates the tin component again causing a poor joint. This step requires practice.
10. Let the solder harden before moving the connection. Then check for a smooth bright joint. A joint that has been moved will have a crystalline appearance, may have a high resistance and will fracture easily.

Good soldering is a matter of practice. If you follow the above hints, it will be only a matter of time till you are making professional joints.

Poor connections look crystalline and grainy, or the solder tends to blob.

Attach the wire.



Heat both the wire and the connection point.

Apply solder to both the tip and the connection.

Let the connection harden before moving the wire. Then check for a smooth, bright joint.

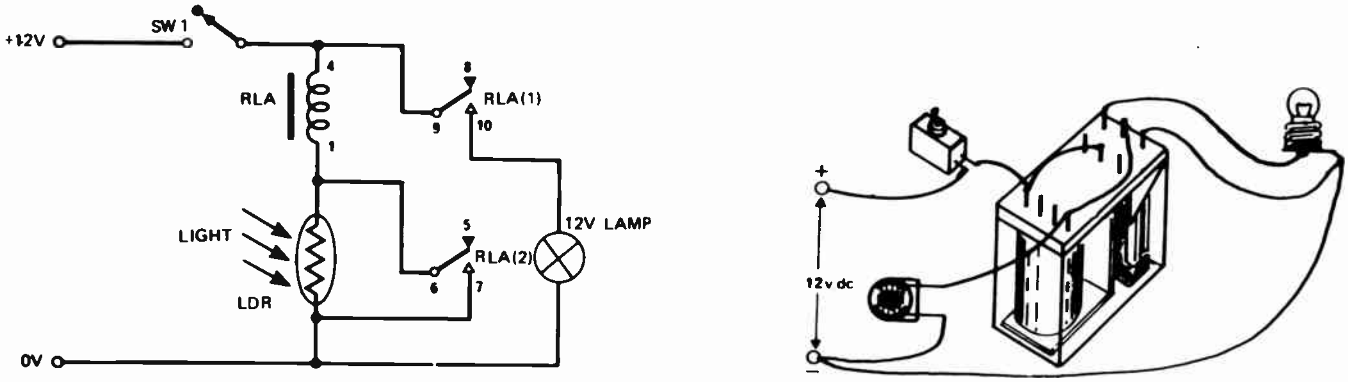


Fig. 14. Replacing resistor R with an LDR (see text), will allow the relay to be operated by a light beam. Relay contacts 6 and 7 are used to hold the relay in once it is operated. Omit connections to these contacts if hold operation is not required.

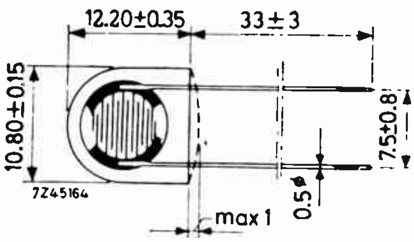


Fig. 13(a). Light dependent resistor (type ORP 12)

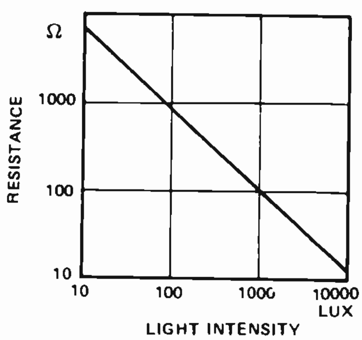


Fig. 13b. Characteristic curve of ORP-12 light dependent resistor.

PARTS LIST

1 – 12 V battery or supply and connector.
1 – 12 V lamp (100 mA to 320 mA) and holder.
1 – Miniature relay, Davall 21/2CA, 185 Ω, two change-over contacts.
1 – Relay holder socket.
1 – Single pole switch.
10 – ½ W resistors, assorted say 47, 100, 470, 1.0k, 1.5k. 4.7k. 10k, 100k, 1M.
1 – ORP 12 light dependent resistor (LDR). Hookup wire of assorted colours.
1 – Mounting board 24 holes by 75 mm.

A well lit room is generally around 100 lux and bright sunlight up to 8000 lux or more. The intensity must be higher than the amount determined from the graph for reliable operation of the relay.

Just as for an ordinary resistor, the power dissipated in the LDR may be calculated. Note however that this must be done for each light level, and that the power dissipation in the LDR rises rapidly as the light intensity increases.

The coil of the relay specified has a resistance of 185Ω so, if you calculate the current flowing when operated from 12 V, you should get 65 mA. Wire the circuit without the resistor R, and check that the lamp lights when the switch SW1 is closed.

Next, place the resistors, one by one, in series with the relay coil and find the value that just enables the coil to pull in contacts, turning on the lamp. In each case calculate the current flowing through the relay, and the power-rating needed of the resistor. You should be able to do this by referring to the theory given. Having found the value of R, calculate the voltage that just operates the relay (having found the current flowing and knowing the coil resistance, it is a case of applying Ohms Law). Knowing the voltage required, and the current, calculate the power required to operate the relay. Confirm your results by then using the coil resistance rather than the current.

Understanding Ohms Law is absolutely essential and tests such as the above will help provide this understanding.

CIRCUIT MOUNTS

There are many ways to construct circuits. The simplest is to use brass or copper plated nails placed in a piece of dry wood. Although this will suffice, it is better in the long run to use a board made specially for the purpose. As the course proceeds we will discuss various other means.

A good start is to use ready drilled matrix board with hole spacings at 0.25 inches. There are other boards with holes at 0.1 inch centres, but the extra space afforded by 0.25 inch centres makes it easier for beginners. Pins are sold that push into the holes ready to take the components.

The spare contacts on the relay can be used to hold the relay on once it has been operated, even though the light has been removed. This method of operation is called latching. The circuit for this is given in Fig. 14. The normally-opened contacts close when the relay operates, shorting out the LDR. To release the relay operate SW1.

Do not cut the leads of components unless the circuit is to be permanent. Also, do not solder too close to the component body as the heat can damage it.

No doubt you can think of many uses for this light sensitive relay. The contacts can switch currents up to 1 ampere at a voltage up to 100 V. *Do not use it to switch circuits connected to the 240 V mains: it would be lethal.*

3 Meters and measurements

WE ARE unable to see electricity directly. But if electrical energy is converted to some other form, then a measurement of its magnitude can be made visible.

For example, the brightness of a light globe is a measure of size of one of the electrical quantities. Such a method is, however, not nearly accurate enough as the eye is a very poor judge of light levels.

A much better way is to use the electricity to drive a mechanical pointer across a scale. The extent of movement indicates the size of the quantity we are measuring. Devices using this principle are called 'meters'.

Meters come in numerous sizes, shapes and types (as shown in Fig. 1) but detailed differences need not concern us as yet for each one of them can be represented by our "black box" concept — an electrical parameter feeds into the box producing an observable pointer movement, proportional to the voltage, current, resistance or power (etc) being measured.

The type of meter known as 'moving coil' is the most suitable (and hence most commonly used) meter for measurements in dc circuits.

A detailed description of the moving coil meter is given elsewhere in this article for those who wish to know more about its construction. However, such detailed knowledge is not necessary in order to make measurements with it, we can consider the meter purely as a black box device having two basic characteristics.

Firstly, it may be considered as having a fixed resistance between its two terminals. Secondly, it requires a certain amount of current to deflect the meter pointer to end-of-scale.

Meters of this type are described in terms of the amount of current required to deflect the meter pointer full scale. (This end-of-scale position is known as — full scale deflection — often abbreviated to 'fsd').

Thus a meter, described as being 0-1 mA, indicates current values between 0 and 1 mA, and requires one milliamp of dc current to deflect the pointer to the end-of-scale (fsd) position.

At this stage, we should point out that the term 'meter' is loosely applied both to the basic *meter movements* (as in Fig. 1) as well as to complete instruments that incorporate switching and other electronics These additions to the basic meter movement are required to measure the different kinds and quantities of the basic variables in circuits, ie, voltage, current and resistance. Each can be measured, using exactly the same meter movements, with the aid of external resistors and appropriate connections.

How the same meter movement can be used to measure the different quantities is easily explained by using Ohms law and the rules of series and parallel connected resistors. We warned you that Ohm's law was basic! All electronic theory is built up logically piece-by-piece. So we reiterate — make sure you consolidate each piece of knowledge as we progress.

MEASUREMENT OF VOLTAGE

In our black-box representation of the meter, we stated that it could be regarded as a resistor. Hence a voltage is required, dependant on the movement resistance, to drive a current through the meter. This may be calculated by our, by now, well known formula V=IR.

Generally, meter movements are sold with the full scale deflection current specified. Some are sold as ready-to-use, specific range units modified to read 0 — 1 A, or 0 — 10 A, for example. These have built-in resistors to modify the basic movement characteristics but for the moment we will consider only basic unmodified units.

Unfortunately it is not common retail practice to quote the resistance of the meter movement. Manufacturers sometimes provide it on the actual unit itself — perhaps written on the meter face below the needle aperture, or in data sheets. (If not given, it can be measured, but as it requires a second accurate voltmeter this will be beyond the beginner).

If, as shown in Fig. 2 for example, the internal resistance of the meter movement is 1 kΩ (1000 ohms), and it requires a current of 100 uA to deflect it full scale (a common specification) it will be fully deflected by applying a voltage of

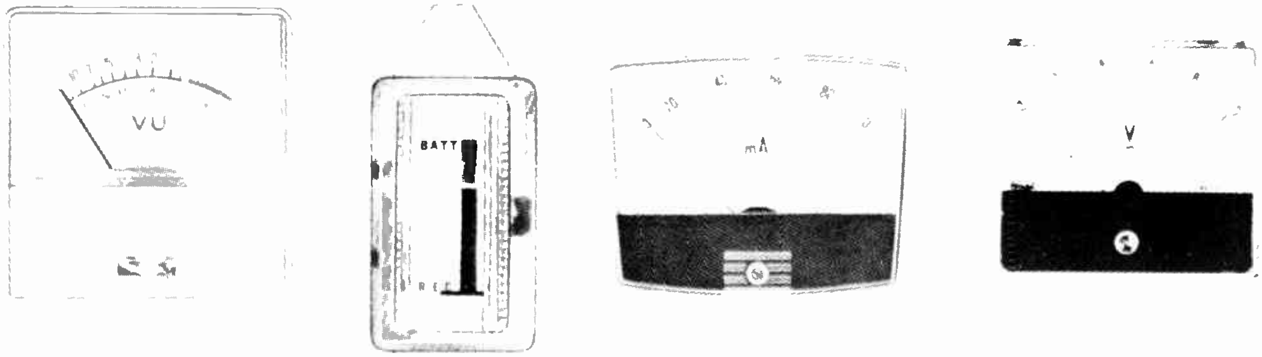
$$V = IR = \frac{100}{1000{,}000} \times 1000 = 0.1 \text{ volts, or } 100 \text{ mV.}$$

So, although described as an amp-meter (we usually say ammeter) it may also be used as a voltmeter having 100 mV fsd.

Obviously the direct application of a voltage larger than 100 mV will cause the needle to deflect past the full-scale value. Excessive deflection can easily damage the movement. The excessive voltage may also cause more current to flow than the movement can handle — remember power is dissipated ($P = I^2R$) so unless adequately cooled, the coil in the meter movement may become too hot and burn out.

So to make our basic movement read higher voltages we add a series resistor, as shown in Fig. 3a. This resistor is called the *multiplier*. Its purpose is to limit the current, at the chosen voltage, to the fsd value of the meter.

For example, if when using the circuit of Fig. 3a, we require a 1 V fsd scale, we may calculate that the total series resistance (from R = V/I must be



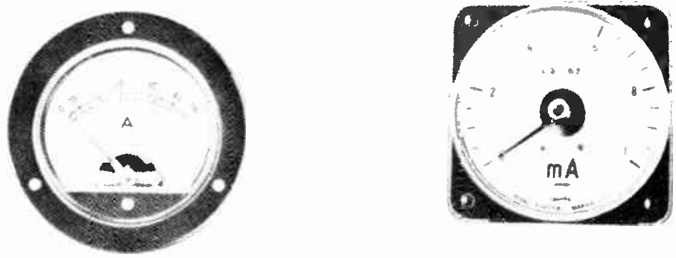


Fig. 1. Basic meter movements have many shapes, sizes and scales to suit different applications.

$$R_{total} = \frac{1\ V}{100 \times 10^{-6}\ A} = 10^4\ \Omega \text{ or } 10\ k\ \Omega$$

Remember the meter movement already contributes 1 kΩ so an additional multiplier resistance of 9 kΩ is needed. By a similar process the multiplier value for any voltage above 100 mV may be found. The upper limit is set by the insulation of the meter, but 1000 V full scale is usually possible.

Let us now see what happens if we require 100 V fsd. We would need a total resistance of $10^6 \Omega$ (1 MΩ). But now it is pointless to subtract the meter movement's resistance value for it is insignificant compared with the multiplier value (0.1 per cent). Note also that for every volt of fsd range required, we need to add approximately 10 k in series (100 V – 1000 kΩ; 10V – 100 kΩ; 1 V – 10 kΩ). We can, therefore, refer to the meter circuit in terms of this – saying it has a sensitivity of 10 kΩ/V. Meters are usually specified this way – not to help us design multiplier values but to enable the effective resistance of the modified meter to be quickly assessed. This is most important when connecting the meter into electrical circuits, as we shall see later.

Voltages up to 100 may now be read by connecting the leads of the meter and its series multiplier across the two points in the circuit between which we require to know the voltage difference. There is no need to break any connections, so the procedure is very simple.

Make sure if you use such a meter in a permanent connection, that the metered voltage *never* exceeds the rating of the unit.

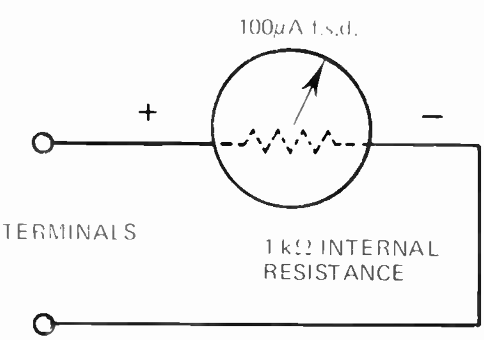


Fig. 2. Regardless of shape or size every meter movement may be represented as a 'black box' having two terminals and an internal resistance. The current required to produce full scale deflection must also be given.

THE MULTIMETER

If we had a meter that was calibrated to read 1000 volts full scale, but were trying to read only three volts, the pointer deflection would be so small that we could not know with certainty whether the reading was two, three, or four volts. Obviously a more sensitive meter is required. Rather than unsolder our multiplier and instal a new one, it is much better to fit a switch that enables us to select a multiplier having a resistance appropriate to the magnitude of the voltage being measured. This arrangement is illustrated in Fig. 3. Such a modification produces a slightly more complex black box – the multi-range voltmeter. We will see below that various ranges of current and resistance may similarly be handled with switched ranges. When all these facilities are built into a multi-range, multi-function unit, it is commonly referred to as a multimeter.

MEASURING CURRENTS

Let us now examine how our 1 kΩ, 100 µA fsd meter movement may be modified to measure larger values of current. Again it is done with resistors, but this time the resistor is placed across the meter and is thus in parallel with the resistance of the movement instead of in *series* as before. This parallel resistance is termed a 'shunt' as its purpose is to by-pass or shunt current around the meter, as illustrated in Fig. 4a.

From Ohms law (again!) we can see that when resistors are connected in parallel, the larger value carries less current than the lower value of resistance. As long as the resistances remain constant so does the ratio of currents.

For example, suppose we need the basic meter to read 300 µA fsd instead of 100µA. As the meter deflects full-scale with only 100µA, 200µA must be diverted by the shunt. Remember that the voltage across each resistor in a parallel arrangement is the same, so our shunt must be of such resistance that 200 µA passes for 100 mV of applied voltage (remember the meter movement is also 100 mV fsd).

The shunt must be

$$R_{Shunt} = \frac{100 \times 10^{-3}\ V}{200 \times 10^{-6}\ A} = 500\Omega$$

Another way to look at it (provided the current range required is much higher than meter movement's normal

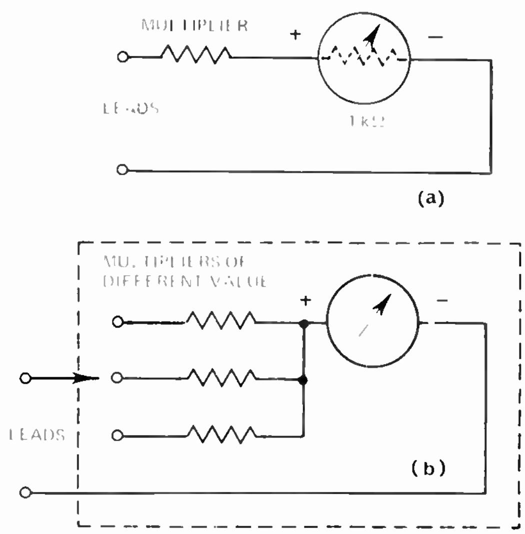
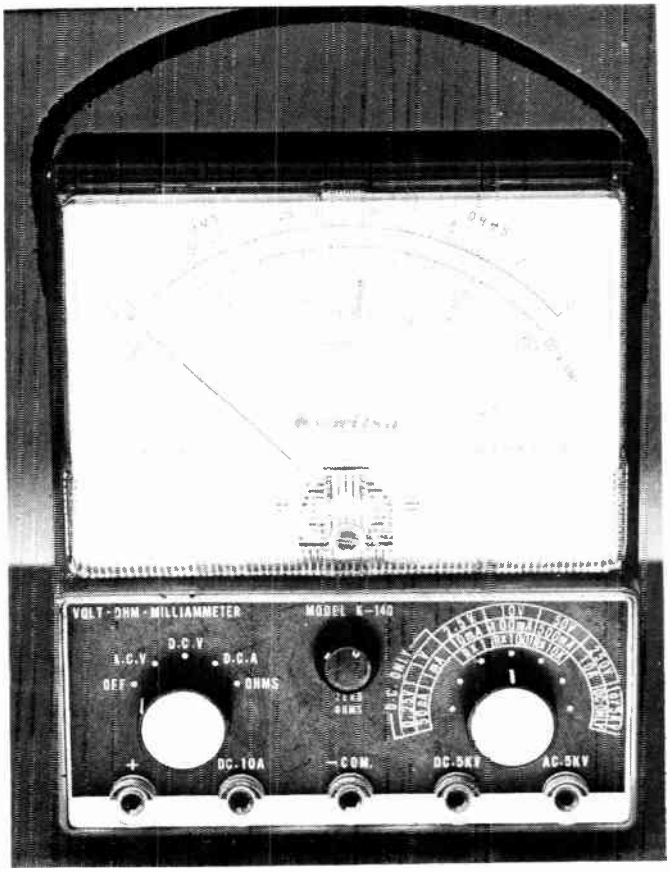


Fig. 3(a). Using a 'multiplier' resistor to provide a single voltage measurement range. (b) Adding a switch and extra multiplier provides more voltage ranges.



Typical high quality multimeter, note the 'ohms' scale in which zero is at right hand end. Meters that indicate the measurement by a scale and moving pointer are known generically as 'analogue' instruments.

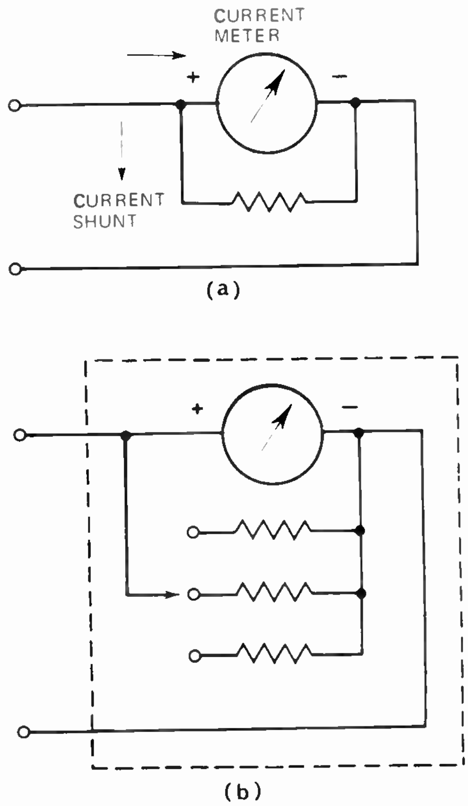


Fig. 4(a). In order to measure currents greater than the meter movement fsd a 'shunt' must be used to divert part of the current around the meter. (b) A switch and extra shunts may be used to obtain more than one range.

current) is to consider what value of resistance would develop 100 mV across it when the desired current flows. For example, 1 A fsd needs a shunt of

$$R_{Shunt} = \frac{V}{I} = \frac{100 \times 10^{-3}}{1} = 0.1\Omega$$

When designing shunts for large currents, thought must be given to the power dissipated in the shunt. In our previous example the power dissipated $P = I^2R = 1.0^2 \times 0.1 = 100$ mW which is not significant. However if we were to require a shunt for 20 amps its resistance would be

$$R_{Shunt} = \frac{V}{I} = \frac{100\text{ mV}}{10\text{ amps}} = .01\Omega$$

Power dissipated $P = I^2R$

$= 10^2 \times 0.01 = 1$ watt.

If our shunt is not to run too hot the resistor must be rated at two or three watts.

In practice the very low resistance required is usually obtained by using a short length of heavy-gauge wire and power dissipation is not usually a problem.

Shunts are usually made specially for meters, as there is little or no call for such low resistance values in the majority of electronic circuitry. Because of this they tend to be more expensive than normal resistors.

Measurement of current is made by breaking the circuit, in which it is required to find the current flow, and wiring the shunted meter in series with the lead and its original connection point so that the normal current passes through the meter and shunt.

We have seen how a shunt lowers the total effective meter resistance — especially as the current range rises. For this reason it is quite wrong to connect a multimeter set to a current range across a component or section of a circuit. It will very effectively short the circuit out causing heavy currents to flow in components probably not able to support them. For example, our above mentioned 1 A meter (which has a 0.1 ohm shunt) placed across points in a circuit between which a voltage difference of 12 V exists, will cause a current to flow of

$$I_{flowing} = \frac{12}{0.1} = 120\text{A}$$

This current, if the supply could provide it, would melt the leads and components instantly!

To create a multi-range ammeter we again use a switch (Fig. 4b) to select the required shunt across the meter. As the same current flows through the switch contacts, the upper current range is limited mainly by the size of

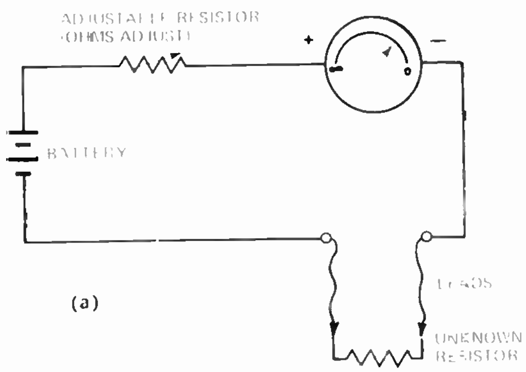
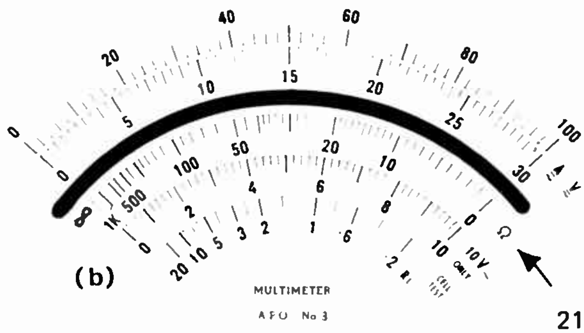


Fig. 5(a). The circuit required to measure ohms. (b). A typical multimeter scale — The Ohms scale is the centre one.



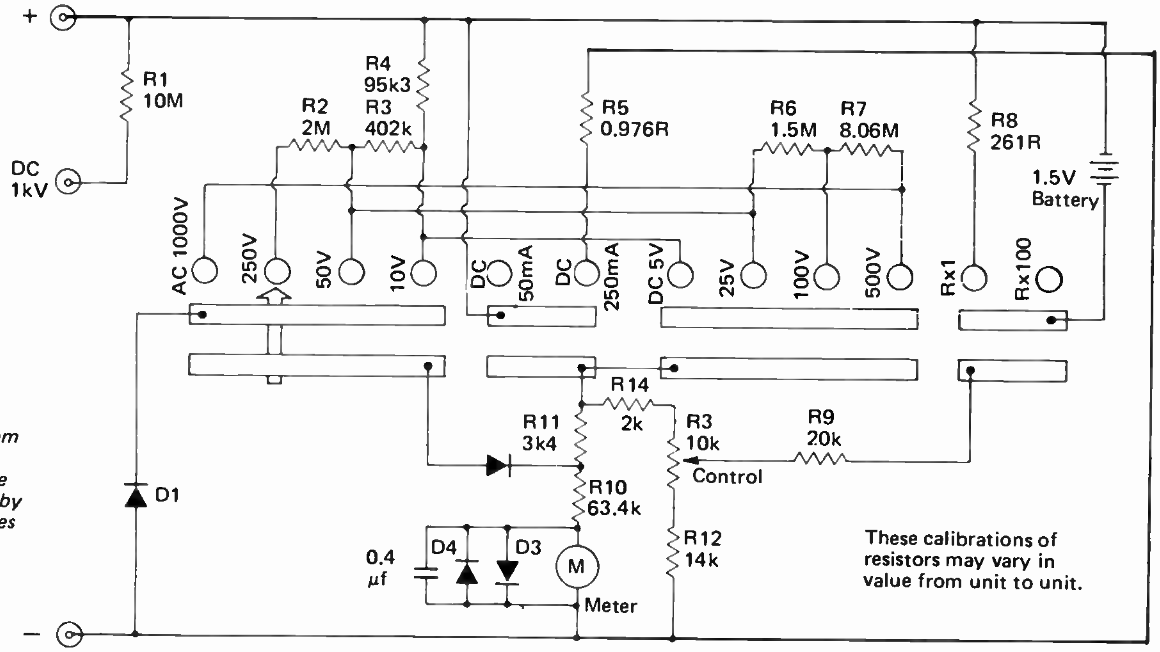


Fig. 6. Circuit diagram of a typical multimeter (JM-20K). The shorting bar, shown by the broad arrow slides along to make three connections at each position.

SCHEMATIC DIAGRAM

> **AC AND DC**
>
> In all our discussion so far, we have considered electrical current flow as being in one direction only. This kind of current flow (as from a battery) is known as "direct current" and we usually abbreviate this to "dc".
>
> There is another kind of current flow which continually changes direction, flowing first one way and then the other. This is known as "alternating current" or "ac". The mains supply used in your home is a 240 volt "ac" supply, and the direction of current flow reverses 50 times a second. We will study ac in more detail later on, for the moment, just keep in mind the distinction between ac and dc.

the contacts and to a lesser extent by the size of the shunts. For this reason multimeters rarely have a range higher than 10 A fsd.

MEASURING RESISTANCE

As well as having voltage and current ranges, multimeters are usually capable of measuring resistance.

As described above, the multiplier of the voltmeter arrangement alters the effective fsd value of the meter to suit various applied voltages. But if we provide a fixed voltage (from a battery) to the meter, and place the unknown resistor in series, we can work backwards from the indicated current to obtain the resistance value. This may be explained with the aid of Fig. 5a as follows:

If the meter leads are shorted together, there is virtually zero resistance between them. This method is in fact used to establish the zero ohms point on the scale.

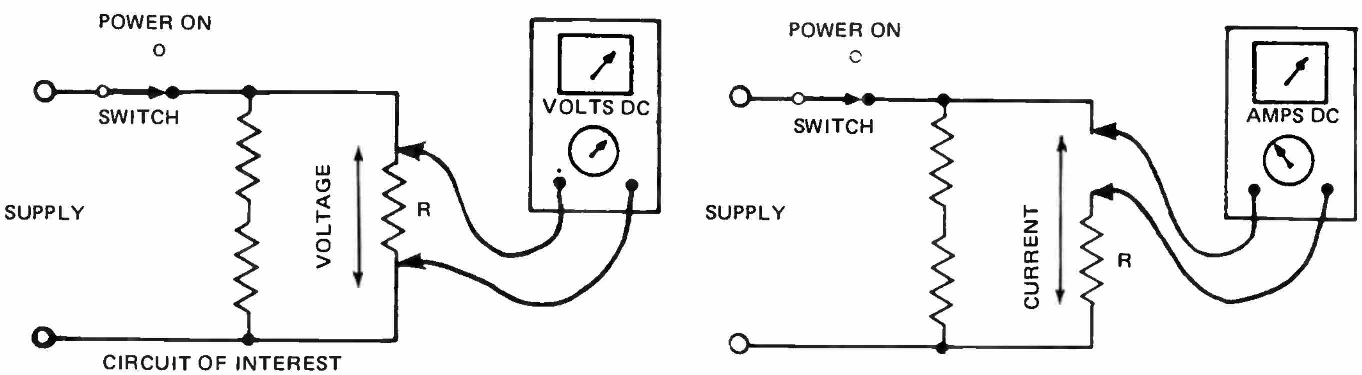
In practice the leads are shorted together and an in-built variable resistance is adjusted to give full scale meter indication (i.e. 100µA is flowing). Unlike the other meter ranges, on the resistance range full scale deflection represents zero.

At the other extreme, if the meter leads are not connected to anything at all, the resistance between them is virtually infinite. Hence no current can flow and the meter pointer will not move from the left-hand end of the scale. This point thus represents infinite resistance and is marked ∞ accordingly.

In between the 0 and ∞ values the pointer will assume a position proportional to the external resistance added. By applying Ohms Law it is easy to work out the position on the scale for any other value. A typical ohmeter scale is illustrated in Fig. 5b.

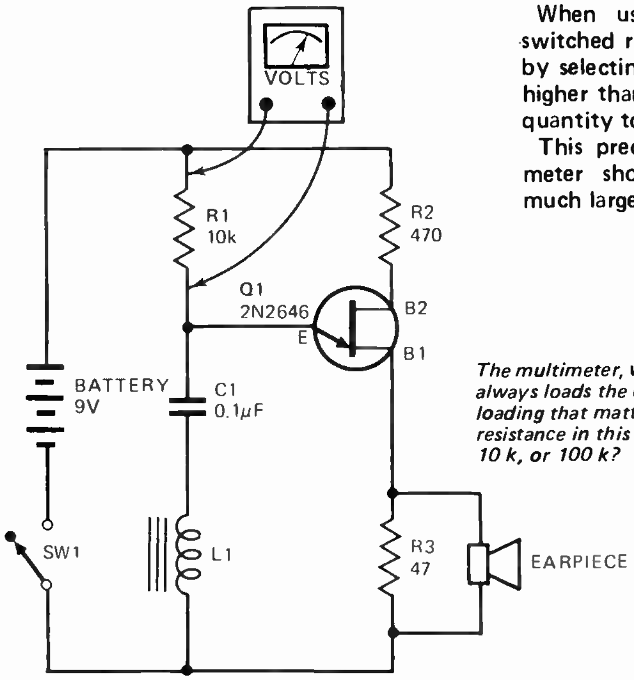
For example, if we use a 1.5 V torch-cell as the voltage source, the adjustable multiplier needed with the

Fig. 7. These diagrams illustrate the three basic measurement techniques for (a) voltage, (b) current and (c) resistance.



(a) MEASURING VOLTAGE ACROSS R

(b) MEASURING CURRENT FLOWING IN R



1 kΩ, 100 μA fsd meter, must be:—

$$R_{total} = \frac{1.5\ V}{100\ 10^{-6} A} = 15\ k\Omega$$

from which the value of the meter (1 kΩ) must be subtracted to get 14 kΩ With shorted leads the meter will go to fsd, now marked 0Ω. The multiplier is adjustable so that the zero point can be reset as the battery becomes discharged (lower voltage). If we now insert 10 kΩ externally we get a total resistance of 25 kΩ. The battery voltage will cause a current to flow in the series loop (remember in a series circuit all components pass the same current) that is

$$I = \frac{1.5}{25 \times 10^3} = 60\mu A$$

Hence, the meter, with 10 kΩ in series with its leads indicates its 60 μA value but is now marked 10 kΩ on our scale. The process is repeated for decade values — 100 Ω, 1 kΩ 100 kΩ, then for spaces in between until the scale is adequately filled. Note that the divisions are not evenly spaced and that this form of ohmeter has better resolution on some parts of the scale than on others. It is also clear that it measures 'back to front'. This is somewhat inconvenient but the simplicity of the arrangement more than compensates for this anomaly. More advanced (and therefore more expensive) multimeters usually read the conventional way.

When using any meter with switched ranges, always start off by selecting a meter range much higher than your estimate of the quantity to be measured.

This precaution safeguards the meter should the quantity be much larger than expected.

The multimeter, when measuring the voltage, always loads the circuit. It is the degree of loading that matters. Calculate the meter resistance in this case — should it be 1 k, 10 k, or 100 k?

It should be noticed that, in the example given, 10 kΩ gives approximately mid-scale deflection but, because of scale non-linearity, it is difficult to read values, around 100 kΩ at all, let alone with any accuracy.

To resolve this difficulty, various battery voltages and/or resistance multipliers may be used to obtain centre scale readings either larger or smaller than in the example given. Use ohms law to determine what multiplier and battery voltage is required to obtain a mid-scale reading of 100 kΩ.

When it is required to measure the resistance of a component in a circuit, it must be disconnected. (One side only will do, but it is often easier to remove it completely).

To make a reading the meter leads are first shorted together and the zero ohms adjuster set to obtain zero ohms reading (this compensates for battery voltage variations). The unknown resistance is then connected between the meter leads and its value read from the meter.

The measurement functions, namely, voltage, current and resistance form the basis of the multimeter. Other ranges may be provided for ac (or alternating current) quantities. These will be discussed later in this series. Figure 6 shows the full circuit for a small meter — can you trace out the various circuits that we have described? Ignore the ac slider position for the moment.

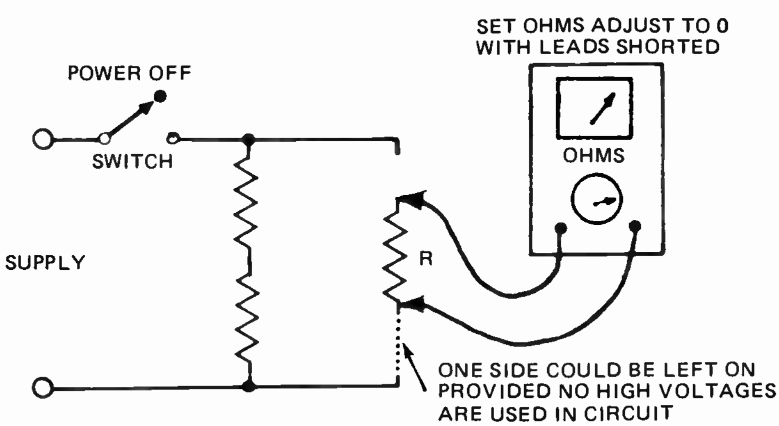
LOADING THE CIRCUIT

As we have seen, connecting a resistor in parallel with another resistor reduces the effective value of both — to something lower than the value of either. Thus, as a voltmeter is in effect a resistor, connecting it across a circuit will inevitably change the resistance of that circuit. In effect one has paralleled one resistor with another, and the meter must shunt current away from the circuit.

Because of this, when a meter is connected across a circuit, the operation of that circuit may well be affected, thus causing the meter reading to be in error.

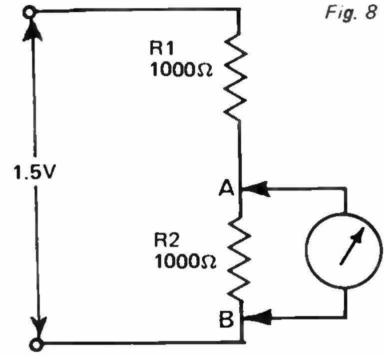
This brings us back to the reason for quoting the sensitivity of voltmeters in ohms per volt. Multiplying the sensitivity by the fsd range in use, gives the resistance of the meter circuit that will be shunting the component. Cheaper multimeters will have sensitivities ranging from as low as 1000 Ω/V volt to as high as 100 kΩ/V.

To illustrate loading effects, consider the circuit in Fig. 8. By Ohms law we



(c) MEASURING THE VALUE OF R

Fig. 8



1000Ω/VOLT METER ON 1 VOLT RANGE

know that the voltage between points A and B is 0.75 volts — don't we?

Now let us see what happens when we use a 1000 ohms/volt meter on the 1 volt range to measure this voltage. The 1000 ohms of the meter in parallel with R2 will produce a combined value of 500 ohms. Thus the voltage read by the meter will be 0.5 volts instead of 0.75 volts — an error of 33 per cent!

It is the *degree* of this shunting effect that is important — in theory it can never be completely avoided, for some energy must flow into the measuring system from that being measured. In electronic measurements the rule of thumb is that for accuracy, the resistance of a voltmeter should be at least ten times that of the circuit — a hundredfold is better still.

This may not always be possible with an inexpensive meter, and some error will have to be tolerated. But remember — this error can be considerable if loading is severe. A check for loading is to make a reading on the next highest range. A much higher voltage reading will indicate that severe loading is occuring on the lower range. When this occurs we must take our reading on the lowest range that does not produce visible loading and put up with the lack of resolution. It should now be clear that the higher the sensitivity of the meter the better.

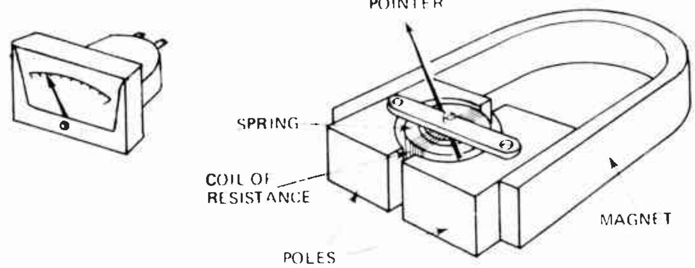
A similar thing happens when measuring currents, the series resistance of the meter may introduce undesirable voltage drops. In this case the combined resistance of the movement and shunt should be one-tenth, or preferably less, of the series circuit resistance.

This discussion of multimeters is restricted to those units that do not include electronic amplifiers — these will be discussed later when we have covered amplifier operation. Amplifier type multimeters are usually characterised by having a quoted fixed input resistance (called impedance in some data sheets) that is typically 1 MΩ or higher. With these there is seldom need to worry about connection loading.

Although we have discussed loading with reference to meter measurements, the same principles apply to the connection of any kinds of circuits — the black-box approach tells us this. Each circuit having input and output terminals will load those coupled to it — output is parallel with input of the next and so on. It is important to remember this, for all so often a circuit stage is developed that fails to operate into the following stage because of loading effects.

THE MOVING-COIL METER

Fig. 9.



When electricity flows through a wire, a magnetic field is produced in a plane perpendicular to the wire. This magnetic field may be concentrated by winding the wire to form a coil of many turns, and still further by winding the coil around a soft-iron core.

Such a device is termed an electro-magnet, when electricity flows through it, it will attract magnetic material — such as iron.

One example of the way such an electro-magnet may be used is the relay in the last practical exercise.

A further example is in the moving-coil meter.

If the electromagnetic coil is suspended in the field of a permanent magnet, (a permanent magnet is made of steel — steel once magnetized, remains magnetized, whereas soft iron loses its magnetism immediately the energising field is removed) it will be caused to rotate, when energized, by a force proportional to the energizing current.

In the moving-coil type of meter, as Fig. 9 shows, the field of the permanent magnet is arranged to pass across a cylinder in which hangs the coil of the meter. A fine spiral tension-spring restrains the rotation by providing a linearly increasing torque as the coil rotates. Attached to the coil is a pointer that moves across a scale, thus indicating current.

As the number of turns is increased, to improve sensitivity, the designer must use finer wire to keep the mass of the coil small. As a consequence of this requirement, sensitive meters usually have a higher resistance, and are more delicate.

CHOOSING A MULTIMETER

Multimeters range in price from ten dollars, to several hundred dollars. The more expensive units incorporate an amplifier to increase the input resistance and/or very elaborate electronic circuitry to provide extremely high accuracy.

For the beginner, extremes of accuracy and input resistance are not necessary and a relatively cheap meter will be entirely adequate. Nevertheless — avoid the very lowest priced meters. These are rarely a bargain.

When purchasing a meter look for the following points:—

1) The sensitivity should be as high as possible, preferably 20 000 ohms/volt or better.

2) It should be capable of reading volts, amps (both ac and dc) and ohms.

3) Voltage ranges should be from one volt to at least 300 volts dc, and 10 to 300 volts ac.

4) Resistance should preferably have at least three ranges, times 1, 10 and 100.

5) Overload protection is highly desirable. This should disconnect, or protect, the meter if excessive voltage or current is accidentally applied.

6) Seek rugged construction. A meter is a valuable tool that should remain serviceable for many years.

4 Frequency and wavelength

WE HAVE seen that a voltage source will provide the force needed to cause charges to circulate in an electronic circuit. Electrons, negatively charged, flow from negative to positive in an attempt to cancel out the charge imbalance created at the supply.

But not all voltage sources create a *static* charge imbalance. It is, in fact, quite possible to produce a condition in which the charge imbalance alternates from positive — negative and then negative — positive, repeating the cycle continuously, thus causing a corresponding alternating direction of current flow.

The principle of alternating current (ac) is simple, and as easy to grasp as that of direct current (dc). Its implications however, go much deeper, for our thinking must allow for the time element present in all ac excited systems.

FREQUENCY AND WAVELENGTH

In a dc circuit, the current always flows in the same direction. It does so with an amplitude that may vary from virtually zero through small to large.

This amplitude — either of voltage or current — may be measured, as we saw in the previous chapter, by using an appropriate meter.

The result may be visually presented in the form of a graph with time on the horizontal axis, and meter reading on the vertical axis. A dc voltage level is shown graphically in Fig. 2a.

In an ac circuit, the current reverses periodically, and voltage sources that operate in this fashion are called alternating current generators.

As electrical energy flows at the speed of light it is possible to regard it as instantaneously following the charge imbalance created at the generator terminals. Note that it is not the electrons themselves that travel at such a speed.

Electrons start moving in a circuit virtually instantaneously when a potential difference is established. They cease moving equally fast when the potential difference is removed.

The effect is rather like turning on a tap.

Water flows immediately, because the pressure from the pump or reservoir is exerting a force on the water in the pipe.

But although *water* flows the moment the tap is opened, it may take hours or even days for any given drop of water in the reservoir to travel along the intervening pipes to the tap.

So with electrons. Although they start and stop moving virtually instantaneously — their actual rate of flow in a conductor is very slow, in fact individual electrons move at mere centimetres a minute.

At first sight it might seem pointless to have an electrical circuit in which electron flow continuously changes direction. The charge flow averages out to precisely zero — so why bother!

Nevertheless, this form of current flow is absolutely essential for the operation of innumerable electronic devices. This will become clearer as the course proceeds. For the time being one can regard the effect of alternating currents (ac) as being similar to the action of a cross-cut saw or double acting steam engine — i.e. work is done during both half-cycles of movement.

In an ac circuit, the current reverses periodically, with a peak amplitude that is usually equal in both directions. As with a dc current, ac currents may vary from practically nothing (fractions of picoamps) to millions of amps. Not only are ac currents variable in amplitude. There are many other ways by which the current can rise, fall, and reverse, with time. Some of these are shown in Fig. 2b.

The behaviour with respect to time is called the waveform of an ac signal. In the case of Fig. 2b, where the reversal of current flow takes place instantaneously, a waveform is produced which, for obvious reasons, is called a square wave.

Waveforms are characterized by three attributes. Shape, amplitude, and cycle time.

If the waveform resembles a common geometric shape it is usually referred to accordingly — square, triangular, staircase — are three examples. If the waveform does not resemble a

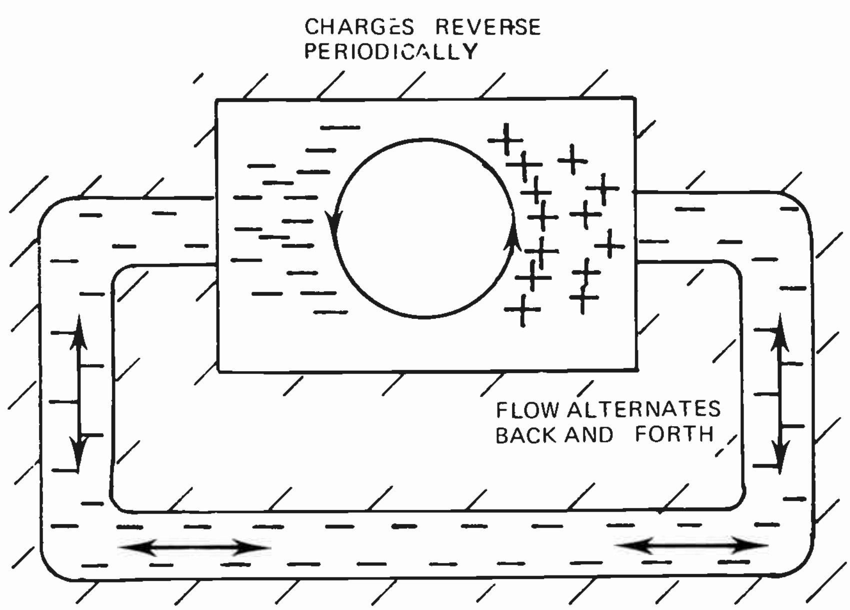


Fig.1. With alternating current, the charges periodically reverse direction.

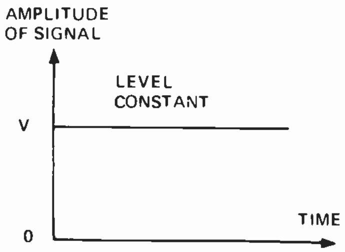


Fig.2a. This form of graph shows how signal amplitude varies with time — this is a dc signal.

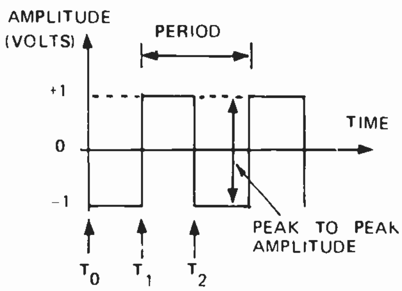


Fig.2b. Amplitude — time graph of squarewave ac signal.

common geometric shape then we define it in other ways appropriate to its characteristic.

Amplitude can be specified in several ways — the most obvious being its value from one peak, through zero, to the opposite peak. This is known as the peak-to-peak value (usually abbreviated to pp or p-p). The square wave shown in Fig. 2b has an amplitude of 2 volts peak to peak.

The cycle time of a waveform — usually known as its 'period' — is the time that a waveform takes to swing from any given point through one complete cycle and back to a similar starting point. It is usually denoted as T. In the case of Fig. 2b the period is from T_0 to T_2.

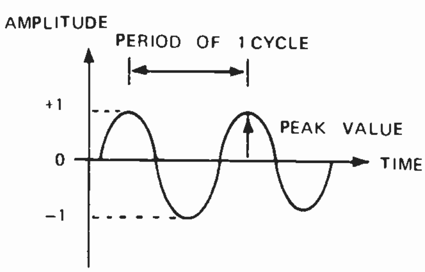


Fig.3. The most basic waveform is the sinusoid.

The period of electronic waveforms varies from one cycle every few thousand seconds to one cycle every fraction of a picosecond (one million-millionth of a second), or less.

The number of such periods occurring in a given time is known as the 'frequency' (usually denoted as f). Thus frequency and period are related — as $f = 1/t$, provided both are measured on the same units of time; seconds being generally used. The number of cycles occurring in one second used to be known (reasonably enough!) as 'cycles per second'. Nowadays the term Hertz is used for the same parameter. Hertz is generally abbreviated to Hz, often with a multiplier prefix such as kHz (one thousand Hertz) or MHz (one million Hertz).

To illustrate the inter-relationship between period and frequency, a 1 MHz squarewave has say, a cycle time (or period) of 10^{-6} seconds. Thus it repeats the waveform one million times a second. At the low end of the spectrum a 1 millihertz waveform takes 1000 seconds to pass through each complete cycle. (Note — although the abbreviation 'm' denotes milli — it is not generally so used as a prefix when denoting frequency. i.e. the abbreviation mHz is correct but millihertz will nevertheless normally be written. This is because frequencies of less than 1 Hz are not commonly encountered in electronics, and because of its rarity, 1 mHz may well be mistaken for 1 MHz — a unit one 1000 million times larger!).

INTRODUCING THE SINE-WAVE

Although the square wave is the easiest alternating waveform to comprehend — it may be produced simply by a battery and a continuously reversing switch — it is not the most basic waveform that exists.

Square waves, and in fact all other waveforms, are really composed of multiple waveforms called 'sinewaves'.

As the name implies, the amplitude/time graph of a sinewave is sinusoidal. It follows the trigonometric sine function as time proceeds. The shape of a sinewave is shown in Fig.3.

The instantaneous values for a sinewave between the two peak-to-peak values can be obtained from a set of natural sine tables by giving it the appropriate sign change during the cycle.

Unlike square waves, a sinewave current does not have rapid transitions (called 'transients') but varies smoothly from zero, rises to a positive maximum, falling through zero, to rise in amplitude in the opposite sense.

It is often necessary to be able to equate the mean energy value of sinewave current to that of dc current. Clearly the total charge moved (energy, therefore) in a given time by a sinewave differs from that moved by a dc current having the same *peak* value (the peak value of a dc current is of course the same as its mean, or average, value). The peak to peak value of a sinewave does not express the true mean energy flowing.

The equivalence occurs when we use a value less than the peak of a sinewave. This lower value is known as the rms value. The term rms is short for Root of the Mean of the sum of the Squares — of the instantaneous values and this turns out to be 0.707 of the zero to peak value.

In other words the rms current (or voltage) in an ac system fed with sinewave power has the same energy level as a dc system fed with dc power of the same numerical value.

Heating illustrates this well. Our household 240 volt 50 Hz ac supply is not 240 volts peak to peak, but 240 volts rms. An electric heater energized by the 240 volt mains ac supply would therefore produce the same amount of heat if fed from a 240 volt dc supply instead.

It is not generally appreciated that our 240 volts 50 Hz ac mains has a zero to peak value of 340 volts. Considerably higher than the value quoted!

In many designs the peak value of an ac waveform must be very carefully considered for it is this value that determines the insulation required. Because of this many components have their voltage ratings quoted in peak not rms value.

Sinewave power is sometimes quoted as an 'average' value. This is the average of the sum of instantaneous values). It works out at 0.637 of the zero to peak value. Average power is

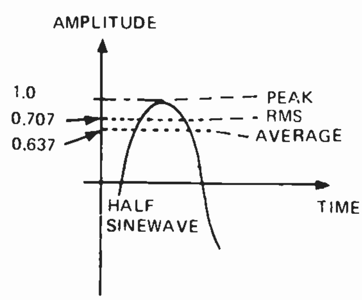


Fig.4. Relationship between average, peak, and rms. of a sinusoidal waveform.

rarely quoted as it has relevance in special applications only. Figure 4 shows the relationship of peak, rms and average values.

HOW SINEWAVES ARE GENERATED

Most of the world's electrical energy is generated as sinewaves.

The basic principle, used in everything from power stations to the alternator now fitted to modern cars, is very simple.

When a loop of wire is rotated in a magnetic field a current is caused to flow in the wire by the influence of the magnetic field.

The direction of current flow is determined by the direction of the magnetic field with respect to the direction of motion of the conductor. By examining Fig.5 it may be seen that each side of the loop will cut the magnetic field alternately downwards and upwards as it rotates. Thus the current must reverse in the loop of wire. The current is at maximum when cutting the field parallel to the North/South pole axis, and at minimum when perpendicular to it. Thus the current is *also proportional to the angle of the conductor in the field.* These two effects combined produce the waveshape output known as a sinewave. The law governing the direction of current flow in a conductor is known as Faraday's law and will be discussed in more detail later.

In practice, the rotating loops of alternating-current generators have many turns and, more often than not many pairs of magnetic poles.

The rotating alternator is an efficient way of generating electricity in large quantity, but only in large quantity – i.e. fifty watts upwards. Because of this and for reasons of practicality, the majority of sinewaves used in electronic equipment are in fact generated by special electronic circuitry powered from dc, that in turn, is either generated by batteries, or produced from the 240 volt ac mains. More about this later.

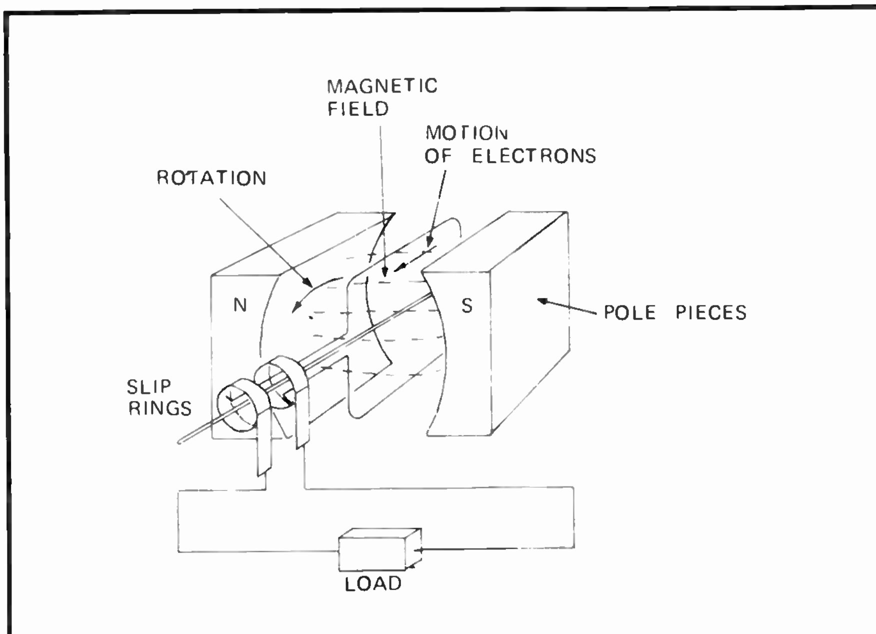


Fig.5. Basic electromechanical sinewave generator.

NOTHING IS PURER THAN A SINEWAVE

In the 1800's, when science was more devoted to thought than hardware, and electricity was still a 'magic' trick, mathematicians were laying the foundations of our present-day sophisticated theories.

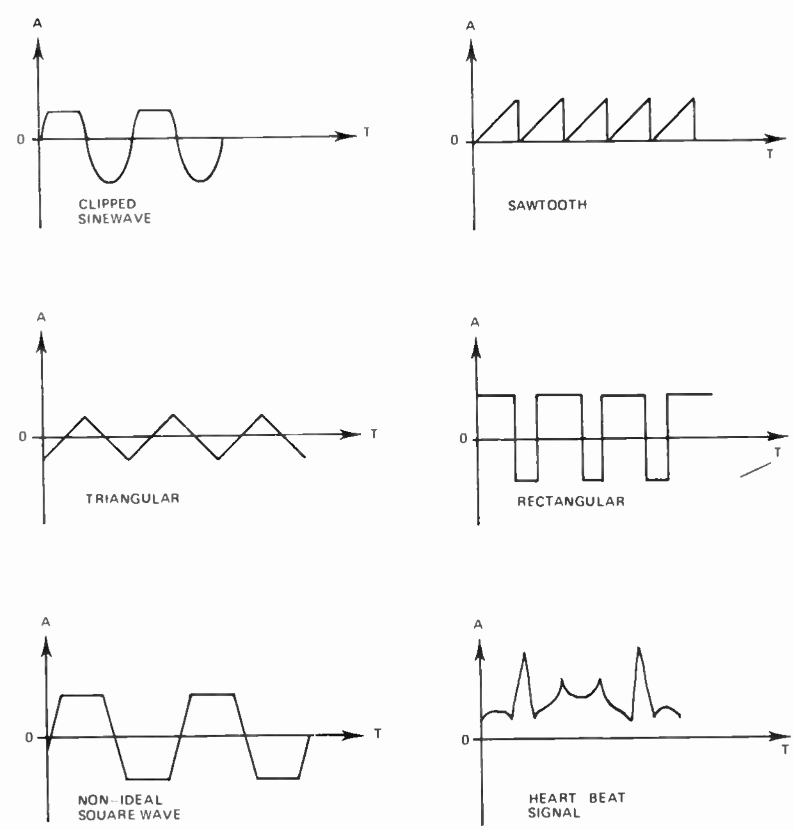
One such mathematician was Jean Fourier, a Frenchman.

Fourier's thing was the solution of numerical equations. He eventually proved that all *periodic* functions (i.e. all functions that repeat periodically) could be broken down into a sum of sinewaves having different amplitudes and frequencies. As each individual sinewave cannot be broken down any further it seems that the sinewave is the purest of waveforms obtainable.

The method, now known universally as 'Fourier Analysis', involves relatively advanced mathematics. Nevertheless the concept can easily be demonstrated using our now familiar amplitude/time graphs.

Waveforms that are not strictly sinusoidal are known as 'complex'. Various complex waveforms are shown in Fig.6. As will be shown, our

Fig.6. A number of complex waveforms.



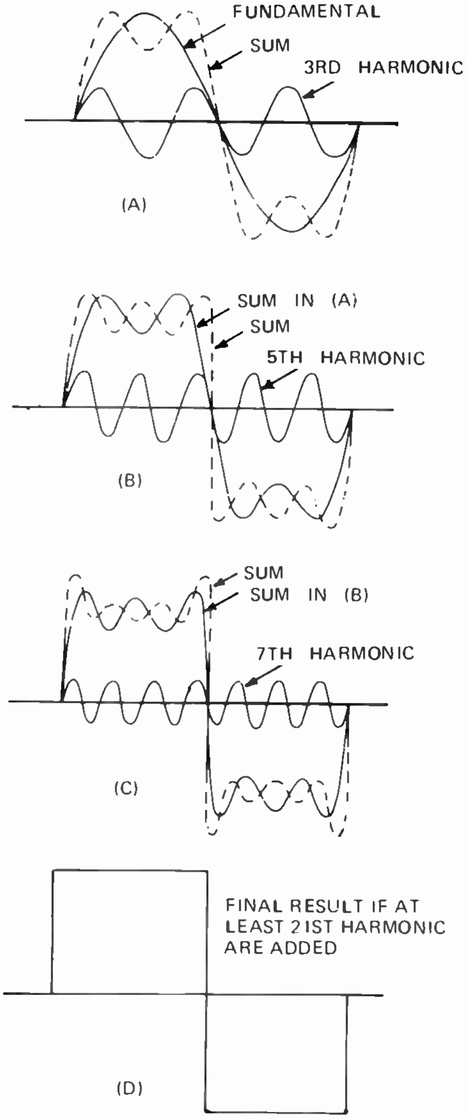


Fig.7. The apparently simple squarewave is really made up of the addition of numerous sinewaves of different frequency and amplitude.

apparently simple looking squarewave is in fact quite complex. It contains sinewave frequencies varying from one having a period identical to that of the squarewave, to one with (theoretically) infinite cycles per second.

Figure 7 shows how a squarewave is made up of a (theoretically) infinite number of sinewaves. In (a) a sinewave, with period equal to that of the squarewave (this one is called fundamental frequency) is added graphically to another sinewave of three times the frequency (known as the third harmonic) but only one third the amplitude. Already the addition of the third harmonic makes the waveform begin to look like a squarewave.

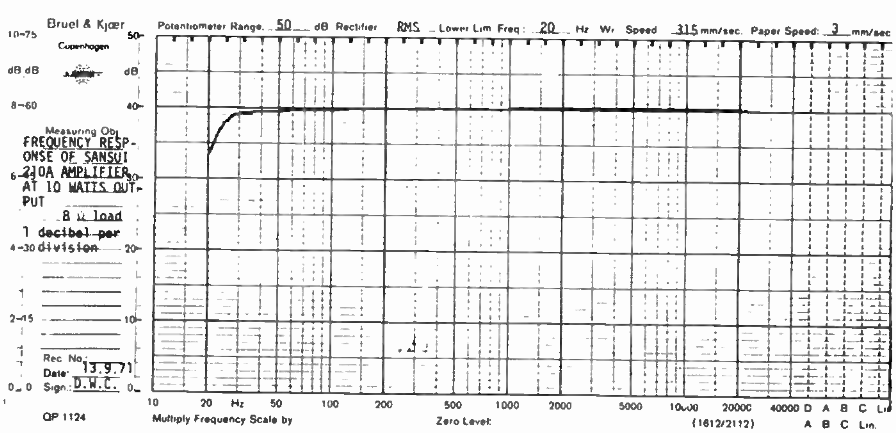
In Fig.7b we have added the fifth harmonic at an amplitude one fifth that of the fundamental – the waveform now looks even closer to a squarewave. In (c) we have added the seventh harmonic – at one seventh the amplitude.

By continually adding frequencies in this way we eventually obtain our final squarewave. In practice, up to the 21st harmonic will be required to obtain a reasonably shaped squarewave.

All complex waveforms can be broken down in this way, not only in theory but also in practice, in fact one way of generating sinewaves is to generate a basic squarewave – perhaps even by using a battery and an automatic polarity reversing switch – and then to use an electronic filter to extract the fundamental frequency.

The concept of Fourier analysis is another basic tool of the person trained in electronics – we will refer to it again throughout the course.

Fig.8. This frequency-response plot (of a hi-fi amplifier) was obtained using an automatic recording chart plotter.



OUR LARGE RANGE OF FREQUENCIES

Electronics in essence may be seen as a vast, tremendously versatile service facility providing the means of transferring energy between input and output devices.

A TV set, as we showed in the first part of this series, receives its signals from the broadcasting station, processes them, and finally uses these signals to recreate the sound and picture of the original programme material. An alarm circuit uses some form of switch to operate a warning device. A computer accepts information – in the form of coded electrical impulses, or holes in paper tape or cards, and then processes this information to produce further holes in paper, tape or cards, printed output data, or whatever.

In radio broadcasting we use high frequency radio waves to carry the audio sound waves across vast distances. In contrast, the same high frequencies may be used for heating and welding plastics, and for deep-heating in medical therapy.

A microwave oven uses very high frequency energy to heat – a case where information transfer is not a factor.

Examples such as these indicate why such a wide range of frequencies is needed to handle our diverse needs. The frequency that we need to use is often related to the speed with which we need to transfer or process information.

At the low end of the frequency spectrum is the simple alarm circuit where one piece of information needs to be transmitted only when needed. Is the alarm contact open? (If not, it must be closed).

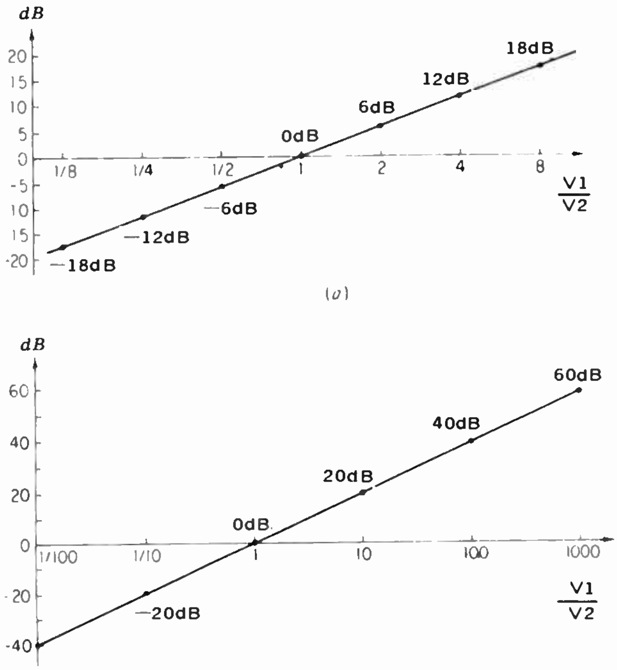
There is little need for extremes of operating speeds as the relatively slow human senses are used to interpret and act on the data received.

Computers may need to process literally millions of pieces of information in as short a space of time as possible. Because of this, their frequency of operation is forever being increased and present day computers have switches capable of operating at speeds in excess of 1000 million cycles per second.

GRAPHICAL REPRESENTATION

At the beginning of this part of the series we introduced amplitude/frequency graphs operational sequence in a system where things are happening faster than our senses can follow.

In many electronic systems, such as



Ratio $\frac{V1}{V2}$	in decibels
1/100	−40
1/10	−20
1/8	−18
1/4	−12
1/2	−6
1	0
2	6
4	12
8	18
10	20
100	40
1000	60

Fig.9. These charts and table give rapid conversion between voltage gain and the equivalent decibel value.

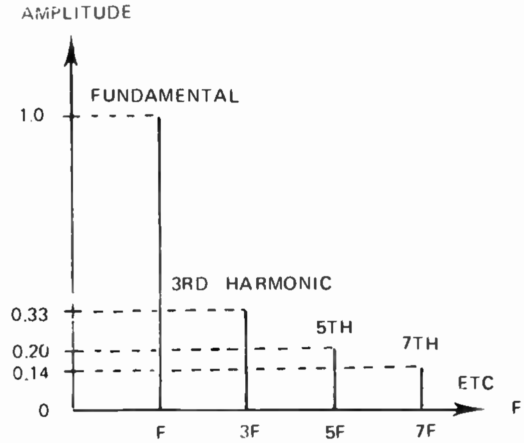


Fig. 10. Frequency spectrum of squarewave signal.

hi-fi equipment, we often need to know the amplitude of a signal at a particular frequency or range of frequencies. To do this we can use a graph on which amplitude is plotted as before, but with time replaced by frequency on the horizontal axis.

Figure 8 shows the frequency response of a hi-fi amplifier; that is an indication of how output level varies with constant level input at various frequencies. Ideally an amplifier should have a 'flat response' — that is, it should amplify all signals within the audio range by an equal amount regardless of frequency.

The first thing to note about Fig.8 (and all similar graphs) is that frequency is not plotted on a linear scale. A logarithmic scale is used instead. The reason for this is that to cover the audio frequency range (that is from approx 20 Hz to 20 kHz) linearly would require several yards of paper. The use of a logarithmic scale compresses the information needed into a more satisfactorily handled format.

The graph shown in Fig.8 starts at 10 Hz and passes through 100, 1000, and 10 000 Hz decades. Paper ruled in this fashion is usually obtained ready printed, but it is worth remembering that a crude log plot may be made by ruling lines at equal intervals for each decade, then subdividing each decade by three equal intervals delineating the 2 and 5 unit positions.

Although it does not appear to be so at first sight, the vertical scale in Fig. 8 has also been compressed logarithmically. Here, although the actual markings are at linear intervals, the response has been plotted in units called decibels (dB) — units used to relate voltage or power levels on a logarithmic basis.

As with frequency, voltage and power levels used in electronics cover an enormously wide range. Several thousands of billions in fact. Thus to show a signal varying amplitude from say, 1 nanovolt (10^{-9}) to 10 volts, would require an enormously wide piece of paper to allow us any resolution at all. So again we need to compress the scale.

The decibel method compresses the values logarithmically *before* plotting and by so doing produces a linear scale — as Fig. 8 shows.

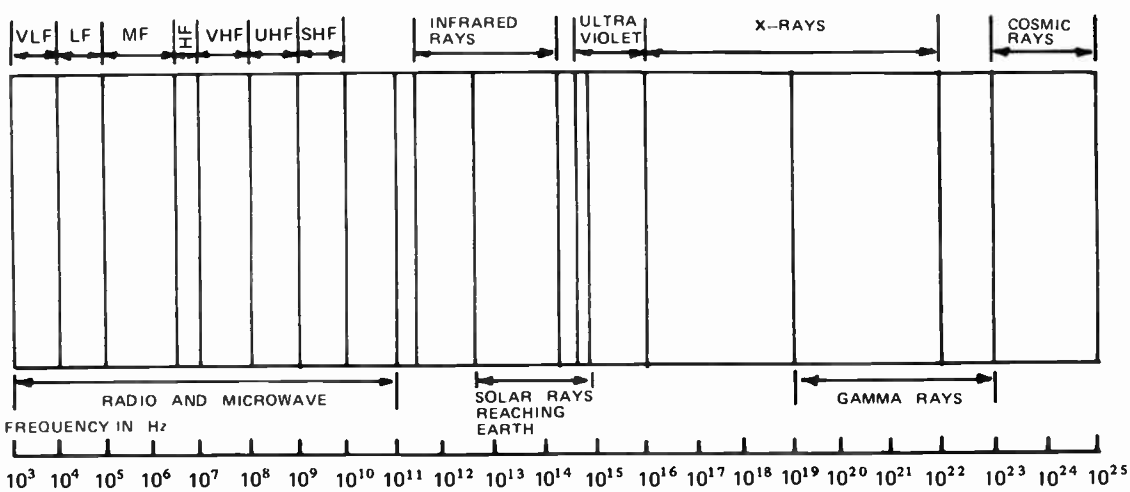


Fig. 11. This chart shows the enormously wide frequency range of the electromagnetic spectrum.

As with many electronic circuits, the response of an amplifier is a matter of relative comparisons between two values, in this case a constant amplitude input signal and the resultant (amplified) output signal. For voltage levels the decibel value is obtained by the expression –

$$dB = 20 \ Log_{10} \ \frac{V1}{V2}$$

Some useful values and charts are given in Fig. 9.

We should now be in a position to interpret Fig. 8. The graph shows us that the output of the amplifier falls off as the frequency falls below 30 Hz. Above this we say that its response is 'flat'. The curve terminates at each end, not because an amplifier cannot be made to work beyond these frequencies, but because it has no useful function beyond these limits.

THE FREQUENCY SPECTRUM

Another form of frequency plot is the so-called frequency spectrum. Previously we have discussed the response of a 'black box' to various input signals. A frequency spectrum on the other hand displays what frequencies are present in a signal and with what relative amplitudes. Figure 10 shows the amplitude/frequency spectrum of a square wave.

A spectrum of particular and general interest is that of electromagnetic radiation. It is the one that involves radiations of energy travelling at the speed of light.

The electromagnetic spectrum is shown in Fig. 11. It gives the names that have been adopted for the various ranges of frequencies that can be used. There is no concept of amplitude here – the graph merely illustrates the range. It is a very good example of the need for logarithmic scaling – for the known extent of the frequency spectrum ranges from zero frequency up to 10^{25} Hz.

PHASE

So far we have discussed relationships between amplitude and time or frequency. Phase is also important.

When two or more signals of the same frequency exist together, their relative timing can be important. If they pass through zero in the same direction at the same time we say they are in phase. If not, they are out of phase. The units used to express this relationship are degrees, fractional cycles or actual time periods. Phase/time and phase/frequency graphs can be drawn. In some applications these are vital, but generally to consider phase in this way is rather rare.

At this point it is appropriate to point out that the discussion above applies not just to electrical and electronic signals but to any periodic phenomena – for example, in mechanical, acoustical or even optical systems.

SEEING THE WAVEFORM

As an ac waveform varies in amplitude with time, a single instantaneous measurement does not necessarily provide the information sought.

If the waveshape is accurately known, as for instance the reasonably closely controlled 240 volt 50 Hz mains supply (sinewave), it is feasible to measure its effective value. This is done with some form of averaging device thus providing what appears to be a single measurement. The ac scale on a multimeter does this.

But if the waveshape and frequency are not known it is necessary to have some means that tells us what is happening.

For extremely slowly varying waveforms, such as squarewaves with periods of minutes, a meter, (reading zero in the centre of the scale) will enable us to observe the amplitude in each direction and the time taken to switch back and forth.

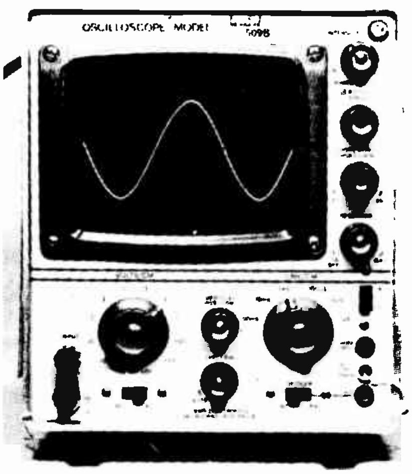
Unfortunately most signals vary faster than we can observe them. More sophisticated methods are necessary.

The most versatile instrument for this (and other) purposes is the cathode ray oscilloscope. Its operation is shown in Fig. 12. In its simplest form it displays what is happening to the signal amplitude as a vertical deflection of a spot of light on a display screen, whilst a signal proportional to time deflects the spot to and fro on the horizontal axis.

By triggering the horizontal sweep at precisely the same place on the repetitive waveform to be observed, the spot traces out a piece of signal wavetrain that persists on the screen long enough to be seen. When the sweep reaches the end it is again re-triggered thus overlaying a second swept pattern on top of the first – and so on.

Oscilloscopes enable us to see waveforms from periods of one cycle per thousand seconds to gigahertz frequencies. Even the cheapest of oscilloscopes will cover the audio spectrum and well beyond – but very wide range 'scopes are expensive devices.

Next to the multimeter, the



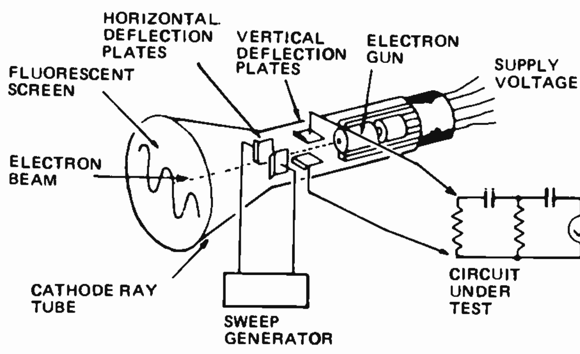


Fig. 12. Typical cathode ray oscilloscope (C.R.O. for short). The schematic block diagram of its internal black boxes is also shown. The sweep generator causes the voltage across the horizontal plates to rise steadily deflecting the electron beam horizontally across the screen. It then rapidly returns, to repeat the process – this is the time base. The signal to be studied is applied to the vertical amplifier input where it causes the spot to deflect in the vertical direction as it sweeps across thus producing the shape of the waveform.

oscilloscope is the most useful diagnostic tool available. Its price (from a couple of hundred dollars upwards) unfortunately rules it out for most amateur work, but any course given by an educational institute would use them. As we do not expect our readers to obtain one this course will not normally require its use in the practical exercises.

If the frequency of the signal to be measured lies in the region from dc to 10 kHz, it is usually possible to 'see' the waveform by using a suitable chart recorder.

These devices use some kind of electro-mechanically operated pen that follows the input signal producing a displacement corresponding to input signal. A mechanism similar to a moving coil meter is often used. The movement of the pen is then recorded on paper which moves under it at a constant known speed.

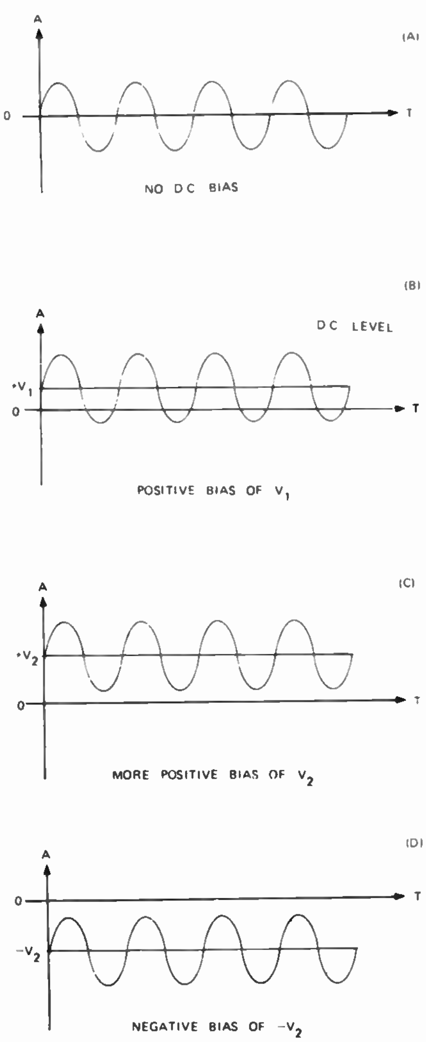


Fig. 13. These graphs show how dc and ac can exist together in a circuit, the dc level providing 'bias' to the ac signal.

Chart recorders come into their own when the signal is varying slowly – such as air temperature variations over a monthly period: they also provide a permanent record.

DC LEVEL OF AN AC SIGNAL

So far we have described ac signals in which the current alternates with equal amplitude in both directions.

It is however possible to have an ac waveform that is *not* symmetrical about zero. An ac signal can be produced, or adjusted, to produce this lopsided situation.

In Fig. 13, the sinusoid is symmetrical about zero in (a) but is biased to progressively greater extents in (b) and (c), in fact the waveform shown in (c) never swings negative at all. Some would argue that waveform (c) is not an ac signal but merely a varying dc signal. But nevertheless it would normally be regarded as an ac signal superimposed on a dc level.

It is possible to separate the ac and dc components of such waveforms very simply – more of this later.

MEASUREMENT OF AC VOLTAGE AND CURRENT

Just as with dc signals, it is very necessary to be able to measure ac signals, such as voltage and current, in order to be able to check circuit operation.

If the signals are sinusoidal (as many are) we can measure the various values, such as rms, peak and average, with a meter made for ac measurements such as the moving iron repulsion type shown in Fig. 14. In this type of meter, the coil is fed with the ac signal, thus producing an electromagnetic field (rather like the relay described in the second article in this series). The electromagnetic field repels a moving iron vane. The greater the amplitude of the signal, the stronger the magnetic field and hence the greater the movement of the vane.

Although cheap and robust, moving iron meters are not very sensitive. Apart from this they have non-uniform deflection characteristics (i.e. vane movement is not linearly related to signal strength). Ac moving coil meters (described in the last chapter) are more sensitive than their moving iron equivalents.

If ac current is fed directly to the meter movement, the needle will attempt to follow the positive and negative excursions of the current. It will vibrate about the zero point indicating little except that the signal is ac not dc.

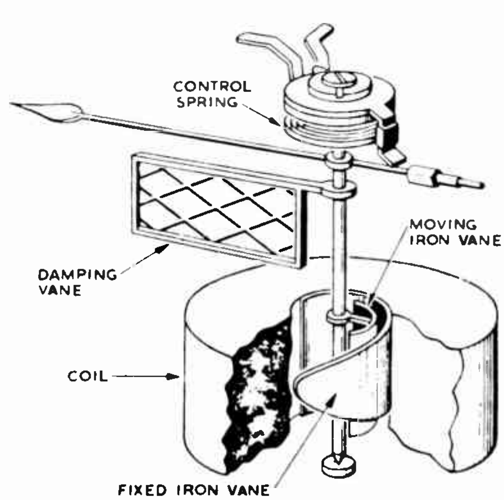
RECTIFYING AC TO PRODUCE DC

Each ac cycle of an unbiased waveform has a positive half cycle and a negative half cycle. If a 'switch' is used to let the positive half cycles through but to block the negative half cycles the resultant waveform will appear as shown in Fig. 15a (centre).

If the positive half cycles passed by the switch are now passed into an 'energy store', the fluctuating ac signal will be smoothed out and dc obtained.

This process is called half-wave rectification and the switch used is called a rectifier.

An ac signal rectified in this way can be measured by a conventional moving-coil meter calibrated in rms ac signals. In meters of this kind, the mechanical inertia of the meter movement acts as the averaging energy store (much in the same way as a flywheel smooths out the individual firing impulses of a car engine). In other circuits a capacitor is used instead.



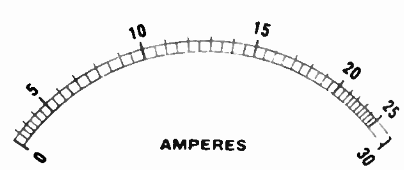


Fig. 14. How the moving-iron meter works – illustrating the principle of attraction. This type of meter will also work on dc signals. The scale, shown right, is non-linear.

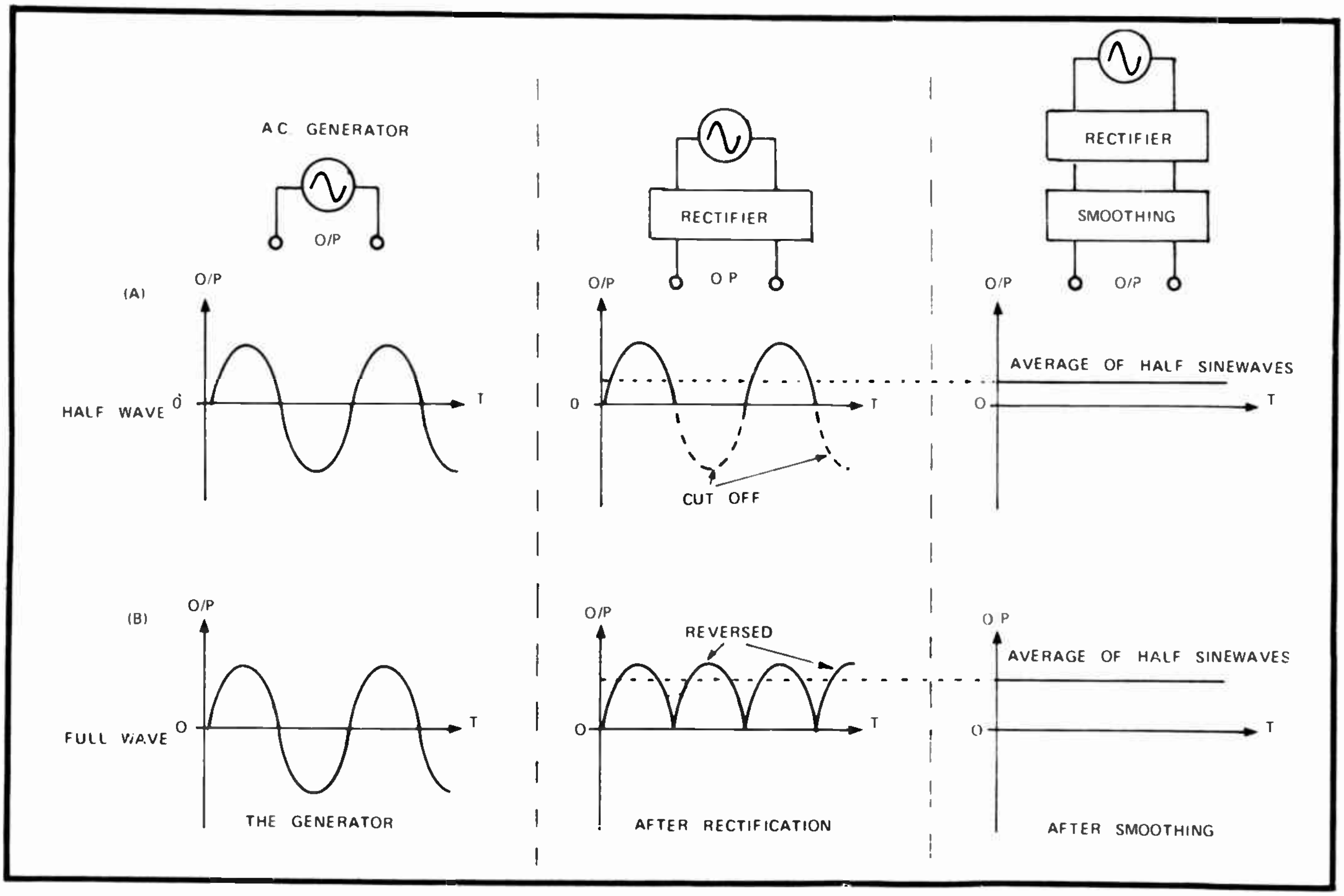


Fig. 15. Halfwave and fullwave rectification – and the black-box representation of the process used. Note the higher output level obtained from the fullwave process.

Although half wave rectification is simple, it is inefficient. Half the available energy is blocked by the rectifier and cannot be used. Also the discontinuity of the waveform adds to the difficulty of adequate smoothing.

A better method of rectification is to use a switch that reverses the polarity of the negative half cycles – thus making them positive. This is illustrated in Fig. 15b (centre).

This process is called full-wave rectification and is the most usual method of converting ac signals or power to dc.

This has taken us back to where we came in – measuring ac on a multimeter. We have seen that when an ac range is selected, the range switch inserts a rectifier between the input terminals and the meter movement, and if need be, adjusts the various shunts and multipliers to obtain the correct rms readings.

Later in this course we will return to look at the various kinds of rectifiers and their use. Their design and operation is relevant when building power supplies to operate electronic equipment powered by the ac mains supply.

A BASIC ELECTRONICS LIBRARY

It is impossible to remember all the facts of electronics. So as the course progresses we will recommend various inexpensive books or data sheets that are worth collecting.

A good start is to obtain an electronics dictionary of terms such as the paper-back –

A DICTIONARY OF ELECTRONICS, Third Edition 1971, S. Handel, Penguin Reference Books, $1.55.

The following are good reference texts –

ELEMENTS OF ELECTRONICS, Parts 1 and 2, by F.A. Wilson, published by Bernard Babani, about $7 each.

ELECTRONICS SELF–TAUGHT, Ashe, published by TAB books, about $8.

HOW TO READ ELECTRONIC DIAGRAMS, Brown and Lawrence, published by TAB books, about $7.

PROJECT ELECTRONICS, by ETI, published by Modern Magazines, $4.75.

GETTING ACQUAINTED WITH THE IC, Rufus P. Turner, published by Howard Sams, about $7.

TRANSISTOR ELECTRONICS, by Gerrish and Dugger, published by Goodeart-Willcox Co. Inc., about $15.

NEWNES RADIO & ELECTRONICS ENGINEERS POCKETBOOK, published by Newnes-Butterworths, about $7.

SOLID STATE BASICS, published by the ARRL, about $9.

A COURSE IN RADIO FUNDAMENTALS, published by the ARRL, about $7.

IN THIS exercise, the intention is to provide experience with ac signals and the use of ac ranges on the multimeter.

One way to obtain this, that is both useful and instructive, is to build a small dc power supply operated from the ac mains. This will provide familiarity with ac components and connecting equipment to the mains.

A MAINS OPERATED SUPPLY

This unit, to which we will later add a circuit that controls its (adjustable) output voltage regardless of changes in supply voltage, provides dc current at a voltage ranging from 18 volts maximum off-load to 10 volts minimum on full load.

The completed unit may be used to replace your 12 volt battery supply in simple electronic applications, such as the relay circuit described in Part 2. It may also be used to power any other normal low current 12 volt devices such as model trains, portable cassette recorders etc.

HOW IT WORKS

The household mains supply delivers power at 240 volts rms 50 Hz.

To obtain the required 12 volt output it is necessary to reduce this voltage. This is done with the transformer shown in Fig. 16.

Inside the transformer are two separate, insulated windings, enclosed by a magnetic iron loop. By appropriate choice of the number of turns in each winding, it is possible to reduce (or increase) the voltage of an ac supply to the level needed.

Transformers will be discussed later in this course, but for now it is only necessary to recognise that in the transformer used for our simple project there are only four leads. Two of these go to the mains, two to the 12 volt circuit. Most transformers have these leads clearly marked. If not it is essential for you to ask a knowledgable person — such as your science teacher, if you are at school.

The reduced voltage ac has to be rectified. This is achieved using diodes arranged in what is called a fullwave bridge.

These diodes are solid-state switches. They allow current to flow only in one direction — as shown in Fig. 17. As an exercise trace out the conducting paths through the rectifier bridge remembering that the transformer provides opposite polarities alternately to the input to the bridge. You will see that the output from the bridge is always of the same polarity.

Although you will not be able to see the waveform without an oscilloscope it will look like Fig. 15b.

Our next and final stage is to arrange to average out the non-smooth waveform. This is done by the capacitor.

It is vital that the leads of this capacitor be connected the right way round as reverse polarity will not provide correct operation and will almost certainly ruin the capacitor — sometimes explosively!

The positive connection is usually marked on the case, either by a red mark or by a positive sign (i.e. +). If

PARTS LIST ETI 217

Transformer 240 volt primary: 12-15 volt secondary (approx 1-1½ amps).
4 — diodes EM401, IN4001 or equivalents.
1 — capacitor 220 μF, 25 volt working electrolytic.
2 — terminals 1 red, 1 black
2 m three-core mains cable.
1 — three-pin plug.
1 — rubber grommet for power cable.
1 m 23/0076 connecting wire.
Tag strips, screws, etc.

Alternatively the negative terminal will be marked with a black band. If not it is again advisable to obtain assistance.

USING THE MAINS SUPPLY

We cannot stress too strongly that the mains supply can be lethal if mishandled.

There is only one safe way in which to work. This is to make all connections and circuit changes with the power plug pulled out. NEVER TRUST THE SWITCH IN THE POWER POINT, for such switches often break just one of the two wires connected to the outlet. If incorrectly wired, as they frequently are, full mains voltage will still be applied to one terminal of the power point even though the switch is off.

This warning should not be taken as a discouragement to use the mains supply. Correct practice will safeguard you at all times.

The power supply unit should be earthed. That is, the earth wire from

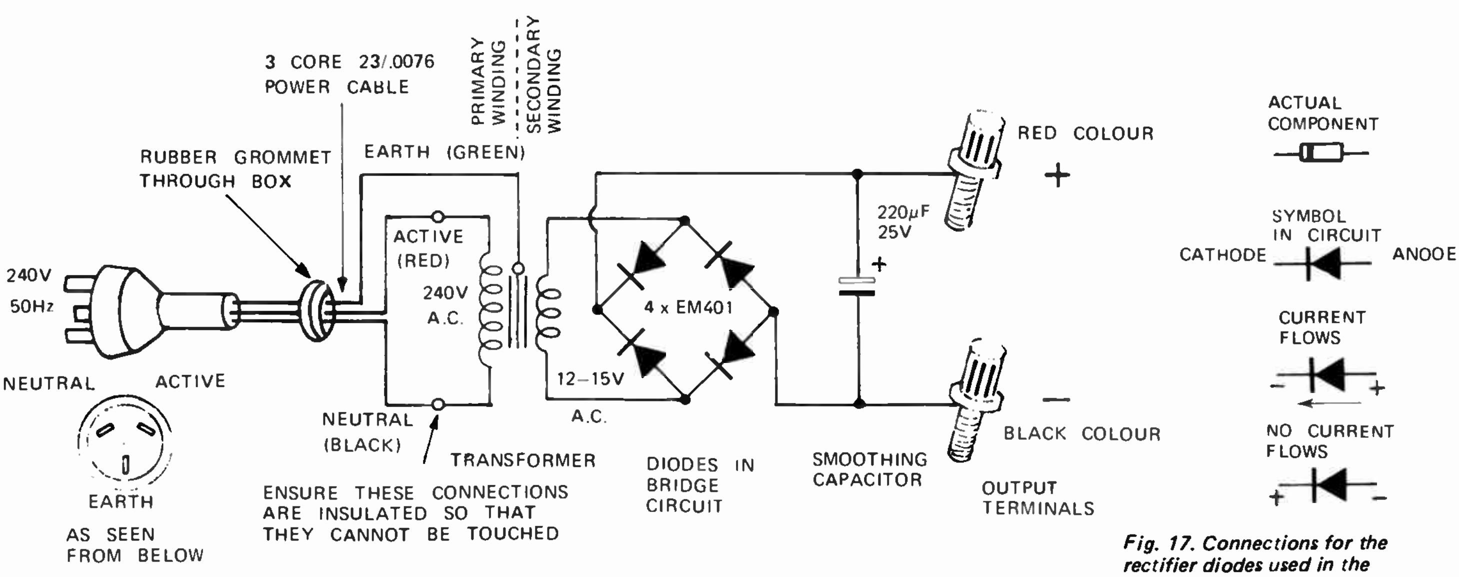


Fig. 17. Connections for the rectifier diodes used in the project shown in Fig. 16.

the three pin plug and cord must be connected to all the metal parts associated with ac. In our case the earth wire must be connected to the transformer case, and to the power supply case – if it is made of metal.

Modern plugs have the pin connections clearly marked.

Once the input (or primary winding, as it is called) is correctly connected there is little danger, for the 18 volt (approx.) secondary winding is insulated from the mains and produces a voltage too low to be dangerous.

Even so, it is good practice never to make or alter the circuit in any way unless the power plug is pulled out of the mains socket.

Having built the power unit you will need to know if it works correctly.

Plug it in and turn on the mains. Then, using the ac volts range set to read 12 volts around mid-scale, measure the voltage across the secondary winding.

It is wise at this stage, not to measure the primary voltage, for a slip of the finger could give you quite a fright. If there is secondary voltage available, then the primary must be operating.

Now switch to a dc volts range and measure the voltage across the output terminals. This should be about 16 volts depending on the transformer used. This is the peak value of the half sinewaves (the capacitor charges up to the maximum peak value with no load applied).

Finally carry out a load test. This is done by progressively adding loads until the minimum allowed measuring voltage is obtained when the load is added.

The results are then plotted on linear graph paper with voltage on the vertical axis and load current on the horizontal axis. This is called a regulation curve – it shows what happens to our 'black box' as load is increased.

Depending on the multimeter used, you can also measure currents in the various wires. Remember that current is measured by placing the meter in series with the lead of interest. (Not all multimeters have ac current scales).

Notes

5 Electronics and communication

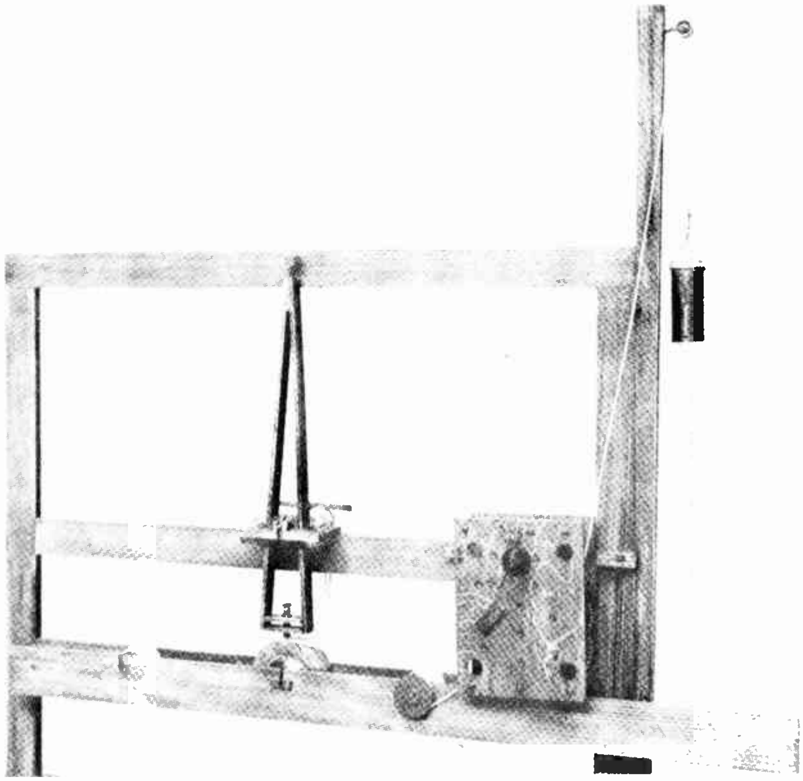


Fig. 1. The receiving terminal of Morse's first telegraph. The falling weight drives the paper strip past the triangular shaped pencil holder. Current received by the coils halfway up this holder cause the beam to deflect making a mark on the trace.

ELECTRICAL methods of sending information have existed on the Earth since the dawn of creatures, for the nervous system, that is so fundamental to life, is based on information transfer by the use of electric currents!

So there is nothing new to the electrical communication systems — but it is only in the past 140 years that man has made use of the principle in his inventions.

The Ancient Greeks knew a little about electricity . . . Thales of Miletus wrote about charged amber, . . . but it was not until recent times that its existence in living creatures was discovered.

In 1791, Luigi Galvani reported his observations (his wife's in actuality) of the effect of applying electricity to frogs' legs, hinting in his writings that electricity exists in tissues. It is very doubtful that Galvani saw the inherent implication that it acted as the means of communication between sensors and muscles.

Forty years passed by, and then, around 1830, a doctor named Jackson was demonstrating how a coil of wire wound around a piece of soft iron caused it to become magnetic, doing it more as entertainment than as a scientific demonstration.

Observing the experiment was Samuel Morse, a well-known portrait painter of the time. He saw the 'trick' as the basis of a method for transmitting intelligence between people located remotely from each other. He visualized electricity, transduced from an invisible current to a visible movement thus enabling messages to be sent using invisible electrons. He was thinking of the telegraph as we know it now.

In Morse's time communication had remained little changed since time immemorial. But the tempo of life was rapidly quickening demanding more speed in communication.

Forms of telegraph existed . . . shuttered light beams, semaphore flags . . . but they lacked speed and versatility.

For several years, Morse developed methods for sending code along a wire circuit using an electro-magnet to deflect a pencil resting on the paper thus producing a scribed line on a moving paper strip (Fig. 1). The method was limited, however, to short distances (devices were not very efficient in those days and most of the power available was wasted as heat) and this drove him to think how he could regenerate the signal). Eventually Morse devised the relay: the electrical input signal, providing enough power to operate the armature, closes the next circuit via contacts and so on. The relay is a simple type of power amplifier, for the power output can be many times the power input.

Eventually Morse devised a system of transmitting information using a code in which combinations of short and long dc pulses were used to represent letters and symbols.

The dc pulses were transmitted along telegraph wires and if the distance between the receiver and sender was very long — intermediate relays were used.

Relays (described in Part 2 of this series) consist basically of a solenoid, which, when energized by an electric current, cause one or more pairs of contacts to close.

As used by Morse, the transmitted signal was used to energize the relay solenoid — this caused the associated contacts to close — which in turn provided the necessary switching for the next stage.

A whole series of such relays were used if necessary to enable the data to be transmitted over thousands of miles.

So much for the telegraph and the use of switched dc currents for information transfer. Let us now look at the early developments in the use of ac currents for this purpose.

In 1887, the son of a Hamburg lawyer, Heinrich Hertz (Fig. 2), wrote in his diary that he had discovered what we now call radio-waves. (Do not bother to search for an explanation of what these *are* for no one knows . . . we do know, however, what they can *do* and how to use them to our advantage). These time-varying oscillations pass through space without the need for wires, being in fact, just a small part of the electromagnetic spectrum we mentioned in a previous part of this series.

Hertz' discovery started the still-continuing search to find ways to generate and sense waves across the entire electromagnetic spectrum. It is said that Hertz thought there was little use for his discovery! The unit for frequency . . . the Hertz . . . is named after this scientist . . . whose pioneering work opened up new vistas in our capability.

Hertzian waves, as they were called at first, could only be detected visually to begin with . . . as minute sparks across the receiving coil's terminals. By 1890 Edouard Branly, in Paris, had devised a technique that could use the waves to operate a relay. It relied on the fusing together of metal filings when they were placed in a field produced by the waves. As the filings joined up they closed an electrical circuit. The method was very crude, and most inconvenient, for the filings had to be shaken apart after each bit of information. It was, however, the start of many experiments that were to eventually lead to the development of radio communication as we know it today.

Fig. 2. A contemporary etching of Heinrich Hertz.

Marconi is generally credited with the first successful radio broadcast but some historians give that credit to the Dane, Johannes Sorensen, who is said to have signalled a ship from the shore even before Hertz reported his findings (in 1866).

Also predating Marconi, was the Russian engineer A.S. Popov, who on 7 May, 1895, demonstrated the possibility of sending messages over the air.

Popov was also the first man to establish a practical regular radio link — between the island of Kuutsalo, near Kotka, and the island of Suursaaren. This link was put into service on Feb 4, 1900, and was used regularly until April, 1900. During 8 days, 440 official telegrams were transmitted.

In 1897, Marconi, helped by several others, succeeded in transmitting lengthy messages over a distance of 7 km between Lavernock Point in Wales to the Island of Flatholme situated in the middle of the Bristol Channel. His success was immediately met with acceptance (not like Sorensen who was publically ridiculed for suggesting he could communicate without wires).

In 1898 Marconi set up a link between the Royal yacht "Osborne" and Queen Victoria's residence on the Isle of Wight.

For many years to follow, the main use of electronics was for communication purposes. The large economic gains that were available using electronic techniques were responsible for the rapid progress that took place in the electronic discipline.

Telegraphy is a simple but efficient way to send information . . . the electrical contact is closed and opened in a certain time sequence to form a code . . . the Morse code for instance. This contact sets the current flowing in the transmission medium, thereby carrying information. At the receiver the signal is used to close another contact that can either be heard or seen. Today we still use the same concept . . . the telex service, short-wave radio and ship-to-ship links . . . but in general the methods are much more complicated as they incorporate such devices as error detection, multiple channels and automatic output recording.

We will now take a look at the forms of signals that can be used to convey information — leaving the actual circuits used until late in the course when we have developed some mastery of the workings of components and sub-systems.

TWO CLASSES OF SIGNAL — DIGITAL AND ANALOGUE

Let us go back to a simple direct current circuit, as shown in Fig. 3. The person operating the key-switch at the transmitting end can cause the device at the receiving end to operate, thus conveying something to the other person. What he has done, in effect, is to set a current flowing, the magnitude of which is decided by the voltage of the battery supply and the resistance of the indicating device. (The

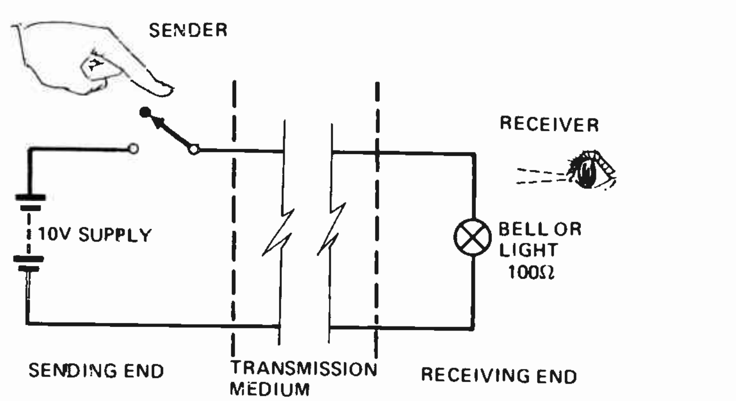


Fig. 3. A simple d.c. signalling circuit using on-off control.

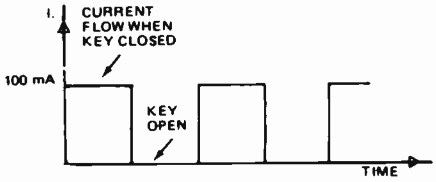


Fig. 4. Amplitude-time graph of current in a switching circuit.

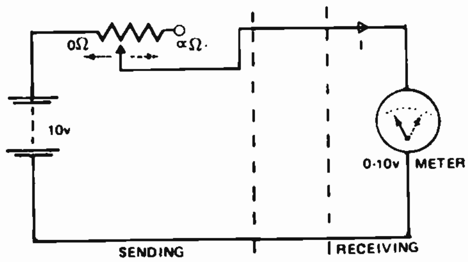


Fig. 5. A simple d.c. signalling circuit using variable current control.

resistance of the cable is assumed to be negligible here, but in practice it must be considered, especially when distances run to many kilometres).

But as shown in Fig. 3, no matter how the switch is closed, it can only provide an ON or OFF action. If the sender repetitively opened and closed the key the current amplitude/time graph would look like that shown in Fig. 4. Note there can be no currents in the circuit between the on and off values.

As this kind of signal has only certain *discrete* values we call it a DIGITAL signal, this word originating from the Latin word for 'finger'. The type of electronic circuit that generates these signals is known as a SWITCHING circuit.

It is convenient here to point out why apparently wasteful resistance is actually so useful in electronics. In the example of Fig. 3. the resistance of the bell or light converts the flowing energy into a useful signalling effect. Without resistance there could be no signal generated at the receiving end. Too little resistance in the device would lead to an enormous current flowing uniformly in the wires; too high a resistance in the device would not provide enough energy to produce the desired indication.

Consider now what happens if we remove the key, replacing it with a variable resistor, as is shown in Fig. 5. Further, at the receiving end we put an indicating volt-meter instead of the bell or light. As the sender varies the *resistance* in the circuit, the current also varies in accordance with Ohms Law. When the variable resistor is set to maximum value, (infinity in our example), no current flows, and so the meter registers zero volts. As the resistance is reduced by the sender the current increases, and the meter reading increases accordingly. Finally, at minimum resistance, current reaches the level at which the meter pointer reads full-scale.

Thus the signal varies *smoothly*, without any evidence of the rapid transitions that we saw in the switching circuit (unless the sender produces them by very fast changes of the resistance). This form of signal is called an ANALOGUE signal.

It is not possible to uniquely define how this signal would look as time varies (as we did for the switching case), for this depends entirely on the sender. If, for example, the resistor is varied uniformly from maximum to zero a ramp signal is produced as shown in Fig 6a. It can also be seen that a sinewave can be generated (Fig 6b) if the resistor is first set to give half-voltage and is then moved back and forth with the appropriate time-resistance relationship. The analogue circuit can be used to produce switching action by very rapid movement, but a digital circuit cannot be used to obtain analogue behaviour (at least not without additional circuits – as we will see later in the course).

The resistance of the circuit plays a vital part in the production of the analogue signals, especially when the value can be made to vary by some means or other. We will see later that the well-known transistor is really little more than a variable resistor – in which the current passing through it is controlled by another current fed to it in another terminal – much in the same fashion as a policeman controlling traffic flow at an intersection in a one-way street.

The next point to consider is how the two forms of signal (digital and analogue) convey information.

INFORMATION IN DIGITAL SIGNALS

Digital signals can only exist at discrete set levels . . . a desk calendar for instance is essentially a digital device. It either *is* Feb 17th – or it isn't. It shows no intermediate stages, such as Feb 17.75th!

By contrast, a conventional watch or clock is an analogue device, in that the passage of time is indicated as a smooth progression of the hands around a dial.

The most basic electronic device for generating a digital signal is a switch. It is either ON or OFF, there are two, and *only* two, possible states.

There are many other devices and circuits (described later) which have only two unique states, and these are known collectively as Binary (meaning two state) devices.

These binary devices form the basis of digital electronics, the digital computer being the most outstanding example, where many thousands or even millions of binary devices are used in combination to perform amazing tasks.

Let us examine how information may be transmitted with a keyed (switched) system such as shown in Fig. 3. Here the light is either ON or OFF. This means that the sender can only signal one piece (we call it a bit) of information at a time . . . Come when the light goes on, etc., the only information that is actually transmitted is that the key is closed. That is, we must *assign a meaning* to this bit of information.

We can however, send the same signal two or more times in sequence and assign meanings to the individual sequences. We can also make our key

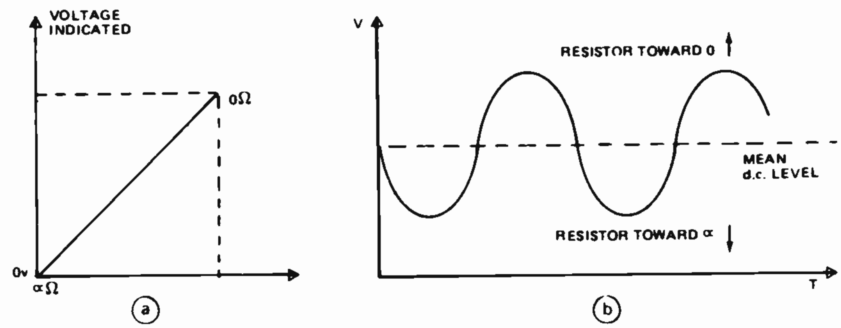


Fig. 6. Circuits with continuously varying current levels are called analogue circuits. (a) A linear ramp (b) A sinewave.

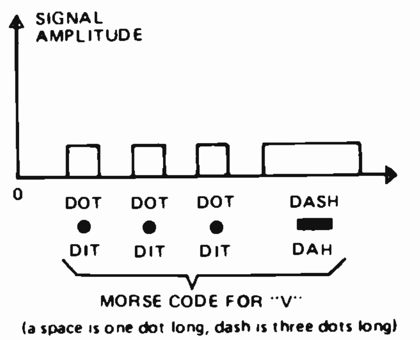


Fig. 7. Amplitude-time graph of Morse code letter V as produced by keying a d.c. circuit.

closures of varying duration.

The first man to construct such an arbitrary code was Samuel Morse in 1837. His code was constructed of agreed sequences of short and long dashes to represent each letter of the alphabet. For example the letter V is represented by dot, dot, dot, dash; its amplitude time graph is illustrated in Fig. 7.

Thus by sending series of such groups we can build up words and hence complete messages in any spoken language. (In computing jargon, each such *group* of bits is called a 'word' even though it may not correspond to any spoken word.)

The Morse code is only one of many possible codes that can be used to transmit information. Many other communication codes are in use, each having unique characteristics most suitable to a particular purpose. Typical examples are the Baudot, ASCII, Selectric and Hollerith codes — to quote just a few in general use.

These codes differ from Morse in that they use groups of pulses (all having the same length) and are based on variations of a fundamental counting system known as the Binary code which we shall now examine.

We normally do all our mathematics (adding, multiplying etc) in a system based on the number 10. For example the number 1285 equals:—

$$\begin{aligned} 1 \times 10^3 &= 1000 \\ +2 \times 10^2 &= 200 \\ +8 \times 10^1 &= 80 \\ +5 \times 10^0 &= 5 \\ &= 1285 \end{aligned}$$

We don't have to count by tens, we can count by two's, eight's, twelve's or any other base number we wish.

Let us now consider how a system with base 2, (a binary system) works.

If we have one switch it has only two possible states — but what happens if we have a second switch? If we let '0' equal switch position OFF, and '1' as switch position ON we can construct a table of the possible combinations as follows.

	4	3	2	1	
SW1				0	0
only				1	1
SW1 +			1	0	2
SW2			1	1	3
SW1 + SW2		1	0	0	4
+SW3		1	0	1	5
		1	1	0	6
		1	1	1	7
SW1 + SW2	1	0	0	0	8
+ SW3 +SW4	1	0	0	1	9
	1	0	1	0	10
	1	0	1	1	11
	1	1	0	0	12
	1	1	0	1	13
	1	1	1	0	14
	1	1	1	1	15

From this we can see that adding a second switch gives us four possible combinations (2^2). Taking this still further, three switches gives us 2^3 = eight combinations, four switches 2^4 = 16 combinations etc. Thus if we were to use six switches a total of $2^6 = 128$ combinations would be possible. We can thus use a group of six bits in a binary code sequence to represent the numbers 0 to 9, all the letters of the alphabet (in both capitals and lower case) plus a number of punctuation marks and other symbols or commands we may wish to transmit.

The length of the code word is thus fixed and the sequential groups of bits (words) are separated by a longer than normal space.

The main differences between the various codes are merely the number of bits in the 'word' and the way in which meanings are assigned to the word.

At first sight these binary codes seem to be a dreadfully slow way to transmit information. But remember electronic switches can open and close millions of times a second, so, in practice, we can send information enormously faster using serial binary codes than can a morse code operator.

Further since all the binary bits are the same length and there is the same number in each we can send each word in parallel.

For example, referring to our table, the figure 8 could be sent on four lines by putting a 'one' on line 4 and 'zeroes' on lines 1, 2 and 3. Thus in this case the transmission rate would be four times faster again. However the use of parallel transmission is impractical (due to the number of lines required) except over short distances, eg, within a computer.

INFORMATION IN ANALOGUE SYSTEMS

Unlike digital signals, that can exist only at discrete levels, the analogue signals can exist at any level between zero and the maximum available.

In theory every minutely different level can be used to represent a specific bit of information, thereby giving us unlimited code capacity at each instant of time. Practice, however, limits the separation between levels that we can reliably detect because Noise (the name given to unwanted disturbing signals) can add or subtract from the signal at each defined level leading us to wrongly interpret the true intended meaning.

In reality then, there is a limited signalling capability in any analogue signal and the capability depends on the level of unwanted signal entering the system. The noisy signal obtained from a temperature measuring thermometer (Fig. 9), illustrates this.

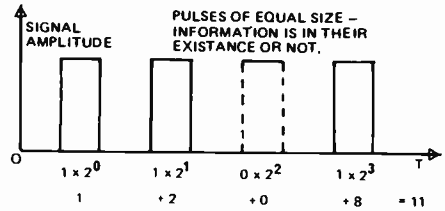


Fig. 8. Sending the number 11 in binary code (digital form of signal).

It tells us how the temperature varied with time but only to a precision limited by the width of the noise superimposed on the record. Within the width of this noise band we cannot say with certainty what the temperature was doing. It may have varied along the mean centre line, it may have varied from the upper to the lower limit or any other way you care to propose. We have no way of knowing what happens when noise swamps the signal.

MODULATION

We saw in Figs. 5 and 6 how the current in the circuit was varied in accordance with the wishes of the operator: (The wiper contact of the resistor was moved with time to accomplish this.) In the parlance of electronics this process is called modulation; the direct current was modulated to produce ac waveforms. The Morse and the binary code are transmitted by modulating a basic dc current.

Looking at the temperature record in Fig. 9 it can be seen that the original signal is modulated by the noise to produce frequencies that probably did not exist in the true temperature signal. Although detrimental in that instance, this process of adding frequencies to others can be used gainfully to transmit information.

If we start with a continuously generated ac signal (instead of the dc case mentioned above) we can modulate the ac waveform in a similar manner by varying its amplitude or its frequency. Let us look at these modulation methods in a little more detail.

Amplitude modulation . . . this is the name given to the process in which the instantaneous amplitude of a constant frequency wavetrain is varied usually in order to convey information. This is shown in Fig. 10.

The original signal is called the CARRIER for it carries the signal

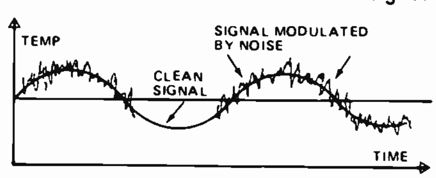


Fig. 9. Resolution of the signal in an analogue circuit is limited by the unwanted noise present with the signal.

information. Amplitude modulation is used extensively in radio transmission, especially the normal broadcasts we are now so familiar with. It is a simple matter to send Morse code over a carrier — the carrier is simply switched on and off to produce short bursts several cycles long. The principle is, however, not confined to radio but finds uses in many other fields of electronics. It is often abbreviated to 'AM'.

Frequency modulation . . . in this type of modulation the amplitude is held constant, and the instantaneous frequency varied instead. The carrier is the same as that for AM to begin with, but after modulation the combined signal has the appearance shown in Fig. 11. (This modulating form is usually known as FM). It is less prone to noise problems than AM but is more expensive to implement, so its use is more restricted than AM systems. No doubt you have heard of FM radio . . . the broadcasting system that uses frequency modulation to transmit the sound signals.

WHY MODULATE AN AC CARRIER?

By now you could well be wondering why we go to all this trouble to modulate an ac signal — it needs a special generator to produce the carrier in the first place and special circuits to recover the signal when received.

Why not use a simple battery-powered dc signal and just vary it with a switch or variable resistor?

To answer this question let us consider the problem of transmitting speech over long distances. As we speak we create pressure waves in the air which another person, reasonably close to us, can detect by means of his pressure sensitive ears — but the distance over which this acoustic communication can take place is strictly limited. How can we transmit a spoken message halfway round the world — or even to the moon?

Of course you know that the means is radio; in a radio transmission we modulate a carrier frequency by the amplitude and frequency of the voice. But how does this technique increase transmission distance? The answer lies in the nature of electromagnetic waves.

ELECTROMAGNETIC WAVES

In a preceding section we told you that when a current flows through a wire there is also an associated magnetic field. In addition where we have two conditions, or charged bodies, insulated from each other and at different potentials, there is an electric field between them.

Thus we can have a magnetic field without an associated electric field and correspondingly an electric field without a magnetic field. However if the fields are *changing it is impossible for either type to exist separately.*

A changing electric-field will produce a magnetic field, and a changing magnetic field will produce an electric field.

This electro-magnetic disturbance, in a similar manner to the ripples caused by a stone thrown into a pond, propogates in all directions.

The remarkable thing about an electromagnetic disturbance is that it propogates at the speed of light and it does not require air, *or any other medium,* for its propagation.

Hence its ability to travel through the vacuum of free space.

As no-one wants to listen to everything that is broadcast, different carrier frequencies are used for different transmission applications.

The carrier frequencies used depend on the specific application, eg, radio, television, amateur radio, radar etc. All use frequencies, appropriate to the type of modulation, in bands allocated by international agreement. Thus AM radio commercial stations use carriers within the range 550 kHz to 1.5 MHz whilst radar may use frequencies in the 1 to 10 GHz region.

In fact communication systems have used electromagnetic radiation with frequencies from 10 kHz for VLF (very low frequency communications with submarines) to light wave frequencies (by using lasers) in the 100 terahertz region.

Do not confuse low frequency electromagnetic radiation (eg 20 kHz) with audio at the same frequency. They are entirely different phenomena. Audio needs a medium such as air and propogates in air at around 334 metres per second. Also note that the speed varies, with the medium. By contrast electromagnetic radiation at 20 kHz propogates at 300 x 10^6 metres per second and does not require a transmission medium.

At the receiving end, special circuitry is used to then recover the modulation impressed upon it.

MULTIPLEXING

Assume that we wished to send four telephone communications over a wire at the same time. The first channel, as shown in Fig. 12, could be sent direct. If we attempted to add the others to the same line the result would be like a party line . . if they all spoke together

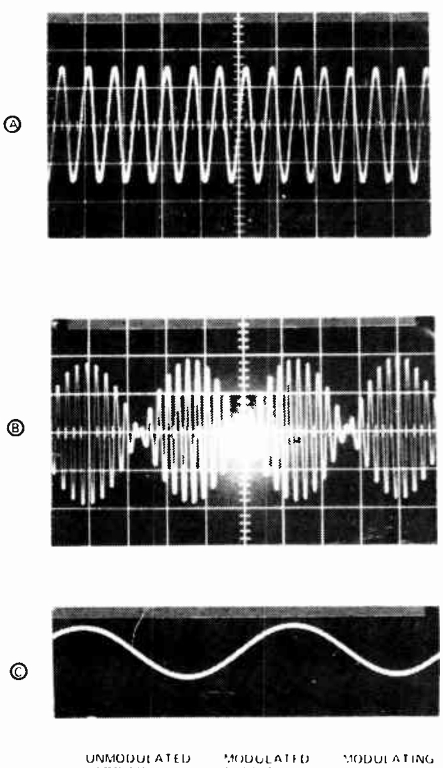


Fig. 10. In amplitude modulation, the amplitude of the a.c. carrier signal is varied, the frequency remaining the same.

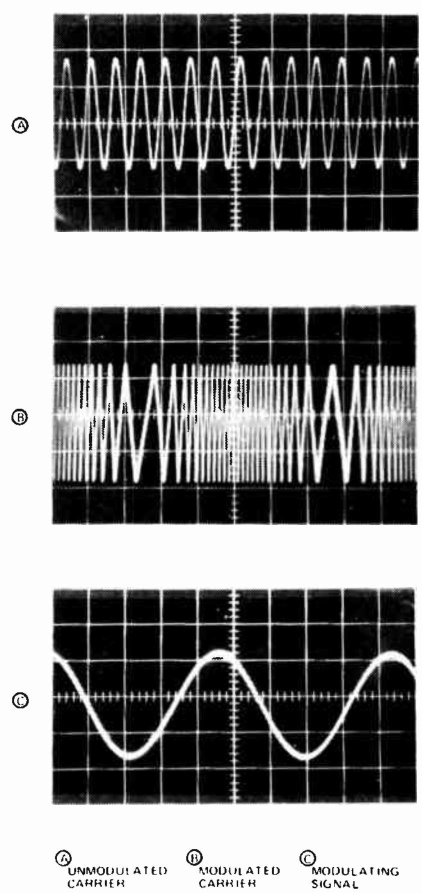


Fig. 11. In frequency modulation, the amplitude of the carrier is held constant, the information signal being used to vary the instantaneous frequency.

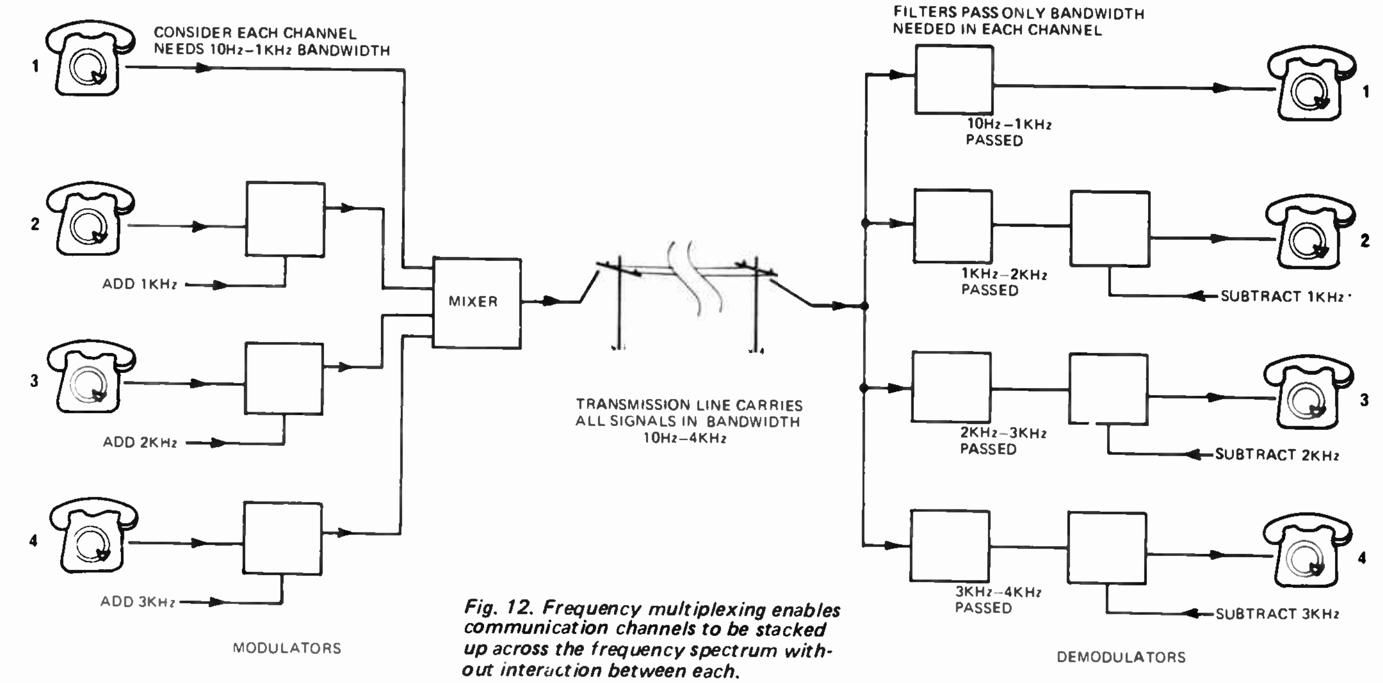


Fig. 12. Frequency multiplexing enables communication channels to be stacked up across the frequency spectrum without interaction between each.

it could become unintelligible and certainly not private. This is overcome by adding the second voice signal to a carrier frequency just higher than any frequency in the first voice channel. This is done by modulation. The other lines are also modified this way placing each voice channel up at a higher frequency than the channel below. Hence the jargon, stacking the channels, for they are being placed across the frequency spectrum, side by side or on top of each other – however you like to visualize it.

For reasonable intelligibility it is necessary to transmit the frequency components of human speech lying, approximately, between 300 Hz and 3300 Hz, i.e. a range of 3000 Hz. This is known as the required BANDWIDTH.

Thus the signal in the interconnecting telephone lines may contain frequencies ranging from the lowest in the first (unmodulated) channel, to the highest in the (modulated) fourth channel. Each channel – as we have seen – requires a *bandwidth* of approximately 3000 Hz, (and it is desirable to separate channels to some extent to prevent overlapping) so that four voice channels will require a total bandwidth of 12 000 Hz (plus channel separation).

In normal telephone line systems however, the number of channels which can be so multiplexed is strictly limited as the total bandwidth that can be handled by a conventional telephone line is seldom much more than about 12 kHz. (Special cables however can handle thousands of channels multiplexed this way).

Having so multiplexed our separate conversations onto one line, it is obviously necessary to separate them at the other end. That is to *demultiplex* them. This is done by using the special electronic circuits known as filters to select the narrow band of frequencies that constituted each individual carrier. Each channel must then have the modulation recovered from the carrier. This process of demodulation is done for each channel and the recovered audio then fed to the individual telephone subscribers.

Frequency multiplexing is certainly a complicated process for sending information. But it is far less expensive to transmit many information channels this way than it is to keep adding new lines to a global communications system, especially if the lines convey television, or if they run under the sea.

As the electromagnetic frequency spectrum is usable up to at least 10^{14} Hz it will be appreciated that an enormous number of communication channels may be used.

The total frequency spectrum cannot be crammed into any one line of course, but hundreds of channels can be multiplexed onto a microwave link. In the future, laser communication systems may allow thousands of TV channels to be transmitted over a single beam of light.

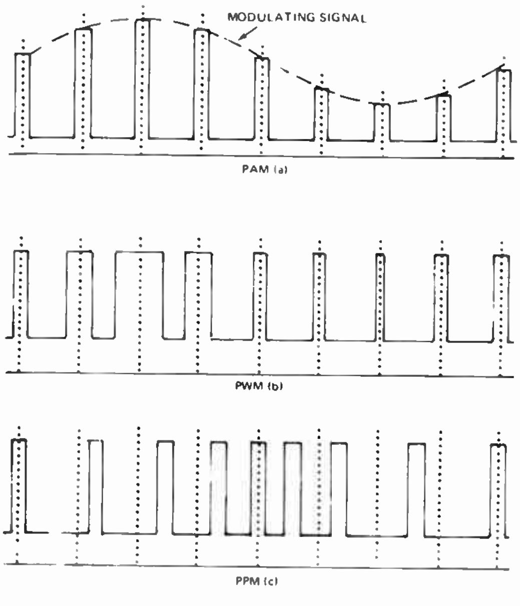


Fig. 13. Modulation of pulse carriers.
(a) Pulse amplitude modulation PAM.
(b) Pulse width modulation PWM.
(c) Pulse position or pulse time modulation PPM or PTM.

Multiplexing is also used in other contexts. Basically its purpose is to send more signals over fewer circuits, as in the multiplexed communication channel discussed above. The reasons, however, can be different.

A first variation would be as found in the digital circuitry of some solid-state displays. Interconnections in digital circuits can be numerous. Multiplexing helps overcome some of the difficulties that arise. More is given about this application in Chapter 26 (Code Converters and Display Systems).

A second variation of the use of multiplexing is found in fast computers. The internal computational speed of many computers far exceeds the ability of one operator to feed in data and instructions. Output devices are also slower than the computer can feed. Time-sharing provides the computer with a constant stream of work to keep it fully occupied. A group of operators works into the computer through a time-sharing terminal. Each works as though no other operator is using the machine — unless the demand exceeds the system capability.

Time-division multiplexing requires a continuous signal to be broken up into a sampled form. This may completely alter the character of the signal to be transmitted. Link capabilities in multiplexed systems need to be quite different from those needed for transmitting the original signal in its original form.

MODULATION IN DIGITAL SYSTEMS

The principle of modulation is not restricted to analogue signal transmission systems, but can also be applied to digital communication links — those that use on-off signals. Again there exist a number of ways by which a basic digital wave-train can be modified to represent signal data that comes in original analogue form.

In our discussion of the transmission of digital codes we saw how a train of pulses could be used to represent all the characters needed for the transmission of messages. In that case, see Fig. 8, the presence, or not, of pulses at certain times indicates the meaning assigned to each data word. It can be seen that the continuously transmitted signal would look like a square-wave train that has pulses missing now and then. This is in fact how signals are sent around inside a digital computer . . . a square-wave train is generated continuously with a generator (called the clock) and the instruction circuits (called logic) decide which pulses are to be there and which are not, depending upon the code value to be sent.

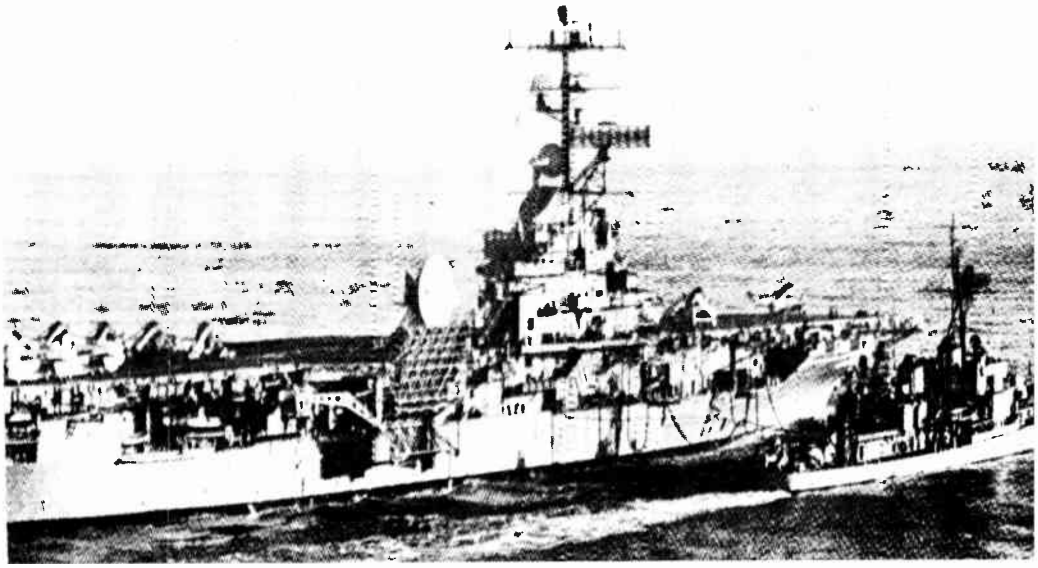


Fig. 14. Communication links sending measurement and control data are known as telemetry systems. In the picture above the round white dish antenna receives data from television cameras in space.

As said before this method of coding may be used for wire or radio communications and in general is known as Pulse Code Modulation — PCM.

Note that the pulses are not always there and each pulse does not carry complete instantaneous information: It takes the addition of several pulse positions to build up the 'word'. Consider now, the case where pulses are continually generated, as before, but where we actually alter some characteristic of each pulse, in a way that is proportional to the analogue input signal to be sent on the data link.

Pulse amplitude modulation (PAM) is one method . . . in this case a square-wave signal has its instantaneous amplitude (of each pulse) varied in accordance with the amplitude of the analogue input signal, as is shown in Fig. 13a.

Pulse width modulation (PWM) is another method . . . the width of each pulse is varied, the height being held constant and the frequency remaining the same, as is illustrated in Fig. 13b.

Pulse position modulation (PPM) . . . also called pulse time modulation (PTM) . . . the remaining available variable is modulated in this. Pulses are identical in height and width but their position is varied within each carrier pulse period. The frequency remains the same, as is shown in Fig. 13c.

The advantage of using pulses to modulate the carrier is mainly that the pulses can be restored (with digital circuits) to their original form as the communication link progresses (in what are called repeaters) thus retaining the quality of the original signal throughout the transmission. This means less errors are sent for in electronic hardware it is possible to maintain timing accuracy far more easily than amplitude accuracy.

Digital modulation methods are used extensively for data transmission in scientific experiments and equipment. One example is the satellite data link (Fig. 14). Digital modulation may also be used in normal voice communications by converting the analogue voice and signals into digital form.

Cost, and the extent to which external unwanted noise is able to upset the system usually decides which method is to be preferred for both analogue and digital systems can convey information. Another factor that may influence the decision is the form of the data when derived, or the form needed on receipt. If already in digital form direct transmission of binary code probably would be preferred to converting it to an analogue equivalent and then back again at the receiving end. (Systems sending data derived from sensors are generally called TELEMETRY systems).

This has been a *systems* introduction to the transmission of data. To understand the design of the black boxes we will need to study many circuits before the operation of the many methods is to be fully comprehended. It has, however, added another significant chapter to the understanding of electronics at the systems level. The principles and terms encountered here are constantly used in electronics.

ELECTRONICS -in practice

THEORY in this part has been concerned with information transmission – so here are two quite different exercises that will illustrate the concepts.

The first is for those who wish to use the Morse Code, perhaps because they ultimately wish to become one of the world-wide group of radio amateurs, who spend their leisure time building and using radio transmitters and receivers for communication across the globe. A working knowledge of the code also enables you to listen in to the many signals sent in the shortwave band of the radio set.

The first thing that must be done is to learn the dot-dash code sequences used to represent each letter of the alphabet. They are given in Fig. 15. Also given are the accepted codes for punctuation and procedure. When you feel you know the code try yourself out by listening to the signals found as you scan the dial of a good short-wave radio set. Don't be too discouraged at first. It requires much practice to reach the speeds used by trained operators.

Another, and probably better way, to learn morse is to build yourself a code practice rig such as that shown in Fig. 17.

The relay coil (from the earlier experiment on the L.D.R. and resistors) is wired in series with the operator's sending key and also in series with a normally closed contact of the relay. When the key contact is made, the relay closes, pulling the contact open, thus opening the circuit releasing the armature. This closes the contact pulling the relay on . . . and so on. The process continuously produces oscillations causing the relay to become a buzzer.

When the relay oscillates in this manner its electrical coil generates a high ac voltage with each swing of the armature. This voltage is sufficient to produce a loud noise in an earpiece.

The 0.005μF capacitor smooths away the harmful peaks of this generated voltage, safeguarding the earpiece. Components in the filter section can be varied as you please to obtain the sort of sound you desire.

The resistor placed in series with the relay coil is provided to reduce the supply voltage to a safe working level. In this mode of use the relay can be provided with a little more than its normal voltage, for the coil is not energised all of the time. Select the resistor that gives the sound you like.

This circuit gives the ardent enthusiast the chance to practice without disturbing the peace of those around (as would be the case if a normal buzzer were used). If necessary the relay can be put into a sound-proof enclosure.

A	di-dah	S	di-di-dit
B	dah-di-di-dit	T	dah
C	dah-di-dah-dit	U	di-di-dah
D	dah-di-dit	V	di-di-di-dah
E	dit	W	di-dah-dah
F	di-di-dah-dit	X	dah-di-di-dah
G	dah-dah-dit	Y	dah-di-dah-dah
H	di-di-di-dit	Z	dah-dah-di-dit
I	di-dit	1	di-dah-dah-dah-dah
J	di-dah-dah-dah	2	di-di-dah-dah-dah
K	dah-di-dah	3	di-di-di-dah-dah
L	di-dah-di-dit	4	di-di-di-di-dah
M	dah-dah	5	di-di-di-di-dit
N	dah-dit	6	dah-di-di-di-dit
O	dah-dah-dah	7	dah-dah-di-di-dit
P	di-dah-dah-dit	8	dah-dah-dah-di-dit
Q	dah-dah-di-dah	9	dah-dah-dah-dah-dit
R	di-dah-dit	0	dah-dah-dah-dah-dah

Punctuation

Frequently employed in Amateur Radio

Question Mark	di-di-dah-dah-di-dit
Full Stop	di-dah-di-dah-di-dah
Comma*	dah-dah-di-di-dah-dah

*Often used to indicate exclamation mark.

Procedure Signals

Stroke	dah-di-di-dah-dit
Break sign (=)	dah-di-di-di-dah
End of Message (+ or $\overline{\text{AR}}$)	di-dah-di-dah-dit
End of Work ($\overline{\text{SK}}$)	di-di-di-dah-di-dah
Wait ($\overline{\text{AS}}$)	di-dah-di-di-dit
Preliminary call ($\overline{\text{CT}}$)	dah-di-dah-di-dah
Error	di-di-di-di-di-di-di-dit
Invitation to transmit (K)	dah-di-dah
$\overline{\text{KN}}$	dah-di-dah-dah-dit

* * *

One dah should be equal to three di's (dit's).
The space between parts of the same letter should be equal to one di (dit).
The space between two letters should be equal to three di (dit's).
The space between two words should be equal to from five to seven di's (dit's).

Fig. 15. Sound equivalents of the Morse Code for letters, numbers, punctuation and procedure signals.

MODULATION

With a working knowledge of dc and ac circuits, signal waveforms, circuit construction, a few basic components and proper use of the multimeter, it is now quite realistic for us to tackle a more ambitious experiment. This time, then, the aim is to build an entire system for sending signals by amplitude modulation and, with a few changes, by frequency modulation. In building this system you will develop expertise in mechanical construction, use some new components, and at the same time gain direct practical experience with dc and ac signals.

To some, the project may appear formidable but, remember, even the most complicated systems break down into familiar sub-system black-boxes which are each made up of basic components in basic circuits.

This particular project will be in several stages with more added in each part of the course.

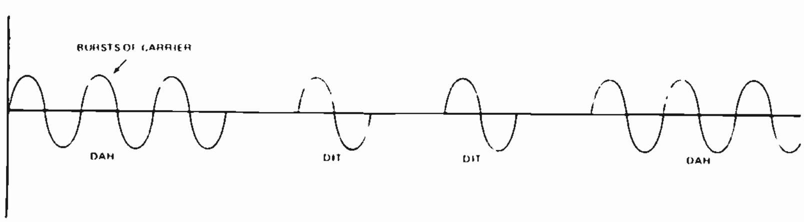


Fig. 16. Morse code sent by radio, travels as bursts of carrier signal. Here is the amplitude-time graph of the letter X.

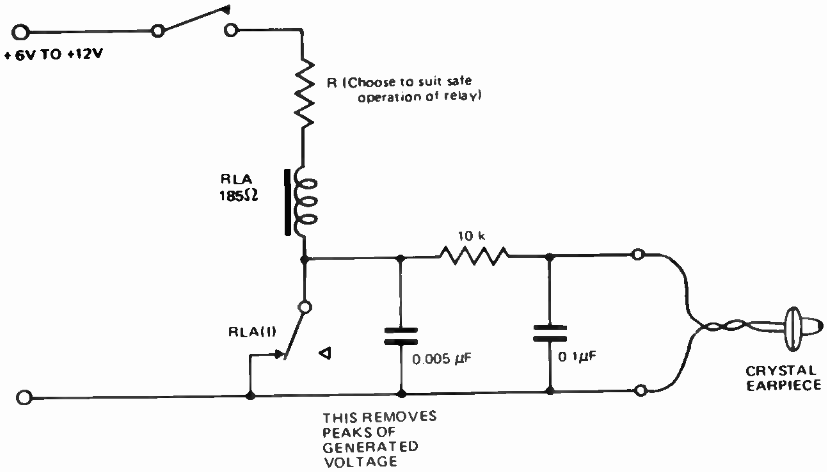


Fig. 17. A simple code practice circuit. The relay acts as a buzzer providing high-voltage ac signals in the earpiece.

We now give details of the construction of the generator and modulating devices. In the next part of the course, the rest of the two systems will be described.

These have been designed to use the components already specified in this course plus a few inexpensive additional parts which, in turn, will find use again as the basis of other circuits later on.

THE LOW-FREQUENCY AC WAVEFORM GENERATOR

In normal transmission systems the carrier frequency signal varies with time at a rate faster than the eye can follow – telephony over open wire trunk lines, for example, operates at kilohertz frequencies; telephony over microwave links is at hundreds of megahertz. Consequently if we attempted to build even the first type of system little could be learnt unless you had an oscilloscope at your disposal to look at the waveforms at various places in the system. To overcome this problem the exercise project described here has been designed to operate with a carrier frequency of around 1 Hz or less, enabling most of the waveforms to be studied by observing the movements of the pointer on a multimeter. The real system works in a similar manner – but at a far faster rate.

The generator is not made entirely from electronic components but uses mechanical motion cyclically to vary the light intensity falling on a light-dependant-resistor (LDR). The schematic diagram is given in Fig.20.

This method gives us a good opportunity to build a composite device in which mechanical, optical and electronic parts are involved. Many sophisticated instruments use all three disciplines together like this.

The 6 Vdc motor (see Fig.21) is of the type found in cheap electrically driven toys or in model trains. Using a rubber band as a belt and a large wheel, the output shaft speed is reduced to rotate at around 10 revs. per second maximum.

Very little precision is needed in the construction – it consists of bent

SYSTEMS USING AM AND FM TRANSMISSION

To send multiple channels of analogue signal information (that is, the continuously varying kind which can have all values between certain limits) over a common line we have seen that we need to generate an ac carrier signal and then modulate this in some way adding the original signal frequencies to the carrier. The combined signal is then fed into the common transmission line and sent to the receiving terminal where the modulated signal is demodulated in order to recover the original data.

The system diagram, given in Fig. 18, portrays this procedure for AM working. A power supply provides dc energy to a 'box' that uses this power to produce a steady dc signal having constant amplitude and frequency: (the carrier). The amplitude of the carrier is modulated, the basic frequency remaining the same, before being fed to the transmission line. This set of input equipment is repeated for each channel to be sent but with each carrier different from the others. Only one input is shown in detail. Upon receipt at the receiving end the varying amplitude of the carrier is used to produce a dc voltage that is proportional to the amplitude of the input signal at the sending end.

The FM system looks somewhat similar – see Fig. 19. The differences are that the modulating input is now derived by altering the frequency of the carrier, leaving its amplitude constant. Demodulation, in this case, (methods vary considerably) is achieved by deriving pulses of uniform amplitude and width (constant energy, therefore) and with one being generated for each cycle of carrier signal. These are smoothed by an averaging circuit. The more pulses received in a given time the higher the average signal level, hence a varying frequency signal produces a varying demodulated output. In this way the dc output at the receiving end is proportional to the dc input that is modulating the carrier at the sending end.

When we get to building the complete FM stage (in the next part) we will also add a dc channel to the transmission wire to demonstrate how both the light circuit (shown in Fig. 19) and the data circuit will operate over the same line at the same time without interfering with each other. Thus we will clearly demonstrate the concept of frequency multiplexing wherein signal channels are 'stacked'.

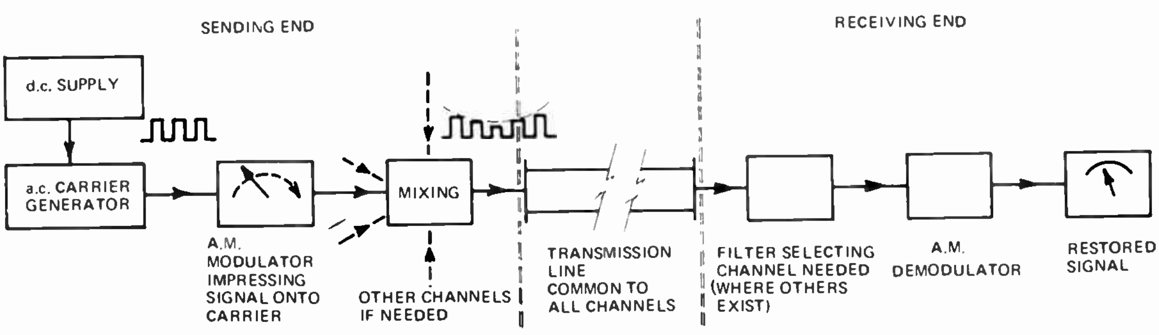


Fig. 18. Block schematic of pulse amplitude modulated communication link.

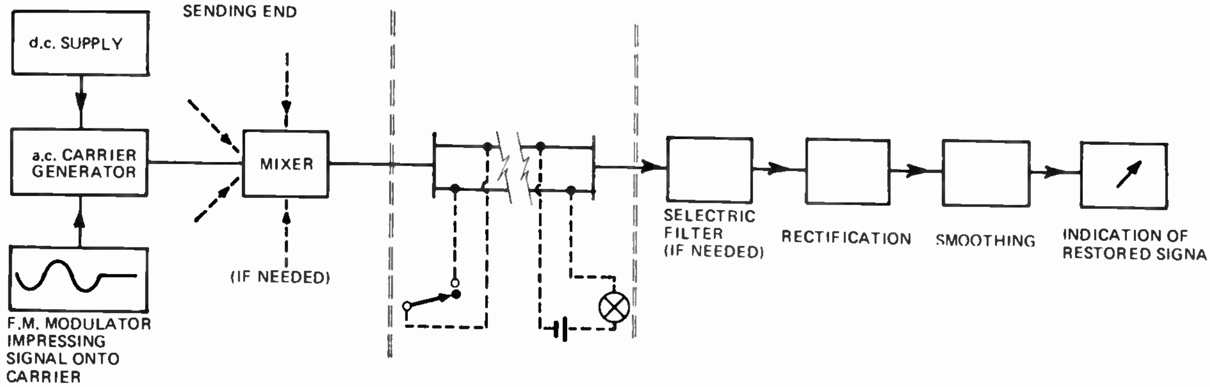


Fig. 19. Block schematic of a pulse frequency modulated link.

pieces of aluminium or brass strip, tag strips and a suitable wheel. We made the prototype in less than three hours using only elementary hand-tools.

Almost any dc motor will do provided its operating voltage suits the power supply. A resistor can be added in series with smaller voltage motors in order to drop the maximum value applied to the motor.

If you have finished the unregulated power-supply, this can be used along with a series dropping resistor.

Pushed onto the shaft is a disk of opaque material (plastic sheet or thick card) cut as shown. This shutters the LDR from the light source as it rotates.

The shape shown will produce square-waves and is usually called a chopper disk. It is the easiest to make. We leave it to you to design other shapes for producing, say, sinewaves, sawtooths, or pulses of higher frequency than the rotational speed of the shaft. (If you drill about 50 holes around the disk and run it at say 10 rps it will generate a 600 Hz signal – this will easily power a loudspeaker, producing a constant tone). The circuits given are designed for use with square-waves.

The speed of the motor, and hence the *frequency* of the signal, is varied by varying the voltage to the motor. Note how the variable resistor is used here as a "potentiometer" giving an output voltage between the wiper and one end which varies smoothly from 0 – 6 V.

The LDR has a 150 Ω resistor in series with it; this enables a voltage swing to be obtained as the light intensity changes — a practical example of how resistors enable voltages to be produced as needed.

By redrawing the LDR and resistor circuit you will see that they form a kind of potentiometer with the mid-point acting as the wiper connection. Ohms law explains why the voltage varies as the LDR changes resistance.

The output of this low-frequency generator is taken from the leads of the LDR and this, in turn, feeds a second potentiometer. By varying the potentiometer the *amplitude* of the carrier is altered from zero to the maximum available (*approximately*) 5V. Hence position of its shaft decides the level of the AM carrier – it is, therefore, our AM modulating signal input. (Although we use a mechanical potentiometer here, the unit could be replaced with, say an LDR and resistor which would enable us to modulate the carrier with the varying intensity of a light input). This potentiometer forms the AM modulating block shown in Fig.18.

If we leave the AM control set to maximum, variations in motor speed will produce frequency modulation of the carrier. The motor speed potentiometer is, therefore, a kind of FM modulator – the FM modulator box shown in Fig.19. As we are using square waves the two cases are more correctly called pulse amplitude modulation (PAM) and pulse frequency modulation (PFM). The working concept remains the same if the chopper disk is cut to produce sine waves.

TESTING THE SYSTEM

Little can go wrong with the generator, the only problem arising

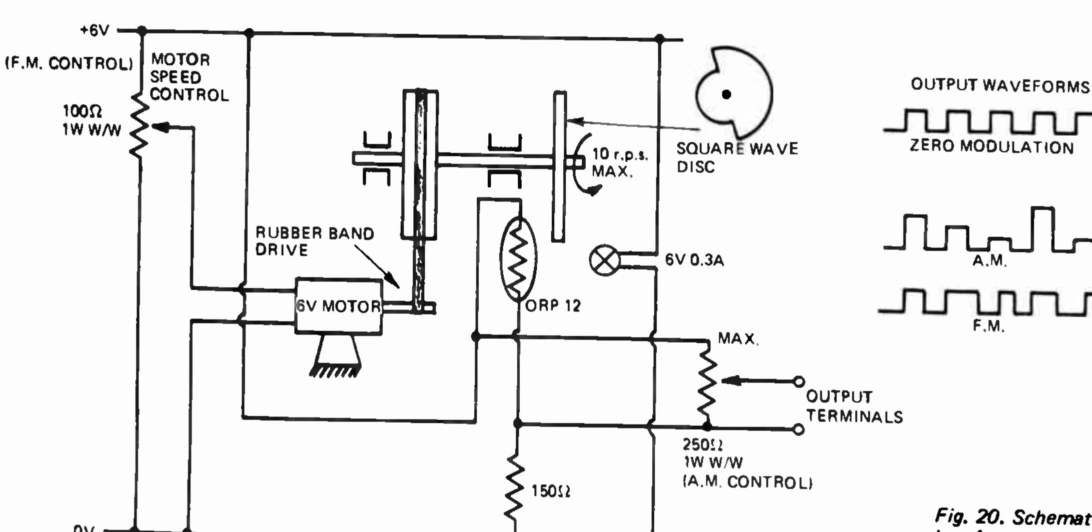


Fig. 20. Schematic layout of the mechanical low-frequency waveform generator used in the telemetry systems discussed.

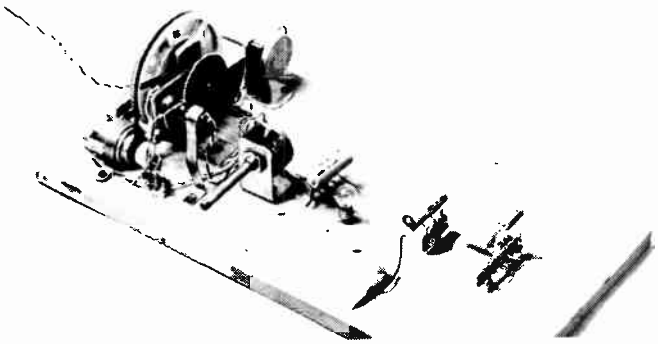


Fig. 21. The generator is simple to build and uses commonly available components.

might be overheating. If in doubt use your multimeter to check currents, voltages and resistances and from these calculate the watts dissipated ($W = I^2 R = EI = E^2/R$). The LDR must not be made to dissipate more than 150 mW. The motors and potentiometers are robust in the sense that they will heat slowly with little risk of rapid burnout. All components in this experiment are robust so there is little chance of failure. Later when we reach a level of using transistors and the like, things are very different. Try to develop a careful approach — "think first — connect the power last" when you are completely satisfied that the design is within safe limits. Once you have the unit running connect the multimeter (set to dc volts) to the output terminals. If the generator is turning at about 1-2 seconds per revolution the meter movement will follow the waveform closely. Next, study the effect on the output of variations in the AM and the FM modulation controls. Note particularly that when the frequency is varied the amplitude remains constant and vice versa.

Most meter units are damped to respond to a full-scale swing in about 1-2 seconds. Consequently waveform frequencies higher than 1 Hz will tend to be averaged — the meter acts as a smoothing energy store. The degree of smoothing increases as the frequency rises.

It will be seen that the ac waveform switches between 0V and +V and is, therefore, not a true ac signal. It 'sits' on a dc level of 0V.

It is possible to observe the frequency changes by listening to the pull-in "clicks" of a relay placed across the output terminals. A lamp circuit can be wired across the relay contacts to enable you to see the varying frequency more clearly than with the multimeter. (The relay specified for previous exercises is satisfactory).

In the next part we will assume the generator is working. We will then add on a dummy transmission line and demodulation circuits for each case. It would be wise to build the generator on one end of a board leaving as much room again for the remainder of the exercise.

REFERENCES

Much can be learned about components, terminology and parts availability if you have a few of the inexpensive up-to-date catalogues listing electronic equipment. The following are worth having in your collection:

Dick Smith — Electronics Catalogue
Applied Technology Catalogue
M.S. Components Catalogue
Radio Spares Catalogue (extensive, but not always available).

6 Capacitance and inductance – the basic concepts

SO FAR we have concentrated on circuits which are composed of pure resistance only.

The relationship between current, voltage and resistance in such circuits follows Ohm's Law E = IR.

Ohm's Law is a linear relationship, that is, if we graph current against voltage for a particular value of resistance we obtain a straight line.

In practical circuits however we find that, when the current is varying with time, Ohm's Law does not adequately explain all the things that happen. This is because resistance is not the only basic property that an electronic circuit has, and the response of a circuit to a varying signal may be far from linear.

All electronic circuits have two further basic properties (other than resistance) which are evident only whilst current and voltage are changing. These two properties are known as INDUCTANCE and CAPACITANCE. They are of extreme importance in electronics.

Both of these phenomena enable energy to be stored in a circuit. An inductance stores energy in a magnetic field. Capacitance stores energy in an electrostatic field.

In this part of the course we will study the nature of capacitance and inductance, how components having given quantities of these properties are constructed, and additionally how all three basic components behave when subjected to transient signals.

THE TRANSIENT

The word transient describes comparatively short-lived conditions that may exist in a circuit when some form of electrical disturbance is applied – a sudden voltage change for instance.

The transient behaviour of the resistors, capacitors and inductors is fundamental to numerous electronic techniques – as we shall see.

Basically the way a transient signal is modified by a component depends on the energy storage capability of that component. Previously we have seen how a switch may be used to turn a source of power on and off causing a current to flow in the load – e.g. a bell or lamp. If however we were to examine how the current in the circuit changes with time for the bell and lamp, we would find that the current does not, in fact, change from zero to maximum instantly, but changes in the manner shown in Fig. 2a and 2b (respectively). The initial part of these

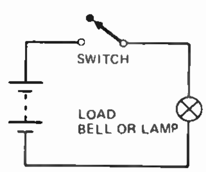
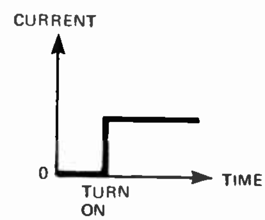


Fig. 1.



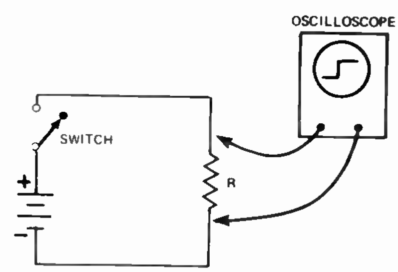


Fig. 3.

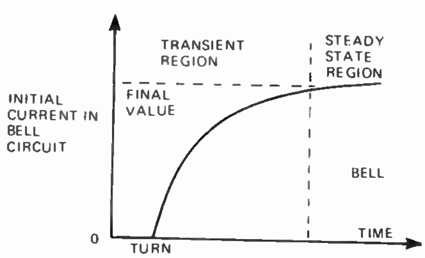


Fig. 2a.

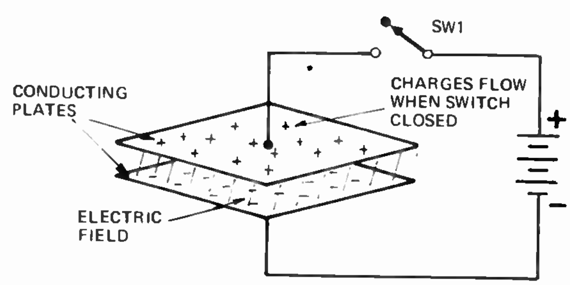


Fig. 4.

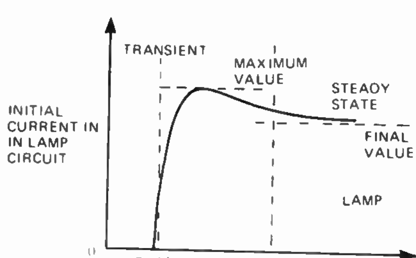


Fig. 2b.

Fig. 1. Basic circuit for switching bell or lamp etc.
Fig. 2a . Typical switch-on characteristic for a bell.
Fig. 2b. Switch-on characteristic for a lamp.
Fig. 3. A step in voltage applied to a resistor will produce a step voltage across. That is the shape is not changed in any way.
Fig. 4. The basic capacitor consists of conductive plates separated by an insulating space or material.

two curves in the transient region. Once the signal has *stabilized* to a constant value it is said to be in the steady state condition.

In the case of the bell the current rise is slowed by the electromagnetic effect of the coil (used to attract the metal armature). The coil of the bell is in fact — an inductor, and energy is stored in the electromagnetic field set up by current flow in the coil.

Current in the lamp, when first switched on, varies for a different reason. In this case the current immediately rises to a maximum value that is determined by the cold resistance of the filament. But as the lamp filament heats up, its resistance increases, causing the current to decrease to a steady-state value which is less than the initial value. Again it is the storage of energy — in heat form in this instance, that produces the initial transient response.

RESISTORS AND TRANSIENTS

We know that pure resistors cannot store electrical energy — they can only dissipate it in the form of heat. Consequently, a properly designed resistor, for use in electronic circuits, will not modify the time behaviour of an electronic signal. A square wave applied via a switch, as mentioned above, will still remain a square wave.

This is illustrated in Fig. 3 which shows the waveform developed across a pure resistor in response to a step-change input. Such a test is called a step-response test.

Any kind of waveform when applied to a resistor (sinewave, sawtooth, pulse etc) will be handled without change of shape in the time dimension.

However although waveshape will not be affected, the amplitude will be, the output value will depend upon the voltage drops occurring in the resistive circuit and is easily calculated using Ohm's Law.

In ordinary electronic circuitry practical resistors behave like pure resistors and hence we need not worry about transient conditions.

But, as the operating frequency of a circuit is raised, to around 50 MHz or beyond, the resistor may have inductance and capacitance as well as its designed resistance value. For such work, special component design techniques must be used to minimize these undesirable (in this case) side-effects.

CAPACITANCE

Every electronic circuit will have current carrying conductors or components running next to other conductors or components. Where adjacent parts of the circuit are operating at different potentials, there will be an electric field between the two parts. This is a basic physical principle.

It takes energy to set up a field and it has been found that the amount of energy stored in an electric field for a given voltage difference is proportional to the area of the adjacent surfaces, and inversely proportional to the distance between them. That is, the capacity to store energy is inherent in the physical arrangement — the larger the conducting surfaces and the smaller the distance between them — the higher the energy storage capacity. Little wonder then that this characteristic of a circuit is known as CAPACITANCE.

In practical circuits the inherent capacitance between components and leads is very small and, unless the circuit is operating at very high frequencies, of little importance. However capacitance is useful, and we can put it to work by building components which have definite known values of capacitance.

The basic construction of such a component is illustrated in Fig. 4. There we can see that a capacitor may be constructed from electrically conductive plates separated by an insulating material. The electrical insulator, known as the DIELECTRIC may be air, oil, insulating paper, plastic film, ceramic layers or special fluids, depending on the properties required.

Assuming the capacitor in Fig. 4. has no initial charge, the voltage source causes charges to flow, the moment the supply switch is closed, creating a charge imbalance between the two plates (negative charges on one side and positive on the other). The charge imbalance will create an electric field in the dielectric between the plates and a voltage across the plates which opposes the source. Thus charges

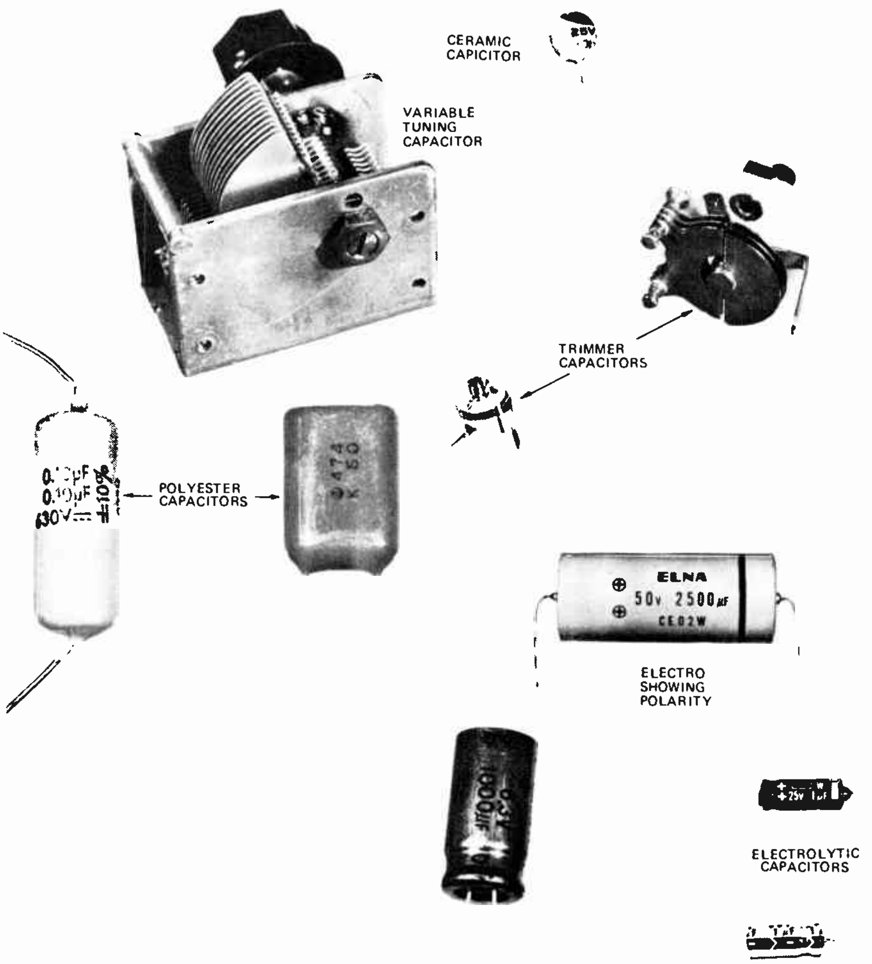


Fig. 5. Typical capacitors as used in electronic circuitry

continue to flow until the voltage across the capacitor and that from the source are equal.

If the supply switch is opened, the charge remains, together with a voltage across the capacitor that depends on the quantity of charge stored. In this condition the capacitor is said to be charged.

Any resistance connected across the plates will provide a path for the charges to flow towards neutrality. Practical dielectric materials and mounting insulators are unable to provide an infinite insulation resistance and hence all capacitors have a finite value of resistance. The charge storage time, therefore, depends upon this resistance which is known as the LEAKAGE RESISTANCE. Quality capacitors have very high leakage resistance and are able to hold charge for many days, but their extra cost is not always warranted.

As mentioned earlier, physical principles tell us that the storage capability of the capacitor is given by:—

$$C = k\frac{A}{d}$$

where k equals dielectric constant
A equals area of plates
d equals distance between plates

The dielectric constant ('k' in the equation above) is a number dependant upon the material used.

For example air is the standard dielectric having a constant of 1. Barium titanate has a dielectric constant of 1143, therefore a capacitor with barium titanate as a dielectric would have 1143 times the capacitance of an air dielectric capacitor with the same plate spacing.

The unit of capacitance is the FARAD and is the value of capacitance by which an applied voltage change of one volt per second produces a current flow of one ampere.

In practice the FARAD is a far larger value than is normally encountered. Instead the smaller sub-units microfarad ($\mu F = 10^{-6}$ F), nanofarad ($nF = 10^{-9}$ F) and picofarad ($pF = 10^{-12}$ F) are most commonly employed.

Physical size limitations prevent the flat plate capacitor, described above, from providing any more than a few picofarads of capacitance.

In order to make components having larger values of capacitance, different construction methods must be employed which utilize as much plate area as possible, have the smallest possible gap and use high dielectric-constant materials as insulation. However, small gaps imply low insulation resistance and large values are only obtained at the expense of increased physical size, and/or, reduced safe working voltage levels.

In the so called solid-dielectric types several manufacturing methods are used — rolls of aluminium foil interleaved with plastic film, layers of deposited materials etc. Knowledge of the actual construction is of little importance to an understanding of electronics however, so we will leave these aspects to the designer. What does concern us is that by using this method of construction, reasonable working voltages are obtainable, but for values of 100μF or more physical size becomes a considerable problem.

A second major class of capacitor, known as electrolytic, provides an answer to the size problem, although they have other disadvantages which prevent them from being a universal replacement.

In the electrolytic capacitor the dielectric layer is produced by electrolytic action (by means of a chemical solution or paste) on the surface of the aluminium foil. By this means very large values are obtainable in reasonable sizes. They suffer however from the fact that they are polarized (meaning that one connection lead must always remain positive with respect to the other). Electrolytic capacitors can be recognized, usually, by the leads being marked with polarity. If the polarity is reversed the capacitor will be damaged — it may even explode!

Electrolytic capacitors must not be used for ac signals unless the signal is biased with a dc level so that polarity of the capacitor is never reversed. (This is not as great a disadvantage as it may seem). Further, the insulation resistance of electrolytics is usually relatively low. However, not withstanding the above disadvantages, electrolytics are used extensively because of their large capacitance per given volume of component.

An assortment of capacitors is shown in Fig. 5 and more illustrations may be seen in the trade catalogues mentioned in the previous article in this series.

As with resistors, capacitors are made in a wide range of values and may be fixed or variable types. In the variable types, those that have a wide capacitance range and have a shaft

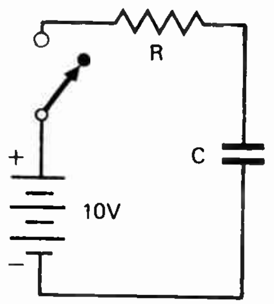


Fig. 6a. Circuit of an RC network.

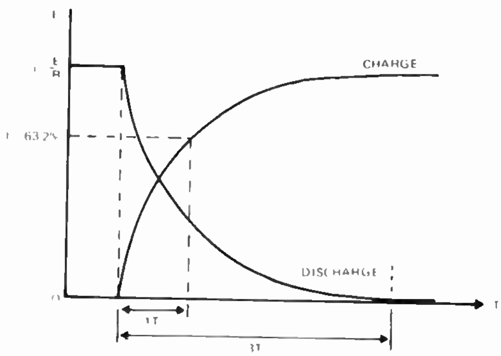


Fig. 6b. The transient behaviour of the RC network of Fig. 6a.

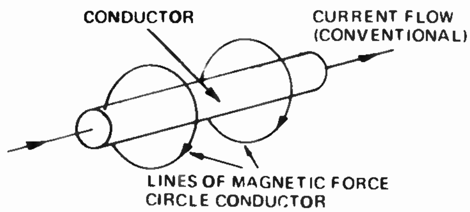


Fig. 7a. The basic inductor is merely a wire with current flowing through it.

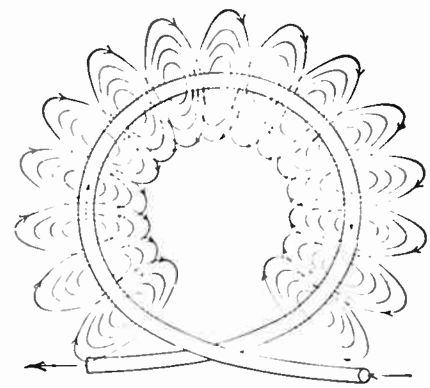


Fig. 7b. The magnetic field about a single loop of wire carrying a current.

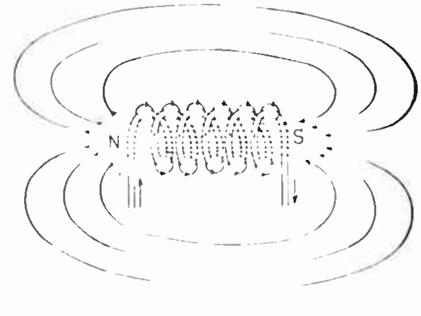


Fig. 7c. The magnetic field about a loosely wound coil of wire carrying a current.

suitable for mounting a knob are known as *tuning* capacitors; those having a screw driver adjustment are known as *trimming* capacitors.

Fixed capacitors are sometimes colour coded and this code is given in Fig. 5. Usually however the value is marked on the capacitor together with its tolerance and working voltage e.g., 100 µF ± 10%, 50 V working.

CAPACITORS AND TRANSIENTS

We previously explained that charges will continue to flow into a capacitor until the voltage across the capacitor equals that of the source.

The important thing to realize about this is that charge flow constitutes a current flow *even though there is no direct electrical path*! Further, current flows only whilst the capacitor is charging or discharging. A little thought will show that if a changing voltage is applied to the capacitor a corresponding change in charge current will occur. Thus if dc is applied, the capacitor, after an initial charge period, will block any further dc current, but *an ac signal will pass through the capacitor*. The great usefulness of a capacitor is therefore in separating various sections of a circuit, as far as dc signals are concerned, but coupling them for ac signals.

The amplitude-time relationship for the charging (and discharging) of capacitors obeys an exponential law (stated simply, an exponential change is one which doubles, or halves, for each unit interval of time) and the

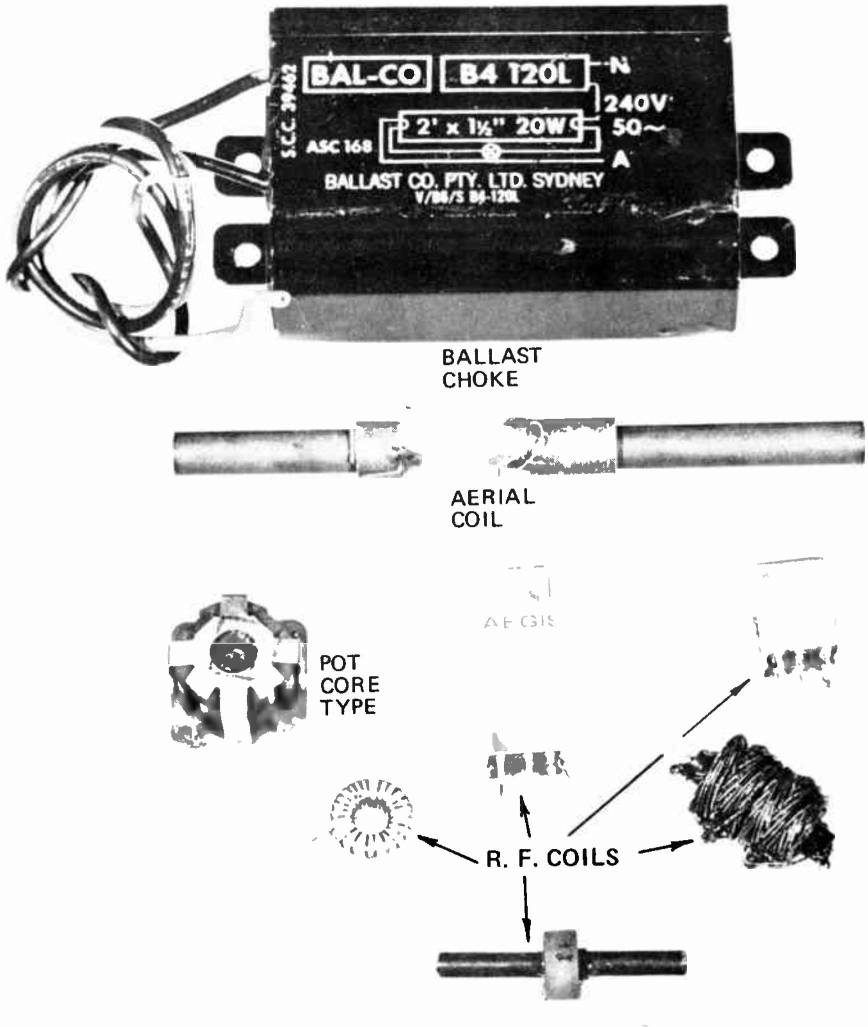


Fig. 8. Typical inductors as used in electronics.

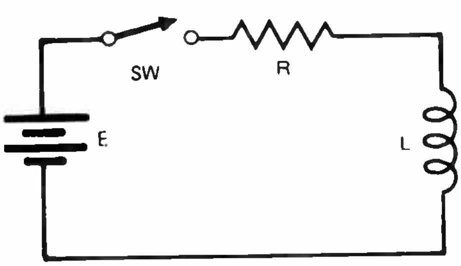


Fig. 9a. The basic resistor and series inductor circuit.

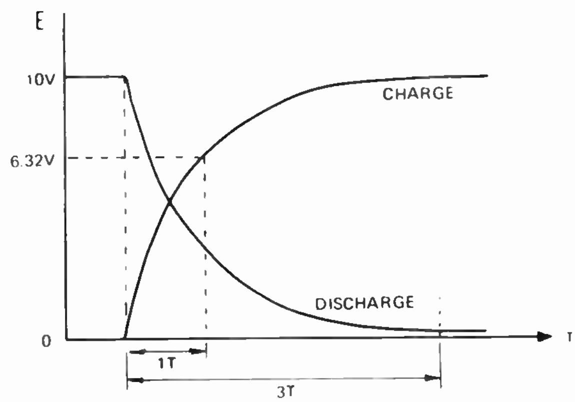


Fig. 9b. The transient characteristics of the series LR circuit.

shape of the charging function is always the same, (see Fig. 6), the only variation being in the scales used. The actual time taken for a capacitor to become fully charged (or discharged) depends on the size of the capacitor (i.e. amount of energy to be stored or released) and upon the amount of series resistance in the charging circuit. Fig. 6a shows a basic charging circuit.

As the charging curve obeys a well defined mathematical law we are able to characterize all charging and discharging operations by what is called the TIME CONSTANT (symbol T or τ).

The TIME CONSTANT is by definition the time taken to charge to 63.2% of the final value, or discharge to 36.8% of the final value. These values are chosen because the time constant is then simply equal to the product of the resistance and capacitance values. That is:–

T = CR

where T = time in seconds

C = capacitance in Farads

R = resistance in Ohms

For example a 1 μF capacitor, charged though a 1 kohm resistor, will reach 63.2% of its final voltage in one millisecond. Note that the actual value of the applied voltage does not alter the relative amount of charge stored in a given time.

A handy rule to remember is that the capacitor is virtually fully charged (an exponential charge never reaches the final value — it merely halves the charge remaining each additional time unit) after a time of 3T.

Similarly when a resistor is placed in parallel with a capacitor that is already charged it will discharge to 36.8% of the final value in accordance with the T = CR rule. As all capacitors have some internal leakage resistance effectively in parallel, they will all become discharged eventually.

The concept of time constant is important when capacitors are used as a means of storing voltages or for smoothing variations on a dc voltage, but more about this later.

INDUCTANCE

Previously, we have briefly mentioned that, when a current flows through a conductor, there is always an associated magnetic field around it, as shown in Fig. 7a.

We can show that the field exists by observing the movements of the needle of a compass when held near it. Again, this is a physical principle for which we have no real explanation — we merely know that it is *there* and have learnt how to make use of it. Thus, as with a capacitor, energy is stored in a field. This time, however, it is a *magnetic* field, not an electrostatic field as with capacitance.

The magnetic field around a simple loop of wire is shown in Fig. 7b. The field is represented by lines of circular form around the wire carrying the current. These lines are called lines of magnetic flux and constitute the magnetic field. The closer the lines are together — the stronger the field.

One way of reinforcing the field is to wind the wire into a coil, as in Fig. 7c. The effect is to concentrate the lines of force through the centre of the coil and thus produce a denser field. The field may be still further concentrated by winding the coil around a soft iron core and by winding several layers of wire. By such means we can produce powerful electromagnets which have many uses (such as in the relay used in an earlier experiment).

If we try to vary the current through the coil we find that the change is resisted, ie, the coil tries to maintain the current at a constant level. This is because the coil generates a voltage, called the back emf, which always opposes the supply voltage change. This effect is known as INDUCTANCE (symbol L) and is only evident when the current tries to change.

The unit of inductance is the HENRY and is defined as that value of inductance which will produce an emf of 1 volt across it when the current is changing at 1 ampere per second.

Thus $e = L \dfrac{di}{dt}$

where e = instantaneous voltage across coil
L = inductance in Henries
$\dfrac{di}{dt}$ = rate of change of current in amperes/second

A single piece of conductor has only a minute amount of associated inductance. It is so small that it may be neglected in circuits operating at low frequencies. However, as with capacitance, inductance is useful, and components may be specially constructed having values of inductance from a few microhenries to tens of Henries, with millihenry and microhenry values being the most commonly used.

The inductance value depends upon the number of turns, and (if used) the iron core. It does not depend on the resistance of the wire. Unfortunately large inductance values, of small physical size, can only be made by using thousands of turns of very fine wire, and hence, the resistance is high. Thus an inductor which is required to carry a large dc current will be bulky and heavy.

Practical inductors, therefore, come in many shapes and sizes and may range from a single piece of wire to a million-turn coil, or a many kilogram-weight unit as used in radio transmitters. Fig. 8. shows a range of units commonly encountered in low power electronics. Variable inductors are also required in some circuits and these may be produced by arranging for, a variable air gap in the iron circuit or, a small slug of ferrite (a type of ferromagnetic material) which can be screwed in or out of the coil, or by using a sliding contact to 'tap off' various parts of the coil.

INDUCTORS AND TRANSIENTS

As said earlier, an inductor will resist any attempt to change the steady-state field conditions. In other words energy being put into, or taken out of, the field experiences a retarding force. This means that the inductor is quite happy to pass a dc current but will oppose any changes in that current. Hence an inductor is useful where it is desired to pass a dc current but block any ac component. This is the reverse of the effect of a capacitor which passes ac and blocks dc.

In a similar fashion to the capacitor, we find that the inductor has a characteristic time constant, in response to a step function input, and the current versus time curve also follows an exponential law.

The time constant for an inductor-resistor circuit is given by:—

$$T = \frac{L}{R}$$

where T = time in seconds
L = inductance in Henries
R = resistance in Ohms

The circuit of an LR network is given in Fig. 9a together with the current versus time behaviour when the switch is first closed. At the instant the switch is closed the value of current is zero but the rate of change of current is very high. Thus the current increases rapidly at first, and then more slowly as it approaches the Ohm's Law value ($\frac{E}{R}$). When the current reaches a steady state value the inductive effect disappears.

Thus we see that in any circuit containing inductance or capacitance *Ohm's Law only applies during direct current (that is steady state) conditions.* Next we will show you how the effects of inductance and capacitance on an ac signal may be calculated. Later we will examine a special kind of inductor known as a transformer. ●

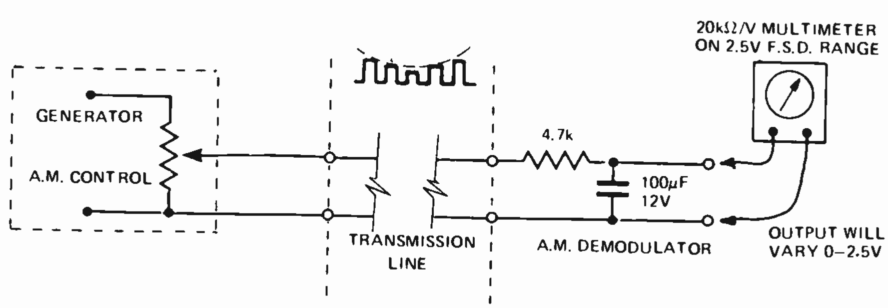


Fig. 10. AM demodulator circuit.

ELECTRONICS -in practice

Continuing the signal transmission project

THE PREVIOUS ARTICLE in this series described how to build a low-frequency mechanical form of signal generator that could be used to examine waveforms and signal transmission.

We now expand this project so that it may be used as the sending end of, firstly, an amplitude modulated (AM) and secondly a frequency modulated (FM) telemetry link. (A telemetry link is one that carries information).

AM DEMODULATION

The generator is connected to the transmission line over which signals are to be sent. The line can be of any practical length providing the resistance of the cable is kept low. Bell wire or twin lighting flex is ideal.

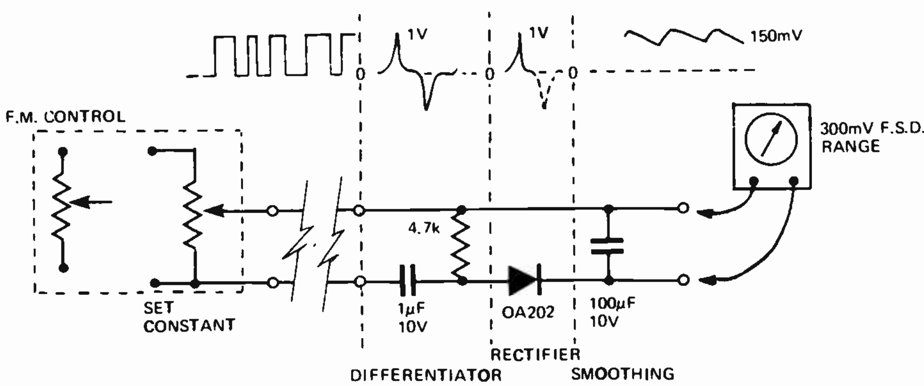


Fig. 11. FM demodulator circuit.

Using the AM control, the received signal will be as shown in Fig. 10. This signal must be processed (demodulated) to regain the original signal — which in this case was the angle of the AM potentiometer input shaft.

Demodulating an AM signal is very easy. All we need to do is just average out the pulses using a smoothing circuit in which the values of the resistor and capacitor have a time-constant chosen to smooth out the 1 Hz carrier frequency but not the lower signal frequencies. The circuit shown in Fig. 10 does just that.

The needle of a multimeter, connected as shown, will rise and fall in sympathy with rotation of the AM potentiometer shaft. If the AM control is left set you should find that a frequency change — induced with the FM control potentiometer — will not alter the output level, showing that the AM link is not affected by changes in frequency.

FM DEMODULATION

This is not as easy to achieve. The requirement is that the demodulating circuit provides energy to the output meter that is proportional to the *frequency* of the signal received, not to its *amplitude* — as was the case with AM.

It is accomplished here using several stages each having a specific purpose — see Fig. 11.

The first stage accepts the square waves as received and produces a pulse of constant width and height (and, therefore, constant energy) at each *transition* of the square wave.

The circuit used is called a *differentiating* circuit for it produces the *differential* of the input signal. This means that *changes* in the signal produce an output signal — but steady-state levels do not. In our circuit, the capacitor is used to let charge through when the signal is varying but not when it is steady.

The choice of time-constant decides the size of the pulses produced.

The second stage has the task of removing the negative-going pulses from the preceding differentiator which produces both positive *and* negative pulses and, unless one or the other is removed, the negative and positive going pulses will effectively cancel out — resulting in zero output.

The pulse-removing stage consists of a half-wave rectifier that lets through positive pulses only.

A more efficient circuit would be a full-wave rectifier as it would let both polarities through, giving twice the energy.

The original signal has now been converted to a train of constant size pulses that occur at a rate depending on the positions of the FM control input shaft. All that remains now is to average these pulses with our, now familiar, capacitor-resistor smoothing circuit.

This circuit is not particularly efficient and only fractional volts are produced at the output. It does, however, incorporate the basic concepts used in many circuits without needing the addition of amplifiers.

The waveforms at each position are given in Fig. 11. If you have an oscilloscope available you will be able to observe them.

This circuit will send shaft *position* information using FM techniques and the output signal is not affected for reasonable variation in AM control.

AC COUPLING

We know that capacitors will pass ac signals, but not steady dc levels. This is easily demonstrated by adding two (electrolytic) capacitors in series with one line as shown in Fig. 12. Provided the values are around those shown, the AM and FM links will work just the same. It is quite satisfactory to increase the size as this improves the coupling but a reduction in size will attenuate the signal.

MULTIPLEXING

So far we have not actually sent more than one signal along the wire at a time. If several generators were available, each operating at a different frequency, it would be quite possible to build a multi-channel link.

A simpler way to demonstrate this instead is to use a dc circuit through the wires as shown in Fig. 13. Operation of the light does not affect the AM or the FM link also working over the same two wires. ●

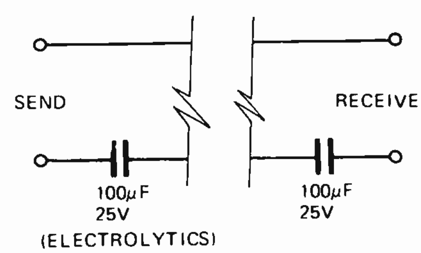


Fig. 12. The addition of capacitors does not stop the transmission of ac signals along the line.

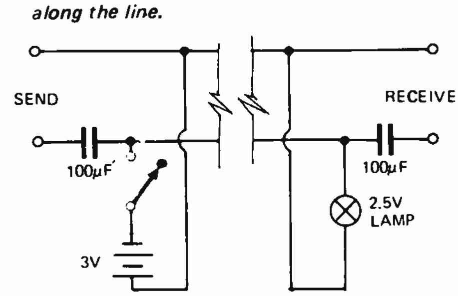


Fig. 13. A dc lamp circuit will operate on the line at the same time as a modulated signal channel thus demonstrating multiplexing.

7 Capacitive and inductive reactance

CAPACITORS and inductors modify any sudden change in voltage that is applied to them. The way in which this happens was explained in the last section.

We will now look at what happens when a square wave is applied to CR and LR circuits that have a time constant that is short compared to the duration of one half cycle of the square wave.

In case (A) of Fig. 1, the capacitive current will be high when the square wave goes positive (time t_1), and will rapidly lessen as the capacitor becomes charged. The same thing will happen when the square wave goes negative (time t_2) except that the capacitor will now supply current back to the supply and hence the current will flow the other way. The current waveform through the network will thus be as shown in waveform 2.

In case (B) the inductor resists a change of current, and hence, the current will initially be low and will increase slowly until the maximum value is reached. When the square wave goes down again the inductor tries to keep the current flowing.

The current will thus gradually decrease until the field of the inductor is zero. Thus the current wave shape will be as shown in waveform 3.

In cases (A) and (B) the output voltage (across the capacitor in (A) and across R2 in B) for both will be as shown in waveform 4. Thus, in terms of voltage-in versus voltage-out, both these arrangements tend to smooth the input waveform. They are therefore known as smoothing (or integrator) circuits.

Now in circuits (C) and (D) we still have the series LR and CR arrangements but this time we have taken the output from across the inductor instead of the resistor as in (B), and from across the resistor in (D) instead of the capacitor as in (A).

The current waveforms will be the same as before as shown in waveforms (5) and (6), and the output voltage in both cases will be as shown in waveform (7).

In these arrangements then, the output is a pulse which corresponds with an input change, and the polarity of the pulse is the same as the direction of the change. These arrangements therefore are known as differentiator circuits (output only when there is a change).

These circuits are fundamental to all electronics and are extensively used to modify an input signal to some different requirement.

As an exercise, see if you can draw the waveforms generated when the time constant of the network is firstly one tenth of the time t, t_1, t_2, and secondly, ten times t, t_1, t_2. You will obtain some interesting results. The time constant in the waveforms given is about one fifth of t_1 to t_2.

We move on now to consider the behaviour of the three basic passive circuit components (R, L and C) when they are excited by a continuous sinewave signal. Our discussion will be restricted to sinewave signals at present — as these are the most basic kind.

RESISTORS AND AC SIGNALS

As resistors cannot store electrical energy, they cannot *alone* affect the time characteristics of a signal. They will however, change the amplitude of the signal if connected to form a voltage-divider network such as is shown in Fig.2. In this example the original 10 V peak to peak sinewave is attenuated to provide an output of 5 V p.p. The attenuation is easily calculated in such cases, for Ohms law (previously used in dc circuits in this course), applies equally as well to ac signals when the circuit is built entirely of resistors — or devices that are effectively resistors. We say such circuits are purely resistive.

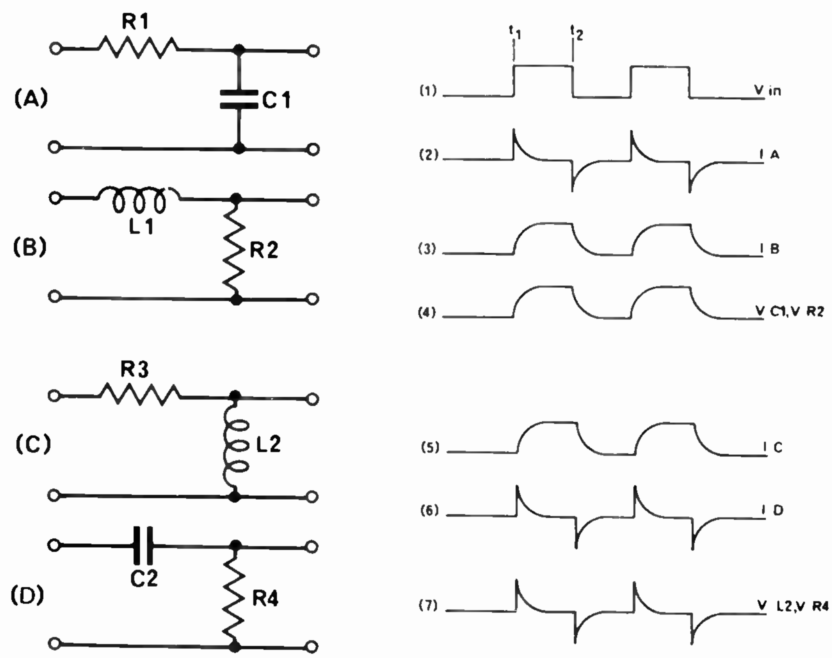


Fig.1. The basic LR and CR networks and their response to a square-wave having a half-cycle duration short in comparison to the network time constant.

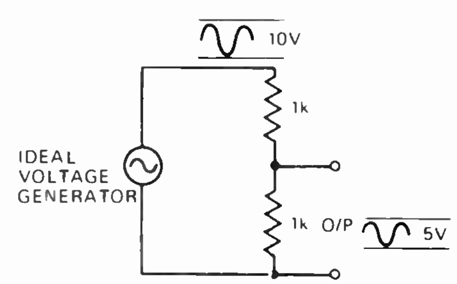


Fig.2. Resistors above cannot change waveshape, they can only reduce the amplitude.

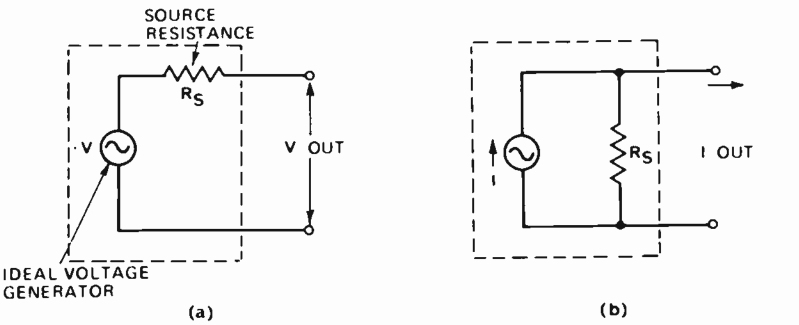


Fig.3a. Equivalent circuit of a practical voltage generator. (3b). Equivalent circuit of a practical current generator.

Note that the voltage-divider cannot provide signals of greater amplitude than the input. This may seem an obvious statement but we will see in the next part that storage elements, when connected in certain ways, can, in fact, magnify the voltage.

INTERNAL RESISTANCE OF SUPPLIES

In Fig.2 the two resistors forming the voltage divider are obvious. In many other cases they are not so easily recognised.

The ideal or perfect source of voltage has no internal resistance and is represented as shown in Fig.2. But in practice all sources have internal resistance so our equivalent circuit is more realistically the ideal voltage generator together with a series-connected internal or source resistance — represented in Fig.3 by Rs. The source resistance Rs is, in fact, the value that is measured (or would be) looking back into the source, and this applies no matter how complicated the power supply is.

Mostly we tend to think in terms of voltages when seeking an understanding of circuitry. But it is sometimes more convenient to make use of currents instead. The perfect current source, again representable as a black box, provides constant value of current regardless of load value. However, practical current generators always have shunt resistance — that seen looking back into the black box.

As with voltage sources we must tolerate resistive losses, this time as a shunt that diverts current from the load. The equivalent circuit of a practical current generator is shown in Fig. 3b.

With a few exceptions, we can regard voltage and current sources as purely resistive devices comprising a perfect lossless generator and a suitably connected source resistor: it is however, not possible to separate the two.

LOADING OF SUPPLIES

In Fig.4 a resistor is connected across a voltage supply. It is clearly forming a voltage-divider chain with the internal supply resistance. In this example the output voltage, will be attenuated to half of that value provided by the supply in the unloaded condition.

The internal resistance of a supply is a vital parameter if the voltage is to hold up and remain constant as the load is changed. A varying load condition imposed on the supply (such as occurs in, say, a hi-fi system as the loudness demand varies would continually alter the system voltage supplying the circuits, with subsequent loss of correct operation.

A little thought reveals that a source resistance very much lower than the minimum load resistance reduces this attenuation effect; at least ten-times less is a good yardstick. Simple power supplies, like that specified earlier as a project, are unable to provide an adequately low internal resistance and, therefore, suffer from loading effects. Special stabilised supplies, although complicated in construction, simply provide a more ideal source by effectively providing a much reduced source resistance — tenths of ohms downward.

Similarly the current supply ideally should provide a constant current with load demand variation, but the inherent shunt resistor diverts the current from the load. The stabilised current supply, therefore, is designed to reduce this effect to a minimum.

It is possible, then, to regard voltage and current sources as black boxes with an internal resistance — in other words as resistive circuits. When studying the coupling of black boxes (stage-to-stage in a circuit or complete sub-system to the next) the preceding one is regarded as the source of voltage or current as is preferred — and the loading effect of that following is easily found from the above reasoning. Fig. 5 illustrates this: it is quite similar to the problem of meter loading discussed in Part 3.

In summary then, when the equivalent resistance of stages has been assessed, or measured, the coupling or loading effect is easily calculated using Ohms law. This concept applies to both dc and ac signals if the circuit is purely resistive.

The voltages and currents flowing in inductors and capacitors can also be handled this way if we use a simple calculation (discussed next) to obtain their effective resistance before using the various circuit laws.

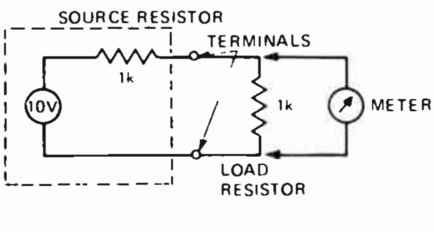


Fig.4. The source resistance, and the load resistance together make a voltage divider. Thus the voltage delivered to the load is less than the maximum available from the generator, and depends on load current.

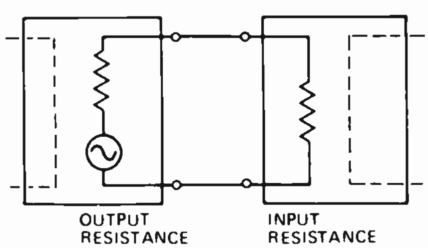


Fig.5. When coupling black boxes, the loading effects must be taken into account. Thus the output and input resistances must be taken into account.

THE CAPACITOR, INDUCTOR AND AC SIGNALS

We have seen how the storage of energy in the electric field of a capacitor, or in the magnetic field of an inductor, modifies the nature of a transient signal impressed across them. It can be said that the capacitor or inductor opposes the transient and tries to prevent its transmission.

If the applied signal is continuously varying from positive to negative — that is, it is an ac signal — it is, in effect, providing a continuous train of transients. We would, therefore, expect storage components to attenuate ac signals in some way. And this in fact is what they do.

INDUCTORS

As just pointed out an inductance opposes sinusoidal current flow. It does not change the time character of the waveform but does reduce the amplitude. The effective resistance is calculated from the formula.

$$X_L = 2\pi f L$$

where X_L is the inductive *reactance,*

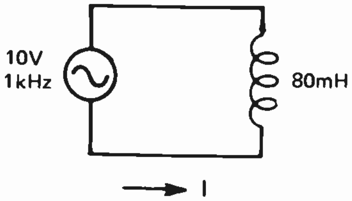


Fig. 6. Current in this circuit is limited by the reactance of the inductor.

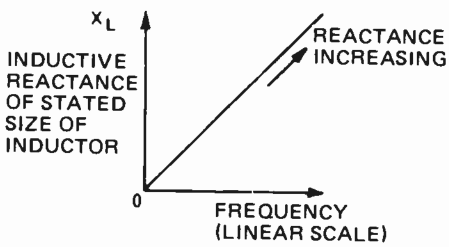


Fig. 7. Variation of inductive reactance with frequency.

the frequency of the sinewave signal and L the inductance in Henries.

Reactance is a term used to describe this particular kind of circuit opposition. It is sometimes called *apparent* resistance. The effect must not be confused with that of pure resistance, for reactance, although limiting current flow and producing voltage drops, *does not dissipate energy*.

The term $2\pi f$ is often replaced by a single symbol ω (Greek omega).

Thus $X_L = 2\pi f L = \omega L$.

A simple exercise illustrates how the formula is used. In Fig.6 an 80 mH inductor is energised by a 10 V rms, 1 kHz source. We wish to calculate the current flowing in the loop.

The formula gives the reactance as

$$X_L = 2\pi \times 10^3 \times 80/10^3 = 500 \text{ ohms.}$$

Knowing the effective resistance to such a signal we can now apply Ohms law to obtain the current

$$I = \frac{V}{X_L} = \frac{10}{500} = 20 \text{ mA}$$

Study of the X_L formula shows that it is frequency dependent so the current will be different if the frequency is changed. For example, if in this example we alter the frequency to 10 kHz, X_L increases to 1000 ohms and the current falls to 10 mA. The frequency effect can be portrayed graphically – see Fig. 7 – showing that X_L increases linearly with increase in frequency.

Practical inductors are made of wire – hence they have resistance as well as reactance. This resistance will deter

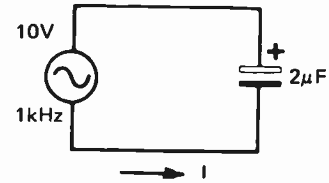


Fig. 8. Current in this circuit is limited by the reactance of the capacitor.

the current flowing by a small amount. So do not be disturbed if calculations do not exactly agree with any measurements made of such circuits. As our background develops further we will see how to take this into account. For the moment it is sufficient to say that – *the resistance value and reactance value CANNOT be directly added to obtain the total resistance.*

CAPACITORS

Having seen how inductors behave when a sinewave is applied to them, we would expect a somewhat similar pattern of behaviour to occur with capacitors. Capacitive reactance X_c is calculated from the expression

$$X_c = \frac{1}{2\pi f C}$$

where C is in Farads and the other terms are as in the inductive reactance formula given above.

In Fig.8 the capacitive reactance is

$$X_C = 1/2\pi \times 10^3 \times 2 \times 10^{-6} = 80 \text{ ohms}$$

and the current I is $V/X_C = 10/80 =$ 125 mA.

This time if the frequency is raised to 10 kHz, X_C becomes 40 ohms and the current *rises* to 250 mA. Thus as the frequency rises the capacitive reactance falls whereas the inductive reactance rises. Put another way, at very high frequencies the capacitor may be considered as a low-resistance link, the inductor on the other hand is a low resistance link only at dc.

Fig.9 shows the variation of X_C with frequency. Note that in contrast with the frequency versus reactance characteristic of the inductor, that for a capacitor is not linear, but hyperbolic.

Although the calculation of X_L and X_c is straightforward, it can become tedious when many values are to be found. To ease this task, a reactance chart may be used from which the reactance at any frequency may be directly determined. A reactance chart is included for your future reference on page 58.

INVALUABLE ELEMENTS

Compared with the simplicity of dc circuits it might seem that the introduction of ac signals makes unnecessary complications. But now we are in a position to see how much of electronic technique is, in fact, based on ac methods.

In an earlier part of this series we saw how signals can be multiplexed onto a common communication channel if ac forms were used. The system design to accomplish this needs circuit techniques that can separate frequencies into individual channels. That is where the inductor and

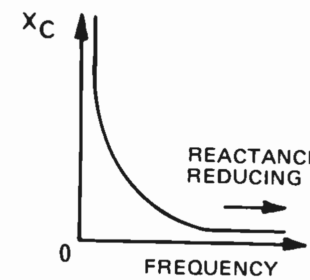


Fig. 9. Variation of capacitive reactance with frequency.

capacitor are of value, for the signal magnitude passed by them depends upon the frequency of the signal applied to them. Using combinations of both components we are able to produce frequency selective circuits that let selected frequency signals through without loss, whilst attenuating those lying either side of the chosen frequency. This is the way in which radio tuners separate the desired programme from all the others picked up by the antenna.

In the power supply project the capacitor was used to smooth out pulsations of the rectified waveform. Thus the capacitor may be seen to provide us with a means of averaging varying signals.

These examples illustrate why it is absolutely essential to have a solid grounding in the behaviour of inductors and capacitors. Like Ohms Law, a knowledge of reactance is absolutely essential. Take time to make sure you understand it thoroughly.

COUPLING BLACK BOXES

Two basic methods may be used to couple circuits together. These are ac coupling and dc coupling.

Where the signal from black box (A) Fig.10 is to be coupled into black box (B), we must first examine the signal to see what frequency range it covers. If it extends down to zero, that is dc, then direct coupling must be used.

In this method the output of (A) is simply joined by means of a wire link or a resistor, to the input of (B).

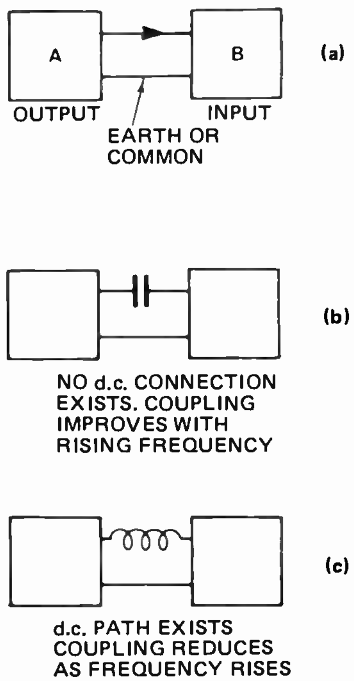
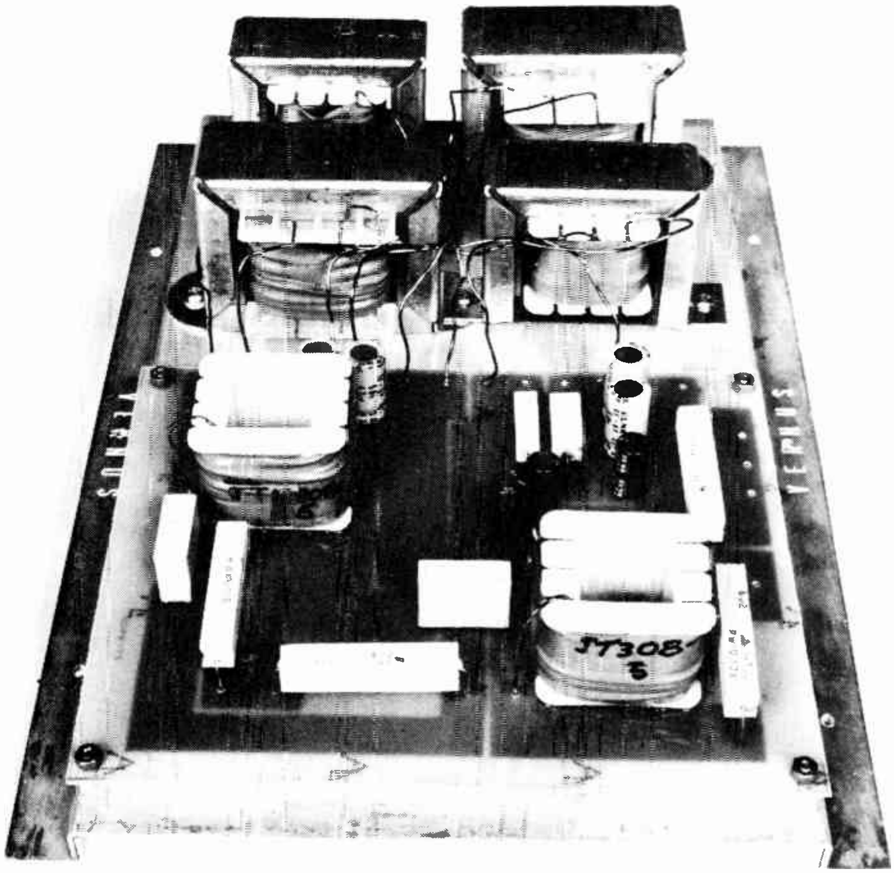


Fig. 10. Coupling methods.
(a) direct
(b) capacitive
(c) inductive



Inductors and capacitors are used to form this Philips loudspeaker crossover network.

However if the mean dc voltages at the output of (A) and the input of (B) are different, a current will flow between them, which may upset the operation of either or both. Thus where dc coupling is required, the black boxes must be designed so that the dc voltage at the output of (A) is the same as that at the input to (B).

Where it is not necessary to operate down to dc, an ac link may be used. This usually takes the form of a series capacitor which blocks dc (and thus allows the dc operating points of (A) output and (B) input to be different) but offers negligible impedance to the ac signal.

It is only necessary to use a capacitor in one lead, in order to block dc, the other can be left as a direct coupling. The capacitor is nearly always wired into the non-grounded (non-earthed) lead.

Ac coupling may however change the signal that is being transferred from A to B.

A signal containing many frequencies – a square-wave, for example – may arrive at B as seen earlier, with shape changed and possibly its amplitude reduced. This is because the various sinusoidal waveforms that compose the signal are each attenuated by differing amounts (for X_C varies with frequency). The net result is a new wave shape. The extent to which the shape is changed depends upon the frequencies present and the value of the capacitor.

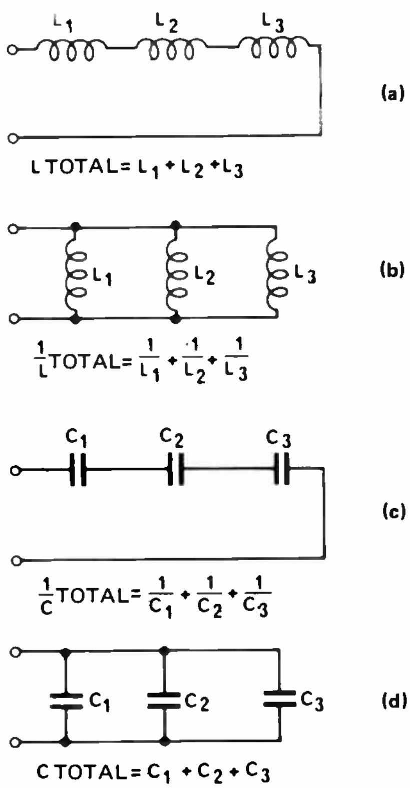


Fig. 11. Rules for combining series and parallel inductors or capacitors to an equivalent single value.

By suitable choice of components this effect may be minimized. Referring back to Fig. 11 on LR and CR networks, if circuit D is used to couple circuits, and the time constant is chosen to be long in comparison with a half cycle of the signal, then the signal will be little changed by the network. Do some sums on this for yourself and see what effects different time constants have. For example, assuming that a circuit has an input resistance of 10 k ohm (R4) what value of C2 would be required to pass a 20 Hz square wave without too much change in shape?

A single frequency sinusoidal signal passes with its shape unchanged. Its amplitude, however, will be altered in accordance with the reactance of the capacitor at that frequency, and any other resistances in the circuit that go with the capacitor to form a voltage-divider chain.

Zero frequency dc signals, as said before, will not pass at all, for the capacitor has no direct coupled path – it only "passes" current when the charges are moving. The capacitor, therefore, provides us with a means to *block* dc whilst allowing ac to flow. This means the steady-state dc voltage level at A can be quite different from that at B yet there is no danger in

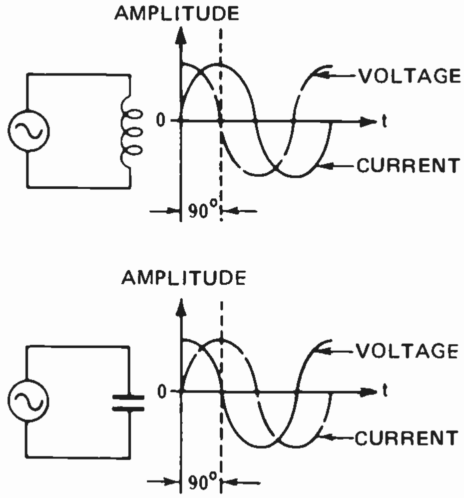


Fig. 12. Amplitude – time graphs for voltage and current in an inductor, and in a capacitor.

connecting them together provided a capacitor is in series.

The higher the excitation frequency the lower is the capacitive reactance. The capacitive coupling therefore, becomes lower in effective resistance as the frequency rises. By appropriate choice of component value (relative to the circuit resistance) it is possible to provide a coupling that is as good as a direct lossless link yet still blocks dc.

Inductors, see Fig. 10c, have the opposite effect (reactance increasing with frequency whilst providing a good dc path). Thus they are commonly used where it is desired to allow dc current flow whilst blocking ac signals. They are also extensively used in combination circuits which separate a signal, or group of signals, from all others. More about this in the next part of the course.

Inductors, then provide increasing coupling resistance with frequency increase and do not block dc. In this role they are able to smooth out fluctuations in a signal: the higher frequencies are attenuated more than lower frequencies. Inductors are often used in this role in which case they are termed chokes.

STORAGE COMPONENTS IN SERIES AND PARALLEL

It is often necessary to calculate the combined effect of inductors or capacitors when they are wired in series or in parallel. The discussion of this section applies only to connections having only inductors, or only capacitors. We will see later that combinations of the two provide vastly different behaviour.

Inductors – the total inductance of series-connected inductors is equal to the sum of each – refer to Fig. 11a.

The total inductance of paralleled inductors obeys the reciprocal law found with paralleled resistors – refer to Fig. 11b. Inductors, then, follow the laws of resistors in this respect. It might be helpful if you remember that inductors in series provide a "bigger" inductor. These rules apply *only when the magnetic fields of each are not interacting.*

Capacitors – these follow the same law but in reverse. Series capacitors obey the reciprocal law, paralleled capacitors obey the additive law – refer Fig. 11c and 11d. An easy way to remember this is that paralleled capacitors effectively increase the plate area thus increasing the capacitor size.

These rules are used to calculate the total component value. From this it is easy to obtain the total reactance as though only one component existed.

But do remember *the laws apply only to groups of similar components.*

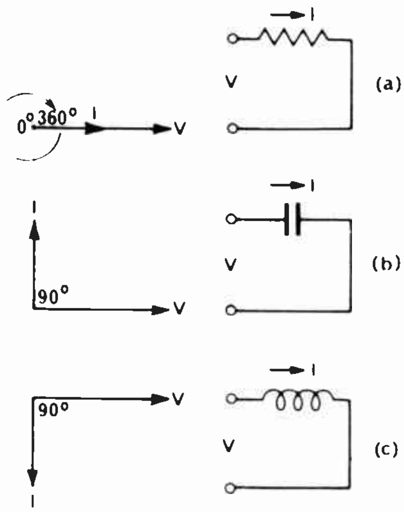


Fig. 13. Vector (or phasor) diagrams for resistance, capacitance and inductance.

If you have a circuit with many capacitors and many inductors the rules work to reduce series or parallel groups of either component to an equivalent, but may not be applied to combinations of different component types.

PHASE RELATIONSHIPS

The voltage developed across an inductor reaches it maximum value when the rate of change of current passing through it is at maximum. This occurs in sinewaves when the current magnitude is zero.

We, therefore, have two distinct components of the signal to consider – current and voltage. They are both sinewaves but they pass through their various levels at different times. A good way to comprehend this is to draw a small piece of the amplitude-time graph of each (as in Fig. 12a). It is clear that the current curve reaches its maximum 90° (or one quarter of a full cycle) behind the driving voltage. In electrical jargon we say the current *lags* the voltage by 90°. (We do not say it leads by 270°: it would only complicate the issue).

The phase effect is opposite with capacitors, for maximum current flows when the charging rate is maximum; this occurs when the applied voltage is zero. Again then, the current and voltage are not in phase and the current leads the voltage by 90° as shown in Fig. 12b.

VECTORS

In order to assess the total effect of combinations of resistors, capacitors

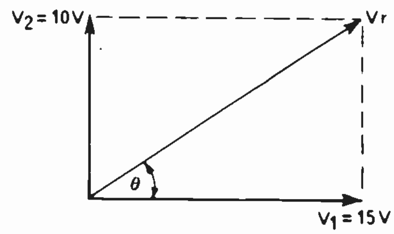


Fig. 14. Graphical additions of vectors.

and inductors, we need to combine their respective signals, making allowance for the different phase shifts in each case.

Special mathematics (called *complex algebra*) can be used to calculate the resultant effect but it is an approach devoid of intuitive feeling for what is happening. It also needs special training to understand it.

Instead we can manage quite well using a purely graphical method in which the length of a line is used to represent longitude of the voltage or current and the *direction* of the line to represent phase. These lines are called vectors (or phasors) to differentiate them from normal lines in which direction is unimportant.

As a sinusoidal signal repeats continously, there is no need to draw each sinewave and add them step by step to see the combined total – this does work but is completely unnecessary. We, in fact, disregard the cyclic changes in instantaneous amplitude and represent the rms or peak value of the signal amplitude by the length of a line. This line is drawn in a certain direction, related to its phase difference. Fig. 13a is a vector diagram for the voltage and current in a purely resistive circuit: the current and voltage are in phase so each vector lies along the same direction. The right-hand, horizontal position shown is always taken as zero phase angle.

The vector diagrams of the basic capacitive and basic inductive loop are also given in Fig.13b and 13c. The 90° phase difference between voltage and current results in the current vector being 90° around from the datum.

When interpreting such diagrams, convention says that the observer moves around the diagram in a clockwise direction — it is wrong to rotate the diagram past the observer.

The reference vector is chosen to be the circuit parameter that is common to each component — current in a series circuit, voltage in a parallel circuit.

VECTOR ADDITION

If two or more compatible signals (eg, a pair of voltages, or a pair of currents, but *not* a voltage and a current) exist with a phase difference they *must* be added as vectors — it is wrong to add their amplitudes directly unless they are in-phase.

Referring to Fig. 14, the vector diagram shows two voltages V_1 and V_2 where V_2 leads V_1 by 90°. They are scaled to represent 10 V and 15 V respectively. The net resultant of the two is not 10 plus 15 because they are not in phase.

The correct sum is, instead, found graphically by drawing lines (at 90° at each axis) out to their intersection point V_r. The distance from the origin to the intersection is the resultant voltage. The resultant phase angle is given as the angle θ also shown in Fig. 14. If they are 180° out of phase they can be arithmetically subtracted.

RESISTANCE AND INDUCTANCE IN SERIES

When resistors are used with inductors or capacitors, the signal across them is similarly involved with phasing problems. To find out what happens requires vector addition of voltages and currents.

The vector diagrams of a resistor and inductor in series is given in Fig. 15. The reference signal is current (common to all components) and the voltage developed across the resistor is in-phase with the current. That across the inductor, however, leads the current by 90° (it leads rather than lags this time as our reference is now current — be careful about which leads or lags what). The parallelogram has been completed to give V_r and θ.

In practice, inductors always possess measurable resistance so the phase angle of the practical inductor never quite reaches 90°. Ignoring this though, when the phase angle is 90° in these diagrams (as can be reasonably assumed for inductors and capacitors) we do not need to draw the vector diagram but, instead, make use of the rules of right-angle triangles to calculate the unknowns. The Pythagoras rule tells us that

$$V_r^2 = V_R^2 + V_L^2$$

from which we can show that the apparent resistance of two combined elements (called the *impedance* Z) is given by

$Z^2 = R^2 + X_L^2$ for inductors.

ie. $Z = \sqrt{R^2 + X_L^2}$

Hence the impedance of a circuit containing inductors is only calculable if the frequency is stated (since X_L is frequency dependent).

The phase angle in degrees is found from the trigonometric formula

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

RESISTANCE AND CAPACITANCE IN SERIES

These are treated in the same way as inductors and resistors giving the same Z (impedance) formula but where X_L is now X_C. Here is an example of how impedance is used to determine currents and voltages in a circuit fed from ac.

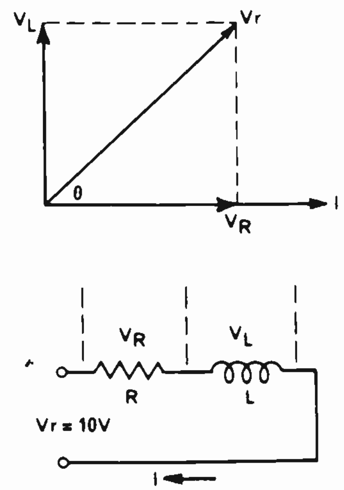


Fig. 15. Vector diagram of current and voltages in a series RL circuit.

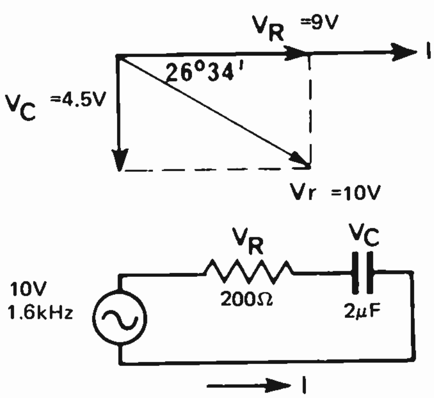


Fig. 16. An RC circuit and its vector diagram.

In Fig. 16 a 200 ohm resistor is in series with a 2 μF capacitor. It is required to decide what voltage is across each component and what value phase angle is seen at the terminals.

Knowing the frequency, first calculate X_C (it will be about 50 ohms). The impedance is then found as

$$Z = \sqrt{R^2 + X_C^2} = 206 \text{ ohms.}$$

The phase angle is given by $\tan \theta = \frac{X_C}{R} = 0.249$, and the tables give the angle as 13.96°. Finally we reason out that the phase angle is leading.

The current flowing in the series loop is found from Ohms law but here we take impedance as the total circuit resistance.

$$I = V/Z = 10/206 = 48.5 \text{ mA.}$$

Ohms law can now be applied (ignoring that we have vector quantities in the current, for this is now allowed for) to arrive at the voltage across each element. Across the resistor will be

48.5 mA x 200 Ω = 9.71 V, and across the capacitor

48.5 mA x 49.7 Ω = 2.41 V.

Note that these do not add up to 10 V and that the sum is always more than the source.

Finally, as a check, it is sound practice to draw a scaled vector diagram. This should agree with your figures. This is done in Fig. 16.

Practical capacitors can be made closer to the ideal than inductors so in most capacitor circuits we do not need to make allowances for their internal resistance.

Q-FACTOR

Practical inductors possess both storage and dissipative capabilities at the same time. As they are intended to *store* energy not waste it, it is useful to form a criterion to express their goodness.

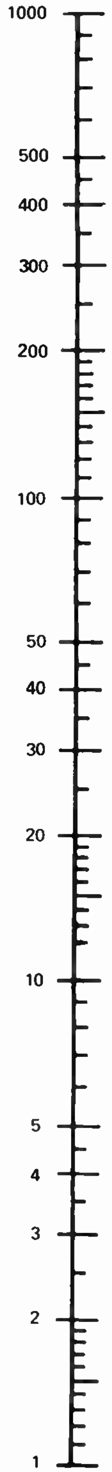
A perfect inductor has no resistance, only reactance. The ratio of these two (for a particular frequency, therefore) is a measure of quality. This ratio $\frac{X_L}{R}$ is called the quality factor or simply the Q-factor (or Q) of the coil.

Practical coils can reach Q's of several hundred. To go higher, special circuits have been developed in which the effective Q is many times higher.

Use of the Q-factor is not confined to inductors. It is used to express the quality of many types of energy storage systems — capacitors, mechanical systems, acoustic cavities, etc.

In the next part of this series we will look at inductors and capacitors in parallel circuits with resistors and then study what happens when both are used together. There we will see some quite astounding effects.

eti REACTANCE CHART



TO USE

Lay a ruler between any two parameters and read off the third eg. to find the reactance of a 10 mH choke at 2000 Hz. Lay a ruler between the two known parameters and read the answer (120 ohms) on scale A.

Note also that 0.7 μF has the same reactance and thus a 0.7 μF capacitor and a 10 mH choke will resonate at 2000 Hz. Resonance may only be read using scale A (values of inductance).

If inductance scales B or C are used, the corresponding reactance scale B or C must also be used.

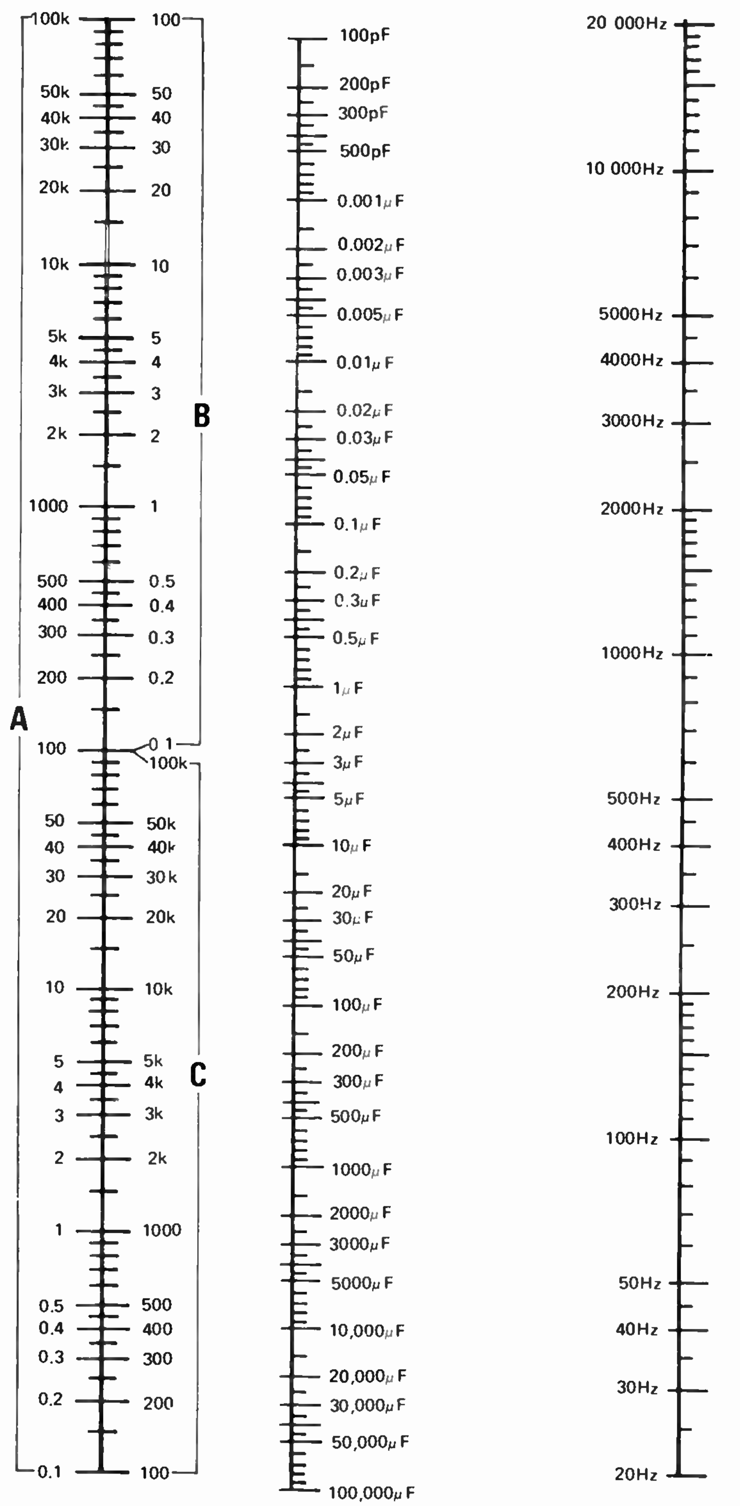
For higher frequencies, multiply frequency scale by 1000, inductance scale by 1000 and divide capacitance scale by 1000. Reactance remains the same.

Capacitive reactance $X_C = \frac{1}{2\pi fc}$

Inductive reactance $X_L = 2\pi fL$

Resonant frequency $F_R = \frac{1}{2\pi\sqrt{LC}}$

Where R is in ohms
C is in farads
L is in henries



INDUCTANCE

SCALE A VALUES IN mH
SCALE B VALUES IN μH
SCALE C VALUES IN H

REACTANCE

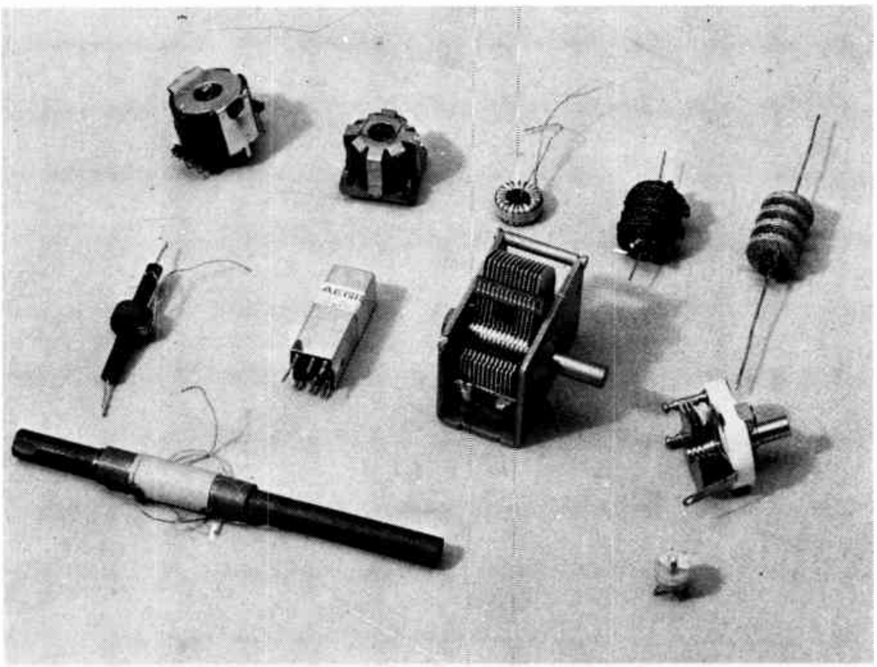
VALUES IN OHMS

CAPACITANCE

USE SCALE A FOR REACTANCE

FREQUENCY

8 Resistance, capacitance and inductance in combination



WE HAVE stressed throughout this course that a good solid understanding of basic electronics is essential if one hopes to understand complex devices.

It is not at all necessary to understand the extraordinarily complex physics going on inside our electronic black boxes. But we must know how these boxes behave in various circumstances and combination.

Hence the fairly solid material that we have presented so far.

Happily this part of the course is now virtually at an end and we are about to get into the more interesting stuff.

That, as they say, is the good news. Bad news is that this last theoretical section is fairly heavy. Do plough through it though — it really is important.

This part of the course deals with circuits that contain resistors in parallel with inductors or capacitors. It also covers the effect called resonance that occurs when inductors and capacitors are used together.

RESISTANCE AND INDUCTANCE IN PARALLEL

Vector diagrams may be used to study paralleled resistance and inductance. This is done much in the same way as series combinations.

In Fig. 1, the signal common to both components is the applied voltage (not current, as in series combinations). So the vector diagram uses a voltage vector as the horizontal reference. The current flowing in the resistor I_R is in phase with the voltage so it is drawn as shown, coincident with V. You will remember from previous theory that the current passing through an inductor lags the applied voltage by 90°. The current vector will therefore point downwards at 90° to the voltage vector.

To find the magnitude of the current drawn from the generator the diagram is added vectorially to produce I_{total}. This procedure is exactly the same as we used previously.

The Pythagoras rule also holds allowing us to compute I_{total} from I_R and I_L giving:—

$$I_{total} = \sqrt{I_R^2 + I_L^2}$$

Similarly the phase angle is found from:—

$$\tan\theta = \frac{I_L}{I_R}$$

A worked example is worth a thousand words, so let us consider the circuit given in Fig. 2a. Here the problem is to work out the current in each component and the magnitude and phase of the current drawn from the 10 V generator. (Remember V is common to both components so we can directly apply Ohm's law to each if we know their reactance values).

$$\text{Hence } I_R = \frac{V}{R} = \frac{10}{25} = 0.4 \text{ A}$$

$$\text{and } I_L = \frac{V}{X_L} = \frac{10}{33.3} = 0.3 \text{ A}$$

By calculation we get

$$I_{total} = \sqrt{0.4^2 + 0.3^2} = 0.5 \text{ A}$$

Alternatively, this result could have been reached by using an accurately drawn vector diagram (see Fig. 2b) in which I_R and I_L are the knowns that lead to I_{total} on completion of the parallelogram.

The tangent of the phase angle is:—

$$\tan\theta = \frac{0.3}{0.4} = 0.75$$

$$\text{Hence } \theta = 36^\circ\ 52'$$

and we know it is lagging as there are no capacitive elements present. The phase angle could also have been found by measuring the angle directly from the graphical vector diagram.

Calculation of current magnitudes and phase angle rarely needs better

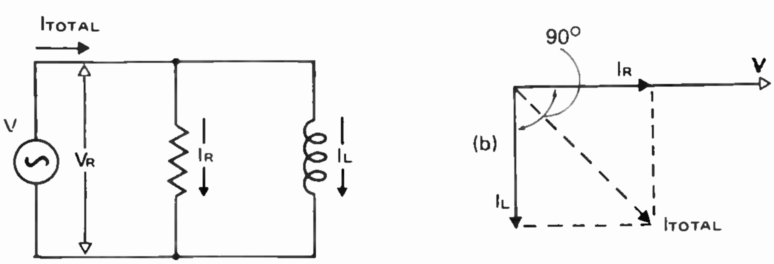


Fig.1. (a) The parallel resistor and inductor. (b). Vector diagram of circuit in Fig. 1a.

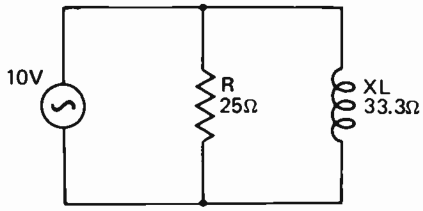
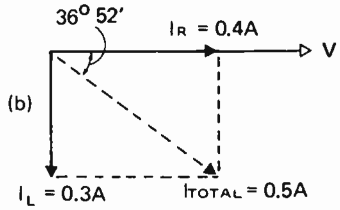


Fig. 2a. A practical example of parallel L and R.



(b) The vector solution.

than 1% accuracy; often 10% is quite adequate.

Indeed, the *majority* of electronic calculations require little precision. There is no point in making long and tedious tasks out of these, often arising, sums. What is more important is that the underlying principle is properly understood. Much of the electronic theory needed in practice is a case of mental arithmetic followed by final adjustment once the circuit is wired up.

RESISTANCE AND CAPACITANCE IN PARALLEL

Figure 3a is the circuit of paralleled resistance and capacitance. The magnitude and phase of the load current may be calculated in exactly the same way as for RL combinations. To check that you have understood the foregoing principles do the figures for yourself and draw the vector diagram as a second check. You should get the values shown in Fig. 3b. Remember, this time, that the current in the capacitor leads that in the resistor.

Now work out the total impedance represented by the two paralleled components – it should be 78.1 ohms. Remember Ohm's law applies to ac circuits provided the impedances are added vectorially to obtain the total – it is quite invalid to arithmetically add the values unless they are in phase (or if 180° out of phase, they can be directly subtracted).

To improve your understanding try it again using firstly, a resistor of 40 ohms with $2\mu F$ of capacitance and secondly, with 40 ohms and $0.1\mu F$. Finally compare the three diagrams and results.

COMBINATIONS OF L AND C

Until now those circuits involving both a capacitor and an inductor have

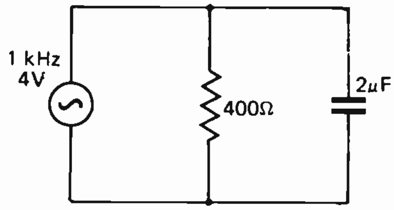
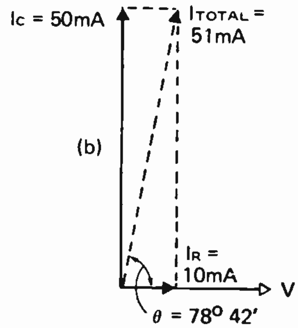


Fig.3a. Parallel combination of capacitance and resistance.



(3b) Vector solution of circuit.

purposely been ignored, for these can (under certain conditions), exhibit characteristics that are strikingly different to those seen so far in our discussion of storage elements.

With the concepts of the vector diagram and the phase of signals behind us, it is now a reasonably straightforward task to gain an understanding of circuits that contain both inductance *and* capacitance.

SERIES COMBINATIONS

When two components are in series, the same current must flow through each, but, as we have previously seen, the voltage across an inductor must lead the current by 90° and the voltage across a capacitor always lags the current by 90°. Thus these voltages *always oppose eath other* (180° out of phase) and the difference between them – is the input voltage! That is, either or both of the voltages across the reactances, may be larger than the input voltage.

To provide a better understanding of what happens in such circuits, let us calculate the current drawn from the supply and the voltages across the reactances in the circuit of Fig. 4a.

Firstly we must find the reactance of each component at the supply frequency of 12 kHz.

$X_L = 2\pi FL =$

$6.28 \times 12 \times 10^3 \times 1 \times 10^{-3} = 75.4$ ohms

$$X_C = \frac{1}{2\pi FC} = \frac{10^6}{6.28 \times 12 \times 10^3 \times 0.1}$$

$= 132.7$ ohms

To determine what the current through the series combination is, we must find the effective combined reactance. As the reactances have the opposite effect, this is simply obtained by subtracting capacitive from inductive reactance (capacitive reactance is always assumed to be negative by convention).

Thus $X_{comb} = X_L - X_C$

$= 75.4 - 132.7 = -57.3$ ohms

The negative sign indicates that the combined effect is that of a capacitive reactance of 57.3 ohms.

By Ohm's Law the current will thus be:–

$$I = \frac{E}{X_{comb}} = \frac{10}{57.3} = 174 \text{ mA}$$

Now that we know the current, we can go back and calculate the voltages across each component

$V_L = X_L I =$

$75.4 \times 174 \times 10^{-3} = 13.1$ volts

$V_C = X_C I =$

$132.7 \times 174 \times 10^{-3} = 23.1$ volts

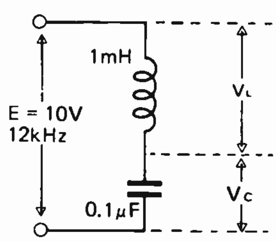
Note particularly the magnitude of these voltages in relation to the input of 10 volts. In fact, due to the subtraction process, the input voltage is always smaller than that across the larger of the two reactances.

The vector diagram for the circuit is as shown in Fig. 4b. We will leave for the moment, the special case where the reactances are equal and study the parallel system.

PARALLEL COMBINATIONS

The parallel combination of L and C is shown in Fig. 5a. In this case the voltage will be common across both components, the current will lag the voltage by 90° in the inductor, and lead the voltage by 90° in the capacitor.

Thus, in this case, it is the two currents which are 180° out of phase so the total current is the difference between them.



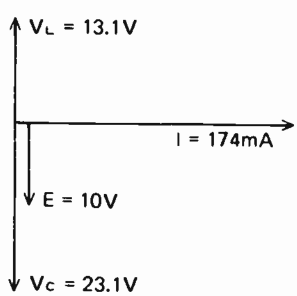


Fig.4a. A series inductor and capacitor combination. (4b) Vector solution.

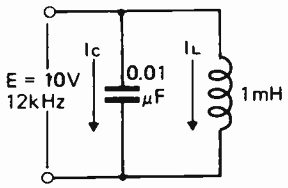
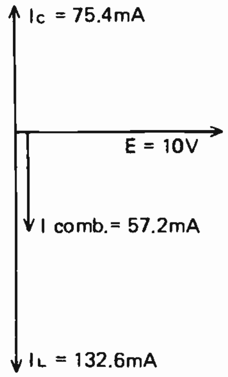


Fig.5a. A parallel combination of L and C.
(5b) The vector solution.



Let us use the same values as for the series case

X_L = 75.4 ohms, X_C = 132.7 ohms

$$\text{Thus } I_L = \frac{10}{75.4} = 132.6 \text{ mA}$$

$$I_C = \frac{10}{132.7} = 75.4 \text{ mA}$$

$$\text{and } I_{comb} = I_L - I_C = 132.6 - 75.4\text{mA} = 57.2 \text{ mA}$$

Compare this current to the previous case. The combined reactance is now:—

$$X_{comb} = \frac{E}{1} = \frac{10}{57.2} = 174.8 \text{ ohms}$$

From this procedure we can deduce that, as the current from the supply is always smaller than the larger of the two reactive currents, the combined reactance, will *always be larger* than the larger of the two reactances. think about it for a while and you will see that this is so.

All practical LC circuits contain some resistance which modifies the behaviour of the circuit. The general circuit of a series LCR combination is given in Fig. 6 and a parallel combination in Fig. 7. These are the most common configurations but by no means the only ones.

In the series case the vector diagram shows how the difference between the reactive voltages is vectorially summed with the voltage across the resistor to obtain the magnitude and phase angle of the supply voltage. Alternatively we can use the Pythagoras rule again to find the input voltage:—

$$\text{Vin} = \sqrt{V_R^2 + (V_L - V_C)^2}$$

and the phase angle

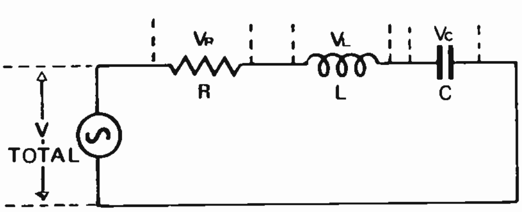
$$\text{Tan } \theta = \frac{V_L - V_C}{V_R}$$

In the parallel case we look at currents instead of voltages. Remember the voltage must be the same across each component, and again a vector diagram of reactive and resistive currents will provide us with the magnitude and phase of the input current.

By Pythagoras:—

$$\text{Iin} = \sqrt{I_R^2 + (I_L - I_C)^2}$$

$$\text{Tan } \theta = \frac{I_L - I_C}{I_R}$$



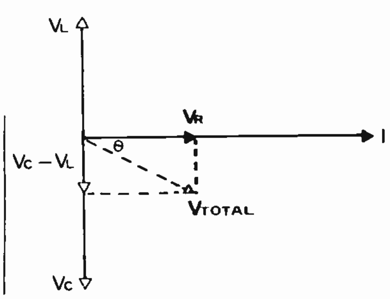


Fig.6a. Series combination of L, C and R.
(6b) Vector diagram of the combination.

APPARENT POWER & POWER FACTOR

In a circuit containing both reactance and resistance, only the energy supplied to the resistor is dissipated. The energy supplied to the reactance is alternately stored in a field and then returned to the supply. Thus no energy is dissipated by the reactance.

The energy supplied to the resistance is called 'REAL' power (because it does work) and is measured in watts. The energy shunted back and forth by the reactance is called APPARENT POWER and is simply equal to the input voltage times the current drawn. The apparent power is measured in terms of volt-amperes — often abbreviated to VA.

The ratio of the real power in watts to the volt-amperes is called the POWER FACTOR of a circuit.

Referring to Fig. 4b we can say that:—

real power in watts = $E_R I$
and apparent power VA = EI

$$\therefore \text{power factor} = \frac{E_R I}{EI} = \frac{E_R}{E} = \cos\theta$$

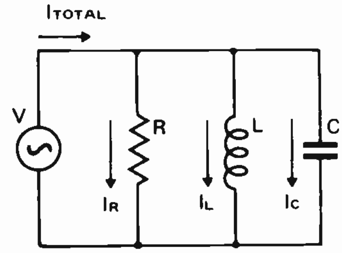
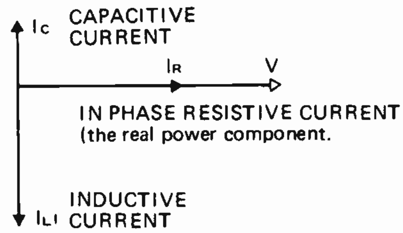
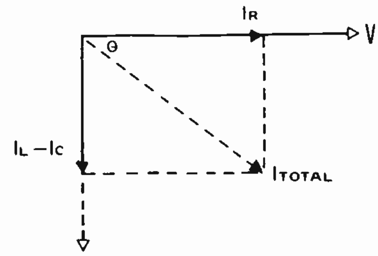


Fig.7a. A parallel combination of L, C and R.



(7b) The basic vector diagram.



(7c) The vector solution — the reactive components having been subtracted.

Thus the power factor in any circuit is equal to the cosine of the phase angle and the power actually dissipated in such a circuit is:—

$$P = EI \cos \theta$$

A PRACTICAL EXAMPLE

An excellent example of the use of reactances is found in fluorescent lights. A basic fluorescent light consists of a gas discharge lamp and a current limiting choke called a ballast as shown in Fig. 8.

Once lit, the complete light appears to the mains as an inductive load and, the current drawn from the mains will lag the voltage by a considerable amount.

The typical four-foot long lamp is rated at 40 watts but, when fed via the correct ballast-choke, draws 0.4 amps from the mains. Thus the VA will be 240 x 0.4 = 100VA approximately! As the consumer only pays for real power, this is of little concern to him, but the extra current drawn causes higher losses in the transmission line, which means the electricity supplier loses revenue. The suppliers therefore, in some areas, insist that large installations of fluorescent lights have suitable power-factor correction.

How is power-factor correction done? Quite easily — because all we need to do to cancel an inductive reactance, is add an equivalent

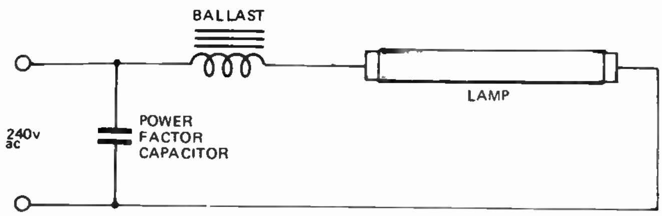


Fig.8. The circuit of a basic fluorescent light fitting (not including starting circuitry). This is an excellent industrial example of the uses of inductors and capacitors.

capacitive reactance in parallel (see section on parallel L and C). Thus a capacitor added across the input terminals will not affect the operation of the lamp but keeps the electricity supplier happy by reducing the input current from 400 mA to about 150 mA.

RESONANCE

As we vary the input frequency to an LC circuit the reactances of L and C change in different directions. That is, as frequency goes up, capacitive reactance goes down, (and inductive reactance goes up). At one particular frequency the reactances will be equal and, when this occurs, we find some very interesting effects – as we will see.

The frequency at which the reactances of L and C are equal is called the RESONANT FREQUENCY, and the circuit is said to be RESONANT at that frequency. Let us now look at the characteristics of series and parallel circuits at resonance.

PARALLEL RESONANCE

In a parallel reasonant circuit the individual currents flowing in the inductor and the capacitor depend upon the frequency at which the ciruit is operated and upon the size of the component (remember $X_L = 2\pi fL$ and $X_C = 1/2\pi fC$). These currents can be plotted as shown in Fig. 9b. The combined current is the direct difference of the two.

At low frequencies the circuit is predominately inductive. As the frequency is raised, more capacitive current flows: at the same time the inductive current reduces. A point is reached where the two are equal and, as they are of opposite sense, the circuit draws no current from the input. It behaves as though the generator is connected to nothing – as would occur if the load was an infinitely high resistance. This happens at the frequency known as the resonant frequency fr, for short. Above resonance the circuit becomes more and more capacitive as the effect of the capacitor becomes more dominant, and the input current gradually increases again.

It is often convenient to consider the impedance of such circuits instead of the currents. Variation of the impedance of a parallel resonant circuit is plotted in Fig. 10. Note the

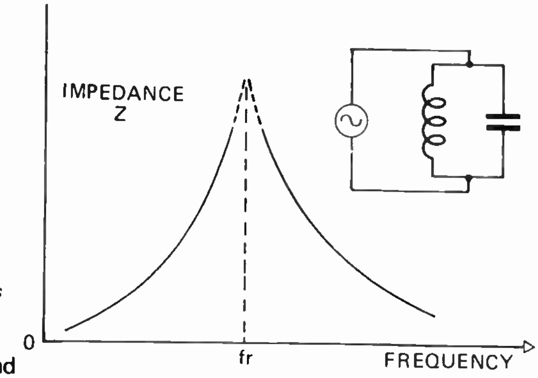


Fig.10. Variation of the impedance of a parallel tuned circuit as the frequency is varied.

theoretical impedance rises to an infinitely high value (zero current flow) at the resonant frequency. However, there is always some resistance in practical resonant circuits and this limits the rise and sharpness of the curve. This resistance is termed the DYNAMIC RESISTANCE.

Circuits capable of resonating in this manner are known as tuned circuits. Tuning is the procedure whereby any of the components is selected or carefully adjusted to achieve the resonant condition at a particular frequency.

SERIES RESONANCE

A similar argument to the above can be used in the case where the two storage components are wired in series. The effective characteristics turn out

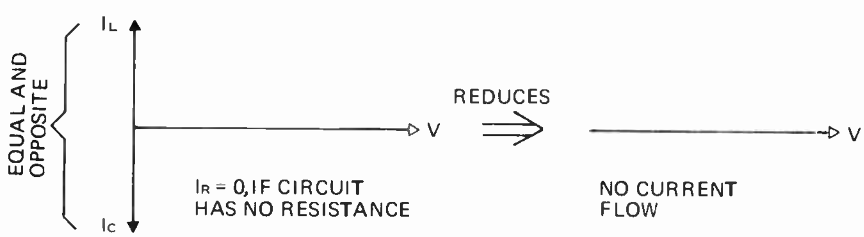


Fig.9a. Vector representation of the condition at resonance, when inductive and capacitive reactance are equal.

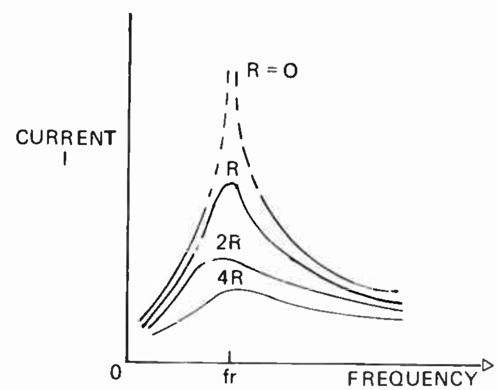


Fig.11a. The dynamic resistance of a series tuned circuit affects the sharpness of the resonant effect as shown.

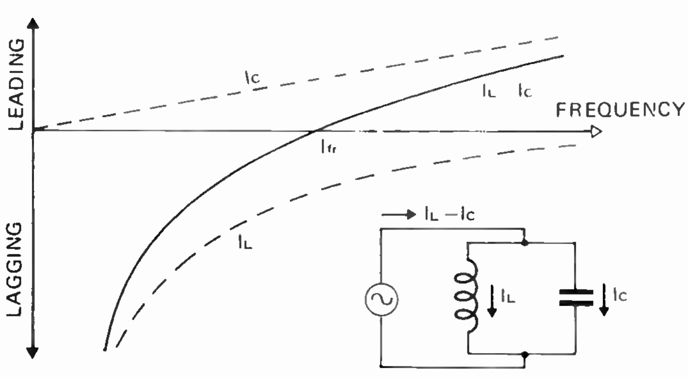


Fig 9b. Currents in the parallel tuned circuit as the frequency is varied.

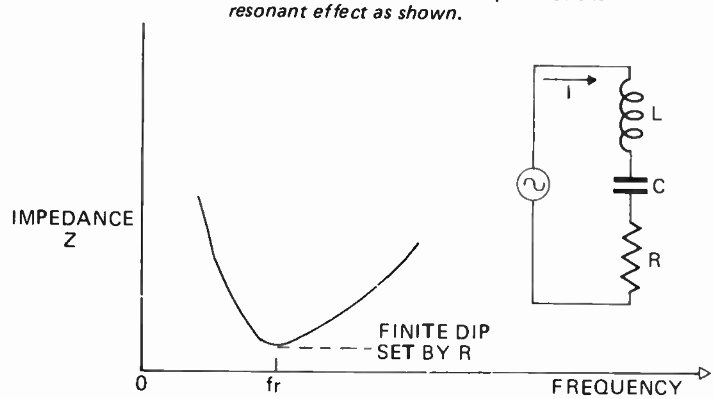


Fig.11b. Impedance of a series tuned circuit drops, at resonance, to a value determined by the dynamic resistance.

ELECTRONICS -it's easy!

to be the reverse of those of the parallel resonance case.

Here the current is common to both components so the typical vector diagram looks like that shown in Fig. 6. The case illustrated has the capactive voltage larger than the voltage across the inductor so the combination appears to be a circuit that has a value of capacitance smaller than that of the component actually in circuit.

It is when the two reactances are equal, at a particular frequency, that interesting things happen for there, the effect of the capacitor cancels that of the inductor and, the source sees only the resistance of the circuit. Ohms law tells us that the current drawn from the source is limited only by the value of the resistance, which in a typical tuned circuit is very small. Consequently the current could well be very large indeed. Fig. 11a shows how the current is limited by various values of dynamic resistance. Impedance variations for a series tuned circuit are given in Fig. 11b. The minimum of the dip is limited by the dynamic resistance.

QUALITY FACTOR OF TUNED CIRCUITS – THE Q

In the series tuned circuit the voltage across the resistor can never be greater than the applied voltage. On the other hand the voltage across the reactive components can rise to values many times that of the supply. The V_L and V_C values in Fig. 6 demonstrate how this occurs.

Consequently the series resonant circuit can be used to produce voltages considerably larger than those supplied to it. The magnification that occurs in this process is expressed as the 'Q', or quality factor of the circuit and is given by

$$Q = \frac{V_L}{V} = \frac{V_C}{V} \text{ (at resonance } V_L = V_C)$$

As the windings of the inductor are responsible for the majority of the resistance a good approximation for the Q factor is found using the realtionship

$$Q = \frac{X_L}{R}$$

introduced earlier in the course.

In parallel tuned circuits it is the current in the reactive components that is magnified, and again the same definition of Q can be used to express the goodness of the tuned circuit. Hence

$$Q = \frac{I_C}{I_R} = \frac{I_L}{I_R}$$

As currents are related to reactances by Ohms law, the Q can also be found from the ratio of the reactance and resistance as for series resonance.

RESONANT FREQUENCY

As pointed out earlier, the resonant frequency is that frequency where the inductive and capacitive reactances, of a series (or parallel) tuned circuit, are equal. That is:–

$$X_L = X_C$$

and hence $2\pi F L = \frac{1}{2\pi F C}$

By transposition we obtain resonant frequency $F_r = \frac{1}{2\pi\sqrt{LC}}$

The following examples will assist. Given a 100 mH inductor and a 0.4μF capacitor, find the frequency at which the two resonate.

$$F_r = \frac{1}{2\pi\sqrt{\frac{100 \times 4}{10^3 \times 10^7}}} = 800 \text{ Hz}$$

This is the frequency for series or parallel resonance of the two.

Often the need is to produce a resonant condition at a given frequency with one component supplied. For example, we may need a circuit resonant at 4 kHz, using a 160 mH choke. The capacitance needed will be:

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

from which $LC = (\frac{1}{2\pi F_r})^2$

or $C = \frac{1}{L} \times (\frac{1}{2\pi F_r})^2$

and putting figures for this example

$$C = \frac{1}{160 \times 10^{-3}} \times (\frac{1}{2\pi \times 4 \times 10^3})^2$$
$$= 0.1\mu F$$

Tuned circuits with zero resistance have the greatest magnification and the sharpest resonance peak. In practice there will always be some resistance present, for the inductor element needs to be as small and light as possible, these factors dictate that the wire used in the coil must be relatively fine in gauge, and hence will have a resistance value that may need to be taken into consideration. However, in systems-level discussions of electronic devices we can usually ignore the effect of the dynamic resistance, we only need to worry about that when actually designing circuits.

If careful measurements of the resonant frequency of a tuned circuit were made, it would be found that dynamic resistance does vary the resonance value by a small amount. In practice, most resonant combinations have an inbuilt variability that enables the capacitor of the inductor to be finely varied to peak up the response.

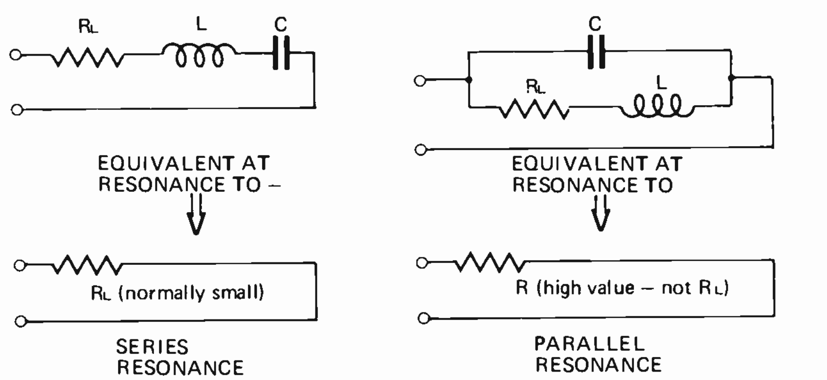


Fig.12. At resonance the two types of tuned circuit become purely resistive. The series circuit becomes a very small resistance and the parallel circuit becomes a very high resistance.

WHAT USE IS RESONANCE?

We have seen the series resonant circuit represents a large impedance when away from resonance but a very small resistance when tuned. The parallel configuration provides the reverse effect. These are summarised in Fig. 12. This way of looking at the resonant circuit is relevant to an understanding of how they are used to select certain frequencies out of a multiple frequency signal.

FREQUENCY SELECTION

Often the need arises to select a known frequency signal (or a narrow band of frequencies) from a wide spectrum. The most common example must be that found in radio transmission where many stations broadcast into the same medium, each at a slightly different frequency. The task of the radio receiver is to tune out the unwanted signals leaving only the required one.

The system to do this is depicted in Fig. 13. A series resonant circuit (Fig. 10a) will provide very little

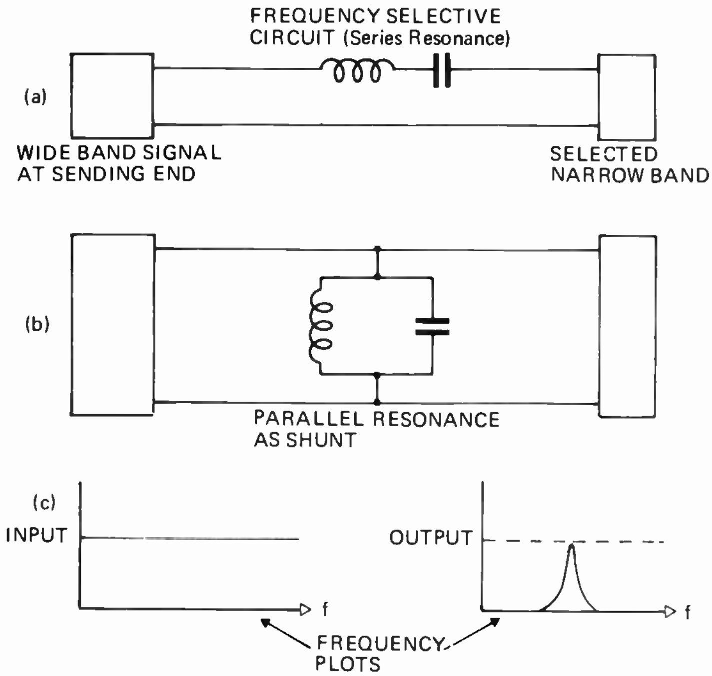


Fig.13. How coupling black boxes with tuned circuits provides frequency selection. (a) coupling by series tuned circuit. (b) coupling by parallel tuned circuit. (c) the input versus frequency compared to the output versus frequency.

attenuation to signals of the required frequency but will act as a larger resistor (actually as an inductive or capacitive reactance) away from the desired frequency. Thus, only those signals near to the resonant frequency

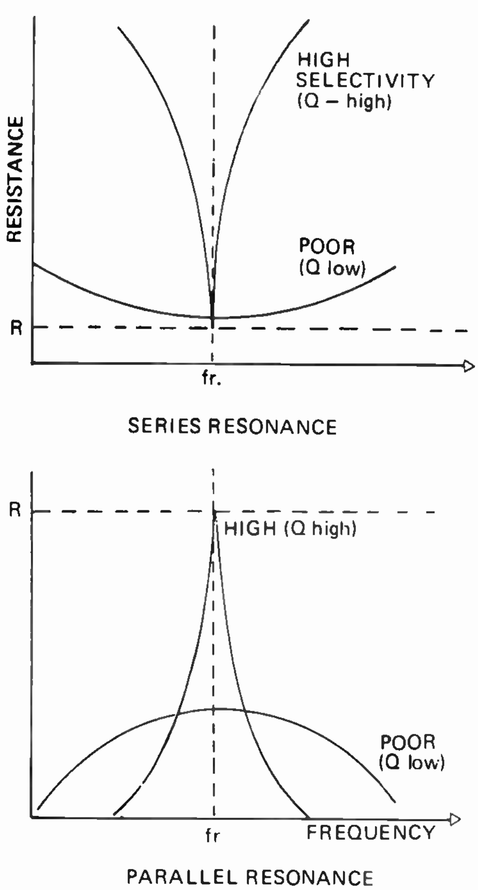


Fig.14. Summary of responses of the two types of tuned circuit with extremes of Q. (a) Series Resonance. (b) Parallel Resonance.

of the combination are allowed through with any signal strength. Although series systems could be used, they seldom are in practice.

A similar effect can .be produced by using the parallel resonant circuit as a shunt across the received output. All frequencies will be attenuated except those required. This form of selection is the one most used in radio work.

The sharpness of this tuning process is dependent upon the Q of the tuned circuit. A coil with a high Q will be more selective (better able to separate two close frequencies) but will of course produce a tuned circuit having a narrower bandwidth. If the signal to be selected is a single frequency all is fine, but most signals must cover a small bandwidth in order to convey information on a frequency as well as time basis. Fig. 14 sums up the various responses. To obtain a wider bandwidth the Q must be adequately low – sometimes resistance is added to spoil the Q to achieve the required compromise between selectivity and bandwidth.

Increased selectivity can be obtained by cascading tuned circuits. Filters used in telephony often consist of many pairs of components. The design of these is very specialised – it is more than merely adding stages one after the other.

When both high selectivity and wide bandwidth are needed, as is the case in radio programme reception, another arrangement is used. Effectively two tuned shunt circuits are used in cascade but with a difference. Each is tuned to a slightly different resonant frequency so that their characteristics overlap as shown in Fig. 15a. The resultant overall frequency response curve is one that has higher gain and a wider bandwidth. The small dip in the middle is not a problem provided the two central frequencies are not taken too far apart.

Rather than use two separate inductors it is, in practice, better to combine them into one component as a doubly-tuned transformer. A transformer is an inductive-coil assembly that can transform ac currents or voltages to smaller or larger values. This is based on the principle of mutual inductance, that is, windings linked by a common magnetic field have voltages induced in them in proportion to the number of turns in each coil.

In the tuned transformer, used in radios, the two windings are wound on a common former; this may be non-magnetic (ferrite, an iron powder material, is now commonly employed) depending on the frequency of operation. Tuning is achieved by screwing-in slugs of ferrite thus slightly altering the inductance. When the capacitance, rather than inductance, of the tuned circuit is to be varied to peak the circuit performance,

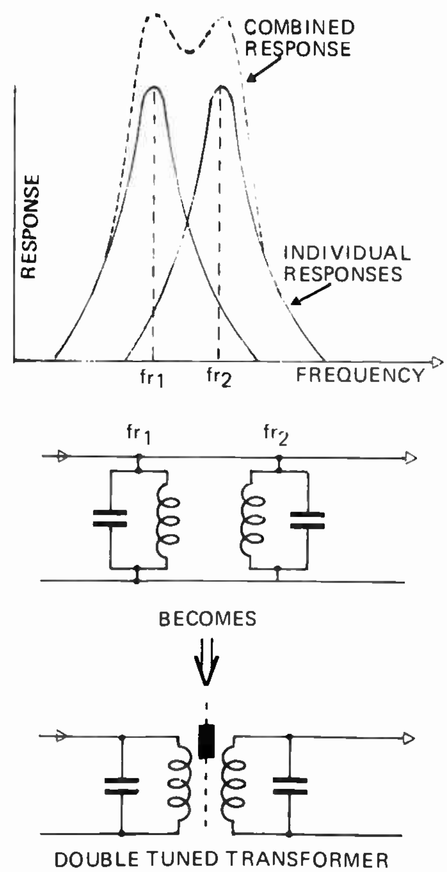


Fig.15. (a) Two tuned circuits may be used to obtain a better bandwidth/selectivity compromise. (b) The two separate-tuned circuits may be combined into a single transformer. This construction is used extensively in radio receivers.

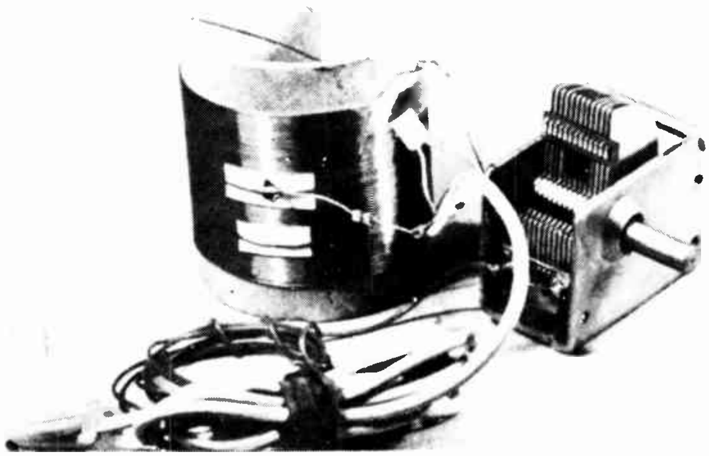
small-range variable capacitors are used. If the range needed is large — e.g. tuning across the AM radio band — the capacitor is invariably made of sets of blades that mesh into each other to vary the capacitance. A range of variable capacitors and transformers and chokes commonly encountered in electronics, is shown in the picture on page 59.

FREQUENCY GENERATION

If a resonant circuit arrangement is given a short impulse of energy — a short period of dc signal, for example — the energy put into the circuit oscillates back and forth between the magnetic field of the inductor and the electric field of the capacitor. This exchange of energy between reactances occurs at the resonant frequency. If the Q of the tuned circuit is high, this process will develop a reasonably pure sine-wave. If no more energy is added after the initial impulse the sine-wave will gradually die away as the energy is dissipated as heat in the coil resistance. If, however, an arrangement is made to add energy to the circuit every time the waveform rises to the same level and phase, the sine-wave will continue to run. High-power radio transmitters make use of this principle to obtain pure signals from highly distorted sources. It is, however, essential that the pulses are delivered to the system at the correct time. Pushing a child on a swing is a good example of pulsed excitation of a resonant system.

Electrically-operated clocks often use the energy pulse concept. The hair-spring and flywheel form the mechanical tuned circuit, as the flywheel rotates it makes a brief electrical contact with a small electromagnet that pulses the flywheel onward with an extra, small amount of energy. Pendulum clocks also often operate this way, gravity providing the restoring force for the mass of the pendulum.

ELECTRONICS - in practice



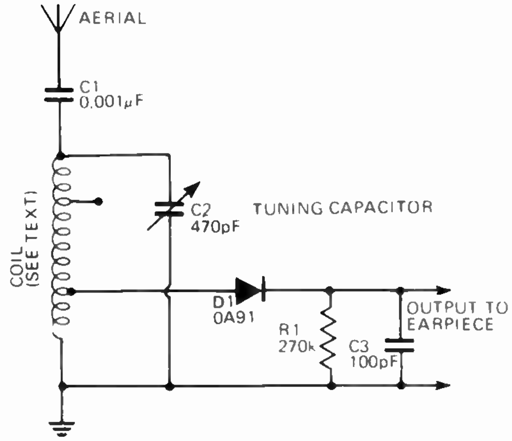


Fig.15. Circuit diagram of the crystal radio.

PARTS LIST — ETI-227

Resistors
R1 .... 270 k ½ watt 5%

Capacitors
C1 .... 0.001uF
C2 .... 100pF

Miscellaneous
D1 .... Diode OA91 or sim.
L1 .... Coil see Table 1.
TC1 .... Tuning gang single section 470pF Roblan or sim.
Coil former piece of cardboard tubing (see Table 1).

TABLE 1
Winding details of air-cored coil (close wound).

COIL DIA.	NUMBER OF TURNS VERSUS WIRE GAUGE			
	22 SWG	24 SWG	26 SWG	28 SWG
1¼"				108
1½"			96	87
1¾"		88	77	70
2"	82	72	67	62
2¼"	71	64	58	54
2½"	61	56	52	49
2¾"	54	52	—	—

Note 1. Tap the coil every ten turns.
Note 2. For former sizes between those stated use an intermediate number of turns. This is not critical.
Note 3. Select the tap for the diode by determining which one gives best volume, whilst still adequately separating the stations.

CRYSTAL SETS were the latest thing in the 20's.

The schematic of the modern version of grandfather's pride and joy is given in Fig. 15. The term crystal-set was coined at the start, for early sets used a small piece of galena crystal which was touched by a fine piece of wire — the cat's whisker — to produce a rectifying contact, if and when you found the right place!

Today, that annoying variable is eliminated by using a germanium diode.

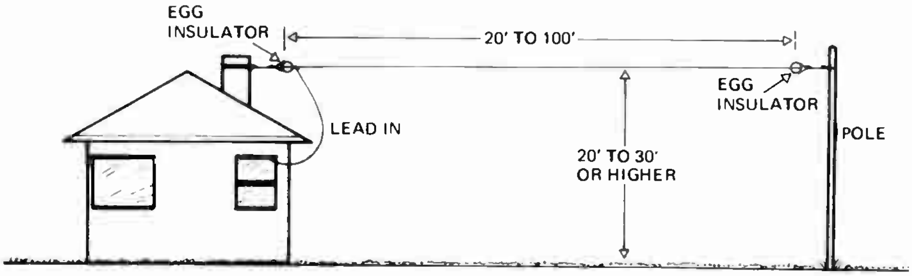
The aerial acts as a conductor in space and couples into the electromagnetic waves sent out by the broadcasting station; minute voltages are induced in it. These small signals, including those from unwanted stations as well, feed energy to the tuned coil and capacitor stage. Signals not at the resonant frequency of the tuned circuit do not excite it and, therefore, go undetected.

The detection and listening circuitry are connected to a tap on the coil, not across the entire inductor winding. This method is used to give the user the ability to make a compromise between signal strength and signal clarity, for the rest of the circuit acts as an unavoidable spoiling resistor that reduces the Q of the tuned circuit. Placed across the full winding the circuit reduces the Q, thus broadening the bandwidth, but reducing selectivity; placed across only a small part of the coil gives the highest Q (the best selectivity) but the smallest signal strength. In use, the taps are tried in turn to find that which gives the clearest signal with the best rejection of unwanted stations.

The diode (virtually any germanium diode can be used) rectifies the amplitude-modulated carrier.

The best headphones to use would be those with high impedance. Impedances of 1 to 5 k are in the correct region. Crystal earpieces can also be tried if you are in a high signal strength area — you might be lucky.

Although this set is not to be compared with modern radios any experimenter who has not built one has missed out on a basic training exercise. It is a must.



This simple antenna is suitable for the crystal radio and may be used for the one transistor radio if required.

A LITTLE BETTER – THE ONE TRANSISTOR RADIO

Since the crystal-set days there have been many changes in radio detection. Apart from more efficient front-end aerials and coils these improvements all involve amplification with active amplifiers. We are not quite to the stage in this course where the operation of transistor amplifiers can be explained, but this simple circuit should present no constructional difficulties.

Note that the input stage is based on a standard modern radio antenna unit. The aerial couples into the tuned circuit by mutual inductance via its own quite separate winding. The resonating signal is taken from the tuned circuit by a second winding, an arrangement that enables a more optimum loading of the circuit to be achieved. It is, in fact, an inductor and transformer combined.

The transistor is used to amplify the signal and the radio frequency choke (R.F. choke) filters out the carrier. We will say no more about the rest of the circuit until the course has proceeded further.

Components for this radio are available as kits ready for assembly from most kit suppliers. It was first published in an early issue of this magazine (Project ETI 406) and subsequently became extraordinarily popular.

If radio receivers are your "thing" a good introductory book covering the practical assembly and operation of the above and more complex models is "Radio" by D. Gibson, Brockhampton Press, 1968. (Illustrated Teach Yourself Series). This inexpensive book is well illustrated and provides the constructional details of sets ranging from a crystal set through to quite advanced receivers. ●

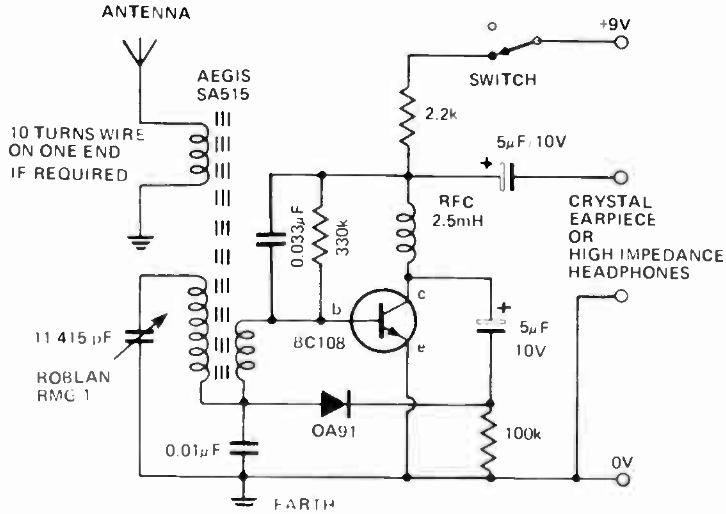


Fig.16. Circuit diagram.

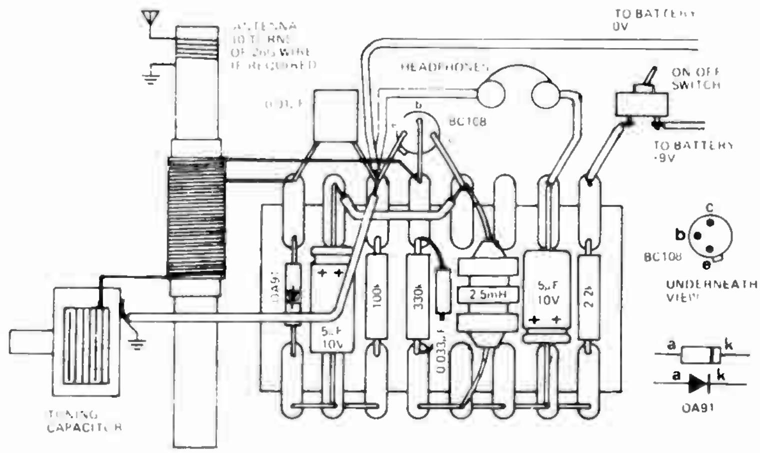
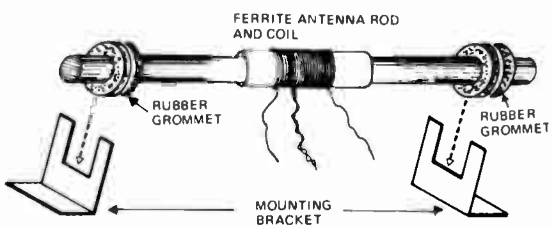


Fig.17. The receiver may be built on tag strips or a piece of circuit board.



The ferrite antenna rod and coil should be supported by two grommets and small metal brackets.

PARTS LIST — ETI-406

Resistors all ½W. 10%
- R1 .... 2k2
- R2 .... 330k
- R3 .... 100k

Capacitors
- C1, C2 .... 5uF 10V electrolytic
- C3 .... 10nF
- C4 .... 33nF

Miscellaneous

1 transistor BC 108, BC 109, 2N3565 etc; 1 diode OA 91 etc; 1 ferrite rod and coil — AEGIS type SA 515 or similar; 1 tuning capacitor 11 — 415pF, Roblan type RMG — 1 or similar; 1 nine volt battery and connectors; 1 toggle switch — single pole single throw; 1 RF choke, 2.5 mH; 1 crystal earpiece or high impedance headphones; 1 pointer control knob; rubber grommets, screws, plywood etc.

9 Detection and amplification

BY THE LATTER part of the 18th century, scientists understood the many obvious effects of electrostatics. They could generate electric charge at will with their "influence-machines". They could store charge in Leyden jars; could make it go where the experimenter wished it to and they could detect it by watching the leaves diverge in an electroscope or the movement of a pair of pith balls (Fig. 1 and 2).

At the same time, social aspects of existence created a faster pace in commerce and in people's way of life, and this in turn created the need for faster means of communication. The mail coaches, travelling at an average of twelve kilometres per hour (at the best!), were just not up to the growing needs of the time.

Against this background, various men of science investigated how they might use electricity to send signals over long distances.

In 1753 Charles Morrison proposed using electrostatic discharge along cables as a means of communication but it was George Lesage who actually first did this – in 1774. But it was one thing to send charge, another to detect its arrival.

IN SEARCH OF A DETECTOR

In 1810 Sommering (Professor of Anatomy at Kassel in German) built an ingenious detector, Fig. 3, in which a minute electrolytic cell (an electrolyte is an electrically conducting solution) produced gas bubbles when the electricity arrived and decomposed the fluid. Seeing these arrivals was a problem so he devised a finely balanced inverted spoon that was situated over the bubble-tube exit in the manner of a see-saw balance. When the gases were generated they collected under the spoon raising it to tip the balance. This in turn set off a mechanical alarm that could be heard with ease. Notice here, particularly, the need for amplification and how Sommering achieved it. As you can imagine the method was somewhat slow but, no doubt, messages in some form of code could be sent faster than by mail coach. We can estimate the delay time would be about 10 seconds, not hours, or even days as required previously, so it was a big step forward. As transmission distances increased, these experiments also provided increased knowledge about the speed of electricity.

Then came another vital discovery. Oersted, by chance in 1819, noticed that electric current deflected a magnetised compass needle. The science of electro-magnetism thus had its beginnings, and was rapidly seen as a tool to provide new forms of detection. Oersted soon had a rig established (in 1820) to detect arriving currents – the needle-telegraph was born. By this time electricity could also be produced with primary-cell batteries – the voltaic-cell (after Volta who reported his design to the Royal Institution in London in 1800) was more easily used than the electrostatic "influence" generators.

Needle-telegraphs were adopted rapidly and numerous types were invented. Figure 4 shows the simplest design of terminal used in the middle 19th century. Many designs were tried in an effort to speed up the reading procedure, to increase the information rate of the communication link, and to reduce the required skill of the operator. Out of this endeavour we have inherited the inkpen recorder, the Schilling code (whereby a number of channels are used most efficiently without sending unused information), the punched tape concept of data storage, the modern teletype machine and frequency and time-division multiplexing. These pioneers too found the need for amplification as we will explore a little later in this section of our course.

The activity in this prior-electronic era – no thermionic valves or transistors then – can be likened to the pace we see in electronics today.

Innumerable forms of telegraph were devised. Indeed by 1850, a practical French telegraph system was clocked officially at a send/receive rate of 40 words in 4.5 minutes (over 350 km). Subsequent tests showed that the system could operate at six words per second over distances of 2000 km. The first under-sea-cable telegraphy channel – across the English Channel

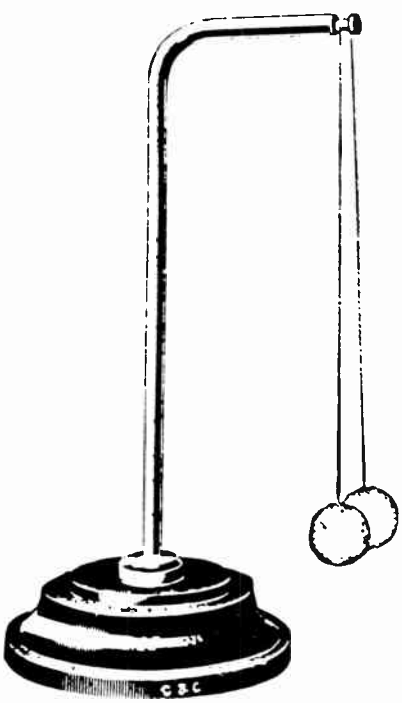


Fig. 1. Pith ball apparatus for detecting charge. If the balls are charged, with the same polarity, they are held apart by the electrostatic field.

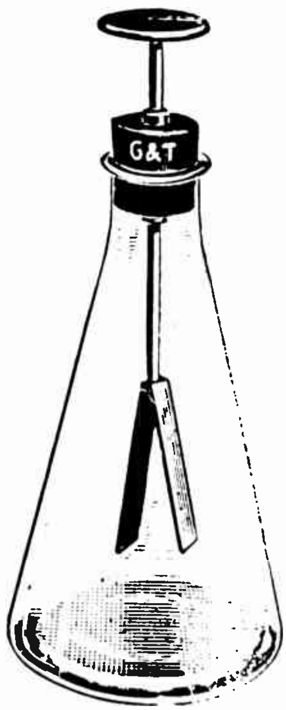


Fig. 2. The gold-leaf electroscope is another charge detector. If the plate at the top of the jar receives charge the gold leaves inside will diverge.

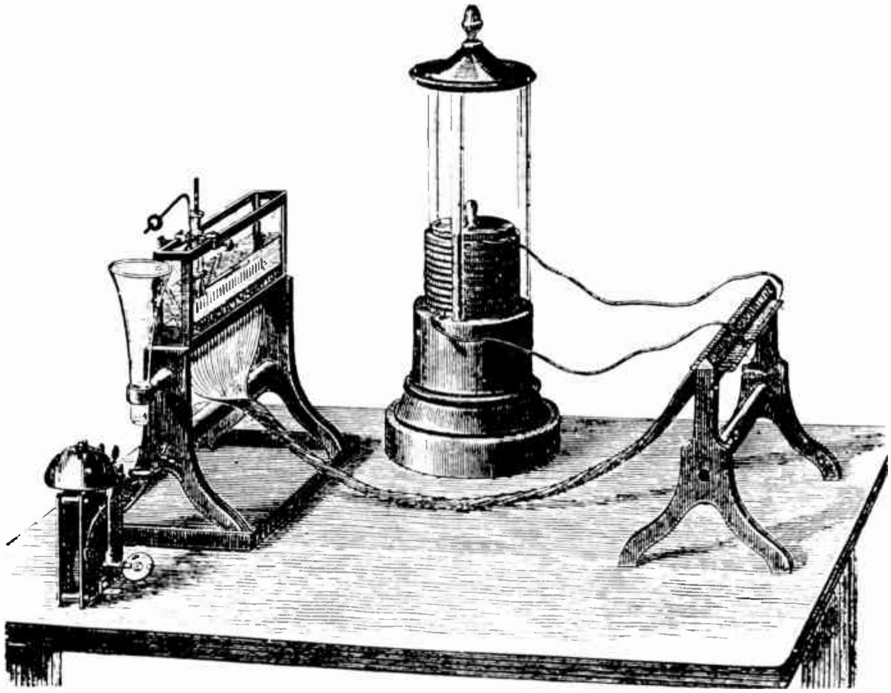


Fig.3. Sommering's 1810 electrolytic detection telegraph. The central voltaic pile provided electric current that was applied to the channel representing the letter to be sent. Current at the receiving end (LHS) produced gas, by electrolytic action; which raised the balanced spoon and caused a lead ball to drop through the funnel onto the bell. It really worked but the problem of finding which channel was in use must have been time consuming.

to France was laid in 1850. It only lasted a few hours but it proved that a submarine cable-link was feasible.

However signals arriving across the (later laid) Atlantic cables were still very poor indeed because of the considerable capacitive effect of the cable. It was like trying to send a square wave signal down a line which was connected to a very large paralleled capacitor — the received pulse was thus badly distorted and the problems of detection were still most severe. A real breakthrough was made when Lord Kelvin, then William Thomson, designed a very sensitive mirror galvanometer by which one could 'see' the arriving code pulses (the moving-coil of this type of meter rotated a mirror which deflected a light spot across a scale — Fig. 5). The use of this instrument enabled the first signals to be exchanged between the U.S.A. and Britain, in 1858.

In the 1880s, Heaviside suggested that loading the cable with added inductance would cancel out the capacitive reactance. This was found to improve performance greatly (remember our related discussion of vector diagrams). Discrete, lump-loading, (that is inductors added at intervals) was used originally but today we add inductance by winding a continuous Mumetal strip around the inner core.

THE BIRTH OF RADIO

By 1900, cable-telegraphy over cables was totally accepted, but by this time another system — radio — was emerging as a contender for speedy communication. It too posed severe detection (and amplification) problems.

In 1886 Hertz made the first practical observations which proved that electro-magnetic radiation did exist. He explored its properties but felt it was of little practical use to mankind! Hertz died in 1894 not knowing what a vast technology he had initiated.

In that same year Marconi, then 20 years old, became interested in the various Hertzian wave phenomena and carried out many experiments using equipment based on other people's designs. Guglielmo Marconi was not an *originator* of ideas, rather an accomplished developer of established technique — he took existing ideas and made them work better and in more useful ways.

Hertz had used arc-discharge to produce high-frequency ac current in his sending aerial, and a minute spark gap across an antenna to see, or detect, its reception.

Later on, Oliver Lodge (in 1894 again) managed to transmit and detect 18 cm wavelength radiation at a 20 metre range (through walls) using an improved detector — which he named the coherer.

In the coherer method of detection E.M. radiation causes metallic filings (the most common type) to change from high resistance to a low resistance. De-cohering was accomplished by a gentle tap with a little hammer. There were many kinds of coherer (see Fig. 6), but Marconi's nickel and silver filings design was

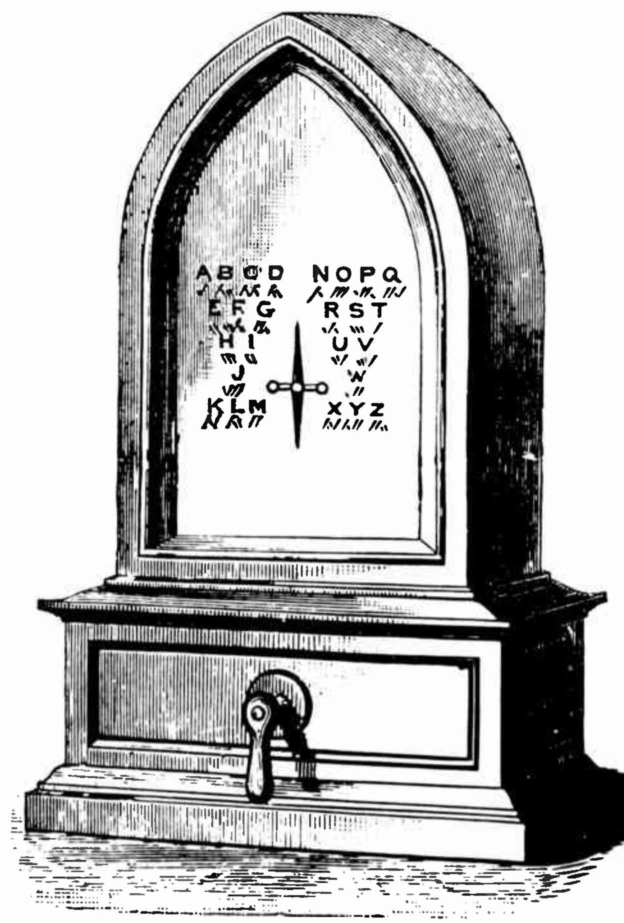


Fig.4. A typical, single-needle, telegraph terminal of the middle 19th century. The handle was turned to the left, or right, causing the needle at both ends to move accordingly. The code sequence of movements decided the letter sent and received.

probably the most sensitive. Improved antenna design increased range to a kilometre and radio telegraphy was started as an alternative to cables which were still proving troublesome across waterways.

The coherer was somewhat limited, it had to be tapped to restore it – (one of Marconi's 1899 detectors incorporated automatic control to provide a restoring tap after a pulse of radiation was detected) – consequently experimenters looked for a better way to 'see' the signals.

Next in line (in 1902) came the magnetic-detector. Compared with the coherer this new device had higher sensitivity, better discrimination against noise and faster overall response and reset time. Its principle (again developed but not invented by Marconi) was based on Rutherford's 1895 discovery that a superimposed high-frequency signal applied to an electro-magnet makes it more sensitive to ac signals. (This principle is still used today in magnetic tape recorders – it is called high-frequency bias.)

The continuous band of the Marconi magnetic detector, shown in Fig. 7, is made of iron wires and was driven as an endless belt by clockwork. In the centre were two horseshoe magnets that provided steady remanent magnetism in the wire as it passed. At the same point were two coaxial coils surrounding the wire; one went to the antenna, the other to headphones. Under normal conditions the wire experienced no change in field strength, just a steady unchanging value, and no voltages were induced in the headphone coil. If, however, a voltage occurred in the antenna coil, this induced a signal in the headphone coil, which then followed the audio frequency, thus reproducing the original audible buzz of each code bit.

The magnetic detector was also investigated by Wilson and Evans in 1897 in an attempt to trigger remote devices – torpedoes, in fact.

THERMIONIC VALVES

Marconi did not play a dominant role in development of the thermionic valves that were soon to replace the magnetic detector. That achievement went to Sir Ambrose Fleming who pioneered the diode or two-element thermionic valve.

In the years closely preceeding 1904, Edison had discovered an effect that he could not explain. His incandescent, carbon-filament lamps blackened with use. To investigate the problem he added a second plate inside the glass envelope (as shown in Fig.8a). He found, to his surprise, that current flowed between the filament and the plate, when the latter was wired to the positive terminal of a battery, but not when reversed. History has it that he did not realise the implications of this finding, but he had in fact constructed the first thermionic valve rectifier. The effect became known as the Edison effect. Ambrose Fleming recognised the useful properties of Edison's device. He went on to improve its performance and apply it to the detection of coded-radio signals. It also enabled analogue, (continuously varying) voice signals to be transmitted and detected with greater simplicity than any then-existing method. At last a really satisfactory rectifier was available.

The valve era of electronics was born. Fleming's diodes (see Fig. 8b) were adopted immediately for weak signal rectification.

But that was not the end of the development for yet another discovery was the rectifying property of a pressure contact made between a crystal, such as galena, and a fine wire. This is, of course, the "cat's whisker" detector mentioned in the previous article in this series. Undoubtedly, this was the forerunner of the point-contact type of semiconducting diode and the junction-diodes of today.

Today, thermionic valves find little place in new designs but they are still used in high-frequency or high-power equipment – we will describe their operation later in this chapter.

Let us now turn to the second great problem of those days – that of amplification.

AMPLIFICATION

The ability to rectify ac signals into a dc form was a great step toward establishing an electronic discipline. But more significant again was the final breakthrough when a thermionic amplifier valve was devised in 1907. To fully appreciate how useful it is to be able to amplify small signals routinely, we need to look at the methods available to designers before this time.

We have discussed in earlier parts how an electronic system is basically a means of communicating one physical effect of the natural world from one place to another, electronic circuitry providing the most convenient energy transmission medium for most purposes.

Many of the physical effects to be transmitted are too small to be sensed by our normal physiological senses. The need might be to hear the noises of insects, to see the behaviour of biological cells, to hear and see each other when out of normal range or to see minute movements. In each of

ELECTRICS may be said to be the application of electricity to passive components (or electric motors etc) where signal amplification is not necessary (eg an electric drill, house wiring etc).

Electronics, in the broadest sense, covers applications requiring the use of active devices (transistors, vacuum tubes, integrated circuits) for controlled signal amplification, eg, speed control, radio, television etc.

No clear cut definition is possible, however, for a relay amplifies (small signal in coil controls large signal through contacts) and thus may be considered as either an electric, or an electronic device. Further, the humble crystal set contains no active devices, nor source of energy other than that received by the aerial, yet it is considered part of the electronic discipline.

Nevertheless, our definition is close enough.

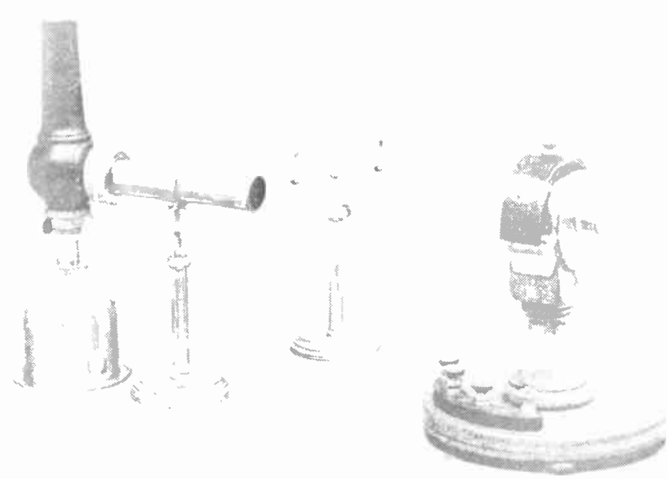


Fig.5. Kelvin's mirror galvonometer (1858). The oil lamp, at left, provided a light beam that was deflected by a mirror mounted on the coil of the galvanometer at right. The deflected beam moved across the calibrated scale at centre.

these, and many other examples, the energy level of the original signal is inadequate to satisfactorily operate our detectors and some mean is needed to amplify the effect. We also refer to this as adding gain to the system. Electrical scientists and engineers prior to 1912 had a tough time, for gain was just not to be had without the application of ingenuity and cunning.

EARLY DEVICES

In 1858 Thomson invented the mirror galvanometer as mentioned above. That provided gain by using the optical lever principle, but it did not provide electrical amplification,for the output form was a displacement, not an electrical quantity.

Mechanical levers were often used to provide increased displacement amplitude. In seismology, minute movements of inertial mass were transformed into considerable deflections of a stylus, by using levers and long arms. Shaw and Laws (around 1900) measured the magneto-strictive length changes of nickel with their 6-lever "electric micrometer": the micrometer screw was turned (see Fig. 11) until a contact was made.

Early designs of gramophones and recorders usually managed with one input or output trumpet but the Columbia quadruple-disc "gramophone" of 1904 had four trumpets to provide enough signal to cope with an audience of 20,000 people. One design of early telephone mouthpiece used two trumpets to couple the speech vibrations to no less than 12 microphone units!

Prior to the discovery of the amplifying valve the dominant electrical gain device was the electro-mechanical relay. We have already met the relay in an early practical exercise. Today their form has little changed from the first unit devised by Wheatstone in 1837, (Fig. 12), it was used to operate a bell.

Today, relays can be made much smaller and with great reliability but the principle remains unchanged.

Relays can only produce digital signals, the contact is either open or closed. Because of this, whilst relays are invaluable in dot-dash type telegraphy, they are useless in voice-telephone work.

Nevertheless, the relay principle played a vital part in early electrical developments for, as well as being able to amplify signal levels, they provided the means of driving equipment — automatic feeds for arc-lamps, clock rewinders, alarm releases, printing telegraphs and step-by-step telephone exchange selector switches. Provided digital operation sufficed, relays could

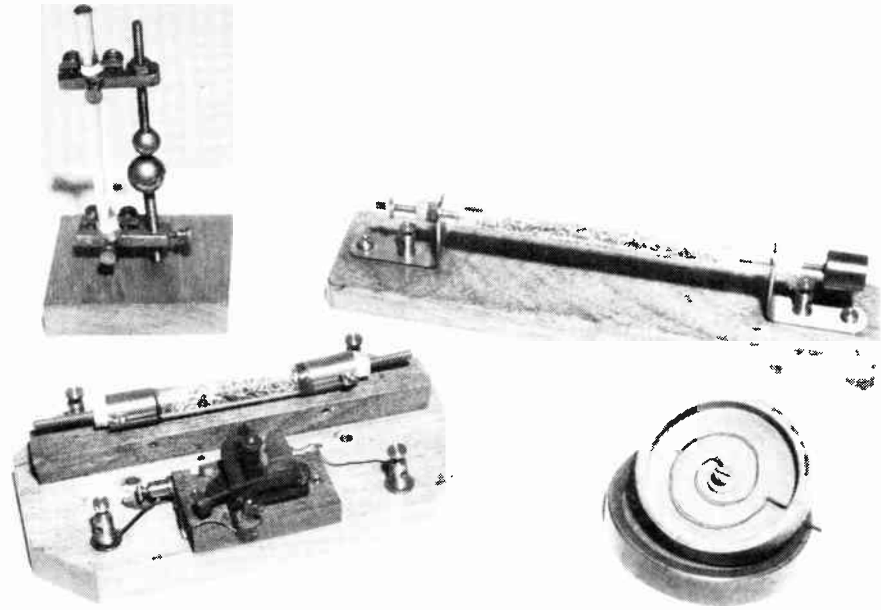


Fig.6. Equipment used by Lodge (1894) to detect electro-magnetic radiation. Two types of coherers, used as detectors, are shown. (top right and bottom left).

easily provide stable gains of a million or more. Brown's signal regenerator of 1899 used a relay, to sense the level of incoming poor-quality pulses from underwater cables, and hence to gate out clean levels thus repeating the original pulse signal.

All manner of methods were tried to obtain amplification of analogue (continuously varying) electrical signals. Probably the most successful before the thermionic valve was Shreeves' electromechanical telephone repeater unit, but it came too late (1910) to help the art. Shreeves' device consisted of a nicely packaged telephone receiver ear-piece mechanism driving a mouthpiece mechanism as a combined single unit. The mouthpiece used a method whereby a dc bias current is modified by the audio-frequency signals of the earpiece. Gain was, thereby, introduced by controlling the rate at which power flowed from the biassing power source into the output circuit. The input energy only had to decide the *rate* of output power flow; it did not have to *provide* it. (This, as was pointed out earlier in the series, is the definition of an amplifier).

THE VALVE AMPLIFIER

In 1907 Lee de Forest conceived the idea of introducing a perforated metal plate, (Fig. 9) between the filament and plate of the Fleming diode valve — this was the first triode valve. They were known as "Audions" and by 1912 were in use as amplifiers.

Their operation is quite straightforward. A voltage of the correct polarity (anode positive) will cause a current to flow between the

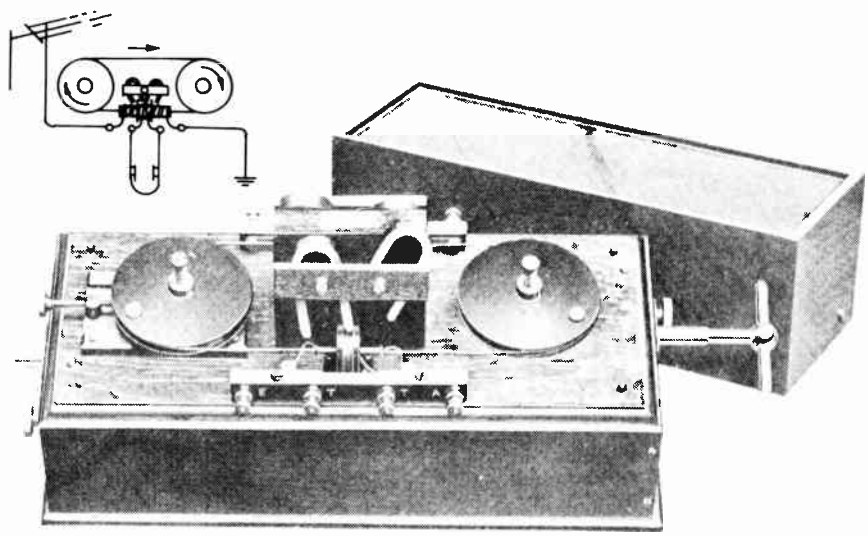


Fig.7. Marconi's magnetic detector of 1902.

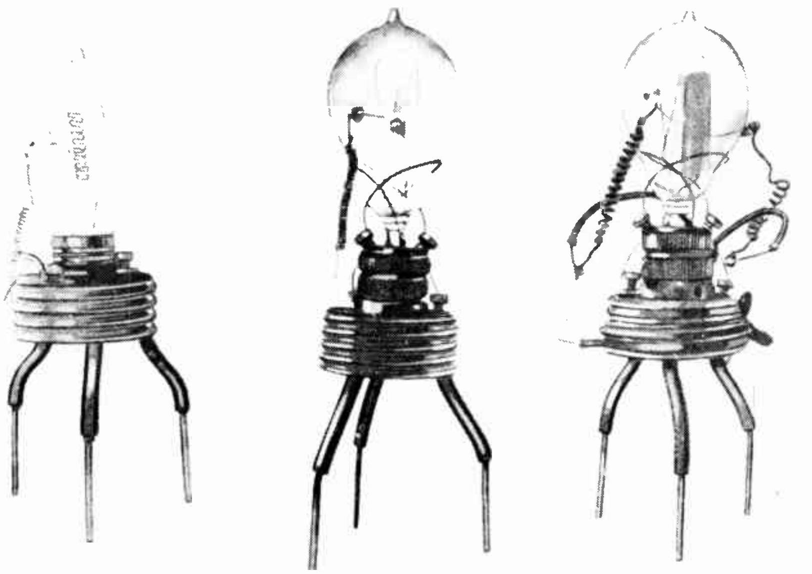
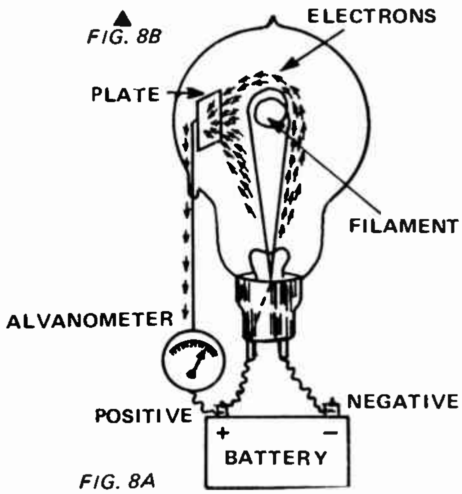


Fig.8a. The Edison effect (b) Three of the actual diode valves used by Fleming in 1904.



filament (cathode) and the plate (anode) if the grid is left unconnected. Signals connected to the grid either allow or prevent this action depending upon their polarity and magnitude — as shown in Fig. 9. A small varying signal voltage applied to the grid controls the flow of a large current in the anode circuit, thus obtaining gain (by producing a signal larger than that fed in).

Early "audions" were not particularly good amplifiers but they could be cascaded to provide increasing signal. They were also incapable of carrying much current to begin with. By 1922, however, 5 kW valves had been developed and by 1930, 1000 kW valves existed along with peanut-sized units for radio receiver work.

The German designed Loewe multiple-valve receiver of 1930 was especially interesting for it contained, in a single glass envelope, three triodes along with the resistors and capacitors needed for the three gain stages. Perhaps this was the first integrated circuit! The advent of valves gave considerable impetus to the development of electronics for they gave designers a new dimension of freedom. It became reasonably easy to build amplifiers, oscillators, digital circuits (the digital computer), measuring instruments (the first vacuum tube voltmeter was probably that originated at Cambridge University by Mallin in 1922), battery eliminators, successful television — the list is virtually endless.

TRANSISTORS

Valve technology continued to improve, but by the early 1950's the shortcomings of valves — excessive size, power dissipation and cost — were becoming an intolerable barrier to further progress. Early digital computers filled many rooms of a building, portable radio sets needed large batteries etc. A new development was needed, and in 1947 the first practical transistor amplifying element was produced to fill the waiting need. The idea had been around for several decades but the necessary production technology had not been available.

This development initiated the so-called solid-state era that we now enjoy. Today, transistors are used by the hundred and even thousand in modern integrated circuits. We now are truly at a systems level, for electronic designers today think more in terms of the *capability* of given circuit blocks than about how to interconnect separate, discrete elements.

The foundation element of active circuit system blocks is the amplifier. In articles that follow we shall discuss this vital component assembly considering it as a black box that behaves in different ways depending upon how passive components are connected around it. Our study of amplifiers will include a brief introduction to the thermionic valve

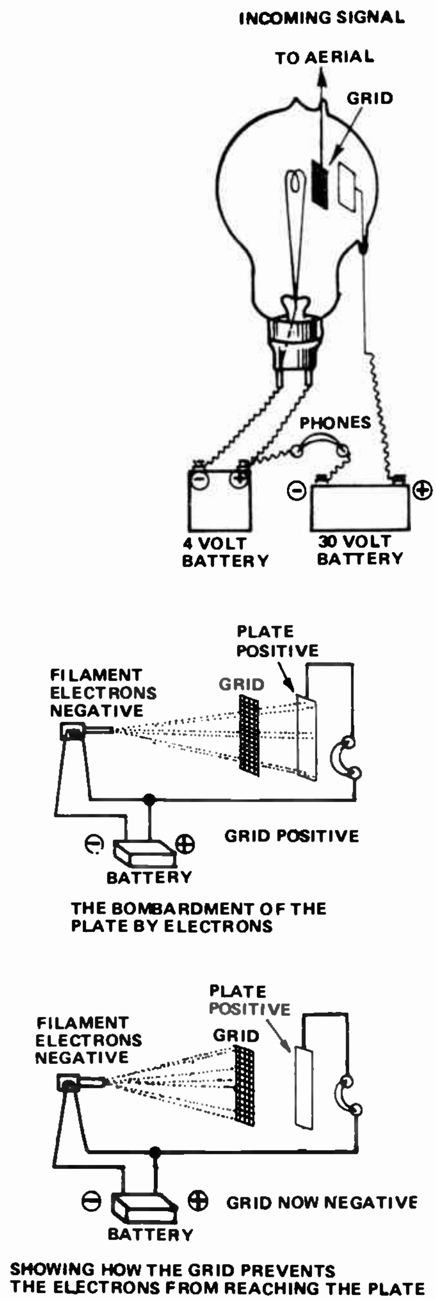


Fig. 9. The triode valve, invented by Lee De Forest, and a schematic of how it works.

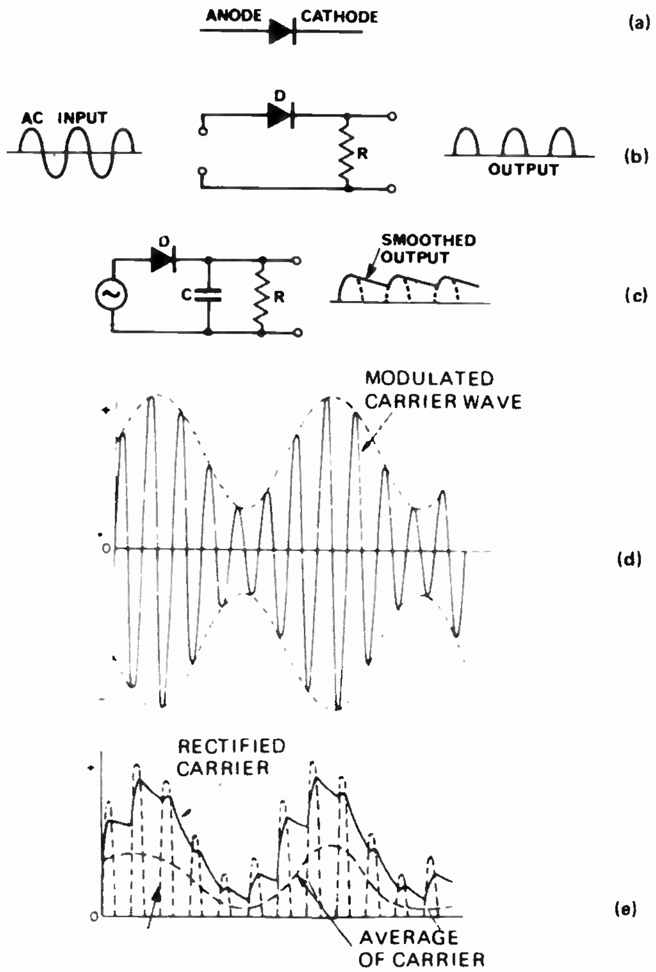


Fig. 10.

RECTIFICATION — THE ONE WAY VALVE

A diode, as we have seen earlier, acts as a one-way valve for current flow. The symbol for a diode is shown in Fig. 10a. The two leads of the diode are named cathode and anode, and the diode will only conduct when the anode is more positive than the cathode.

One major application of diodes is in the conversion of ac to dc. When sine waves are applied to a diode circuit (Fig. 10b), the negative half-cycles will be blocked. The diode will only conduct when the anode of the diode is positive with respect to its cathode. The result will be pulses of current which may then be smoothed into a steady, direct current by the action of a capacitor (Fig. 10c).

This process, of converting ac into dc, is called RECTIFICATION, and a diode constructed for this purpose is called a RECTIFIER.

As most electronic apparatus requires a source of smooth dc power, and as the mains power supply comes to us as ac, the process of the rectification is fundamental to modern electronic systems.

Further, if an ac waveform, modulated as shown in Fig. 10d, is applied to the circuit of Fig. 10c and the time constant of CR is chosen to be much longer than the period of the carrier signal but short compared to the minimum period of the modulation frequency, the circuit of Fig. 10c will effectively demodulate, or detect, the signal thus recovering the original modulation as shown in Fig. 10e, such a circuit is called a DETECTOR circuit.

Diodes may also be used to protect equipment against reversed connection of the power supply, to protect against excessive input voltages and for a whole host of other applications.

amplifier: it will be brief because the technology is now outdated. Nevertheless, it still is used in many measuring instruments and electronic devices and, the principles involved align with those used in solid-state amplifier circuitry. It will also help those trained in valve technology to better appreciate the operation of transistors.

The course will then describe the necessary basics of semiconducting amplifier components without undue explanation of semiconductor theory – that will be left to added reading, for a thorough knowledge of the physics of semiconductors is not necessary for an appreciation of the electronic discipline.

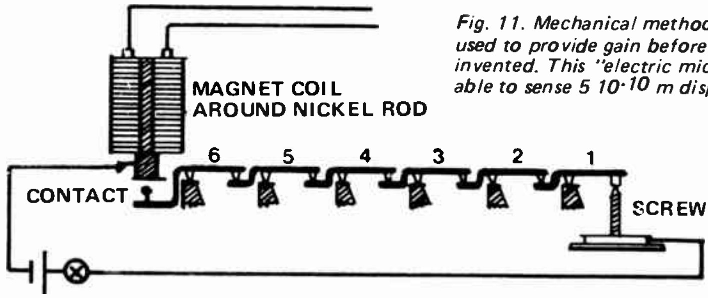


Fig. 11. Mechanical methods were often used to provide gain before valves were invented. This "electric micrometer" was able to sense 5 10^{-10} m displacements.

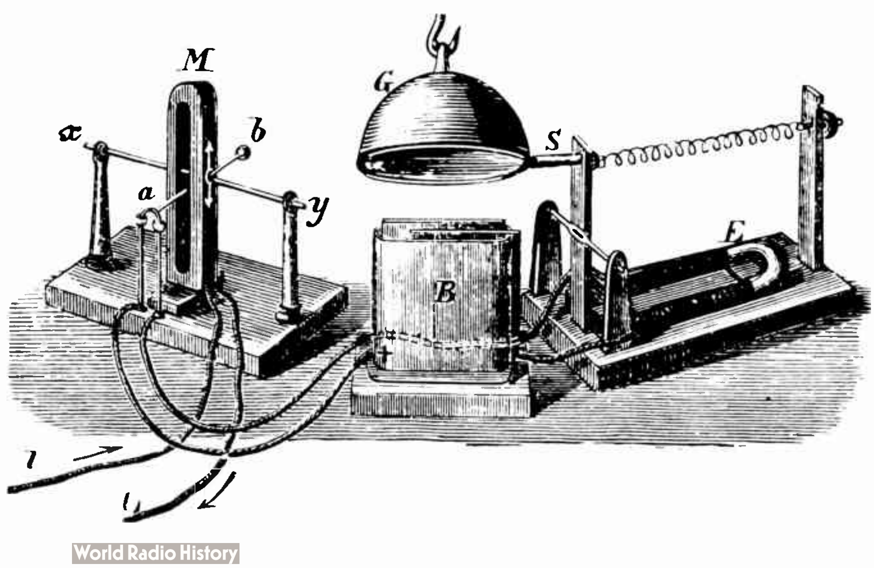


Fig. 12. Cooke and Wheatone's relay of 1837 – the first. Current entering the wires on the left caused the compass needle 'ab' in the coil M to rotate making two mercury contacts at a. This in turn operated the bell hammer by means of an electro-magnet (E) and battery (B).

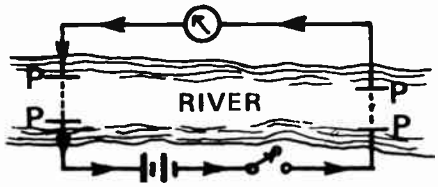


Fig. 13. In 1842 Morse demonstrated wireless electrical communications using the ground as "wires".

ELECTRICITY flows through the ground with ease. In 1842 Morse showed that he could communicate across a river in New York without wires. He used the principle shown in Fig. 13. The plates P were made of copper and were deeply sunk into the mud. He established that the current flowing across the river was proportional to the plate size, and to the distance between the plates on each side. (The current paths do not short out!) Lindsay signalled across the 1.5 km width of the river Tay, Scotland, in 1854 by this means. After this, many attempts were made to bridge larger distances but as the method is very inefficient little came of such effort.

This intriguing technique is easily demonstrated using the circuit of Fig. 14. It uses a buzzer oscillator similar to that shown in Chapter 5. It is left to you to experiment with plate size and spacing. The idea is not restricted to spanning water only — it works for continuous ground.

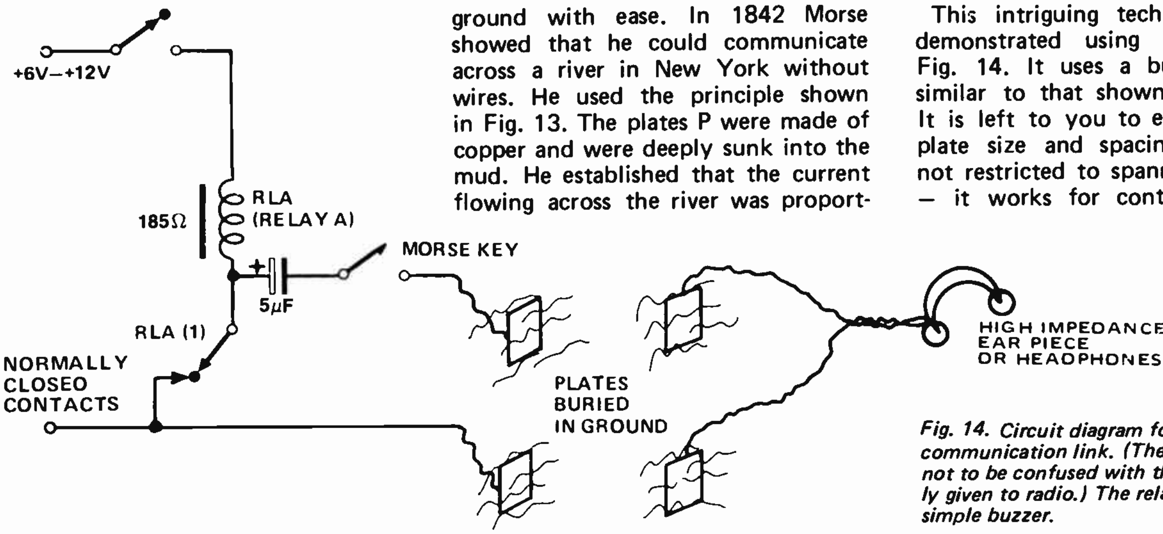
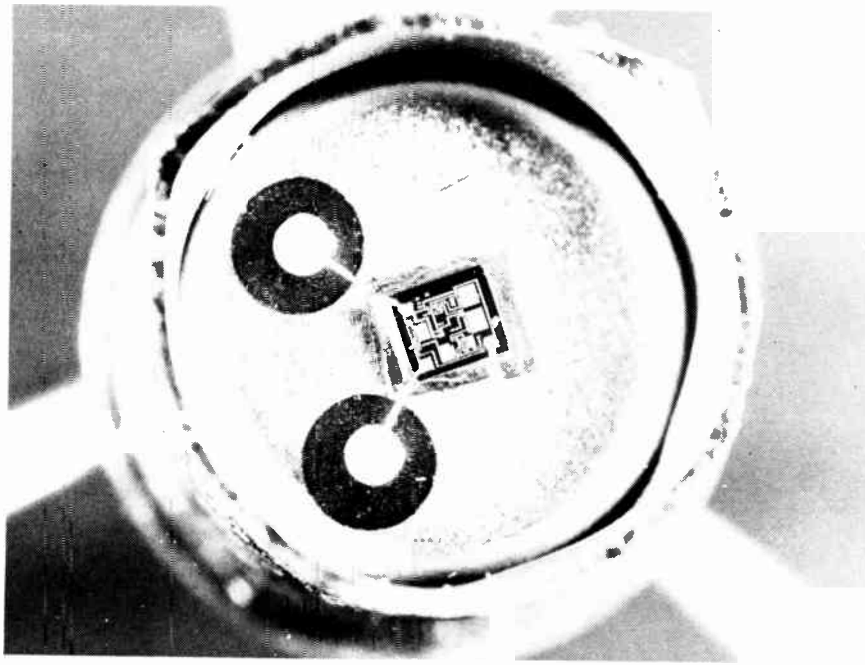


Fig. 14. Circuit diagram for a wireless communication link. (The term wireless not to be confused with the name formerly given to radio.) The relay acts as a simple buzzer.

CHRONOLOGY OF SOME KEY DATES IN ELECTRONICS

Ancient World knew of static electricity
1730 S. Gray discovered the use of wires to convey static electrical charges.
1745 Leyden jar devised by Kleist, Muschenbrok or Cuneus — the first capacitor.
1746 Winkler discharges Leyden jar using water return circuit.
1786 Galvani's work on animal electricity.
1800 Volta's primary battery invented.
1819 Oersted discovered current in a wire deflects the needle of a compass.
1825 Sturgeon made first electro-magnet.
1827 Ohm announced his law relating current, voltage and resistance.
1831 Lindsay telegraphed via ground and water circuits (probable date).
1831 Faraday discovered laws of induction.
1843 Wheatstone's bridge circuit published.
1852 Kelvin related L, C in resonant circuits to natural resonant frequency.
1853 Kelvin published paper "Oscillatory discharge of a Leyden jar".
1858 Kelvin's sensitive mirror galvanometer used on Atlantic cable.
1861 Feddersen's work experimentally proved oscillatory nature of discharge.
1863 Clerk Maxwell suggested EM waves exist using theory only.
1867 Clerk Maxwell produced formulae describing EM waves.
1867 Kelvin's syphon recorder devised and built.
1873 Clerk Maxwell published treatise "Electricity and magnetism".
1879 Berlin Academy of Science offered prize concerned with nature of Maxwell's theory of EM waves.
1883 Edison noticed rectifying effect in lamp (described in an 1884 patent).
1884 Preece began radio experiments including induction between wires.
1885 Preece experimented with induction wires.
1886 Hertz obtained EM induction with close coils and Leyden jars.
1888 Hertz demonstrated existence of EM waves.
1894 Marconi's personal claim to first recorded message through space by EM waves.
1895 Popoff worked on electrical storm detection using Branly coherer and lightning conductor; incorporated de-coherer device.
1897 Marconi's own claim for first telegraphy between ships in motion (20 km range).
1898 Braun used tuned circuits as coupled resonance.
1898 Tesla's patent for controlling route of a distant ship by radio.
1898 Zickler reported ultra-violet link over 1.3 km.
1899 J.J. Thomson explained the Edison effect of 1883.
1900 Dudell's patent on singing-arc continuous wave production.
1900 Car radio patented.
1904 Fleming invented thermionic diode for signal detection.
1904 Nussbaumer's musical transmission system.
1904 Hueslmeyer's radar using radio waves reflected from ships.
1905 Lieben's valve work (patented in 1906 as a relay) using magnetic field control of internal current.
1906 Dunwoody discovered crystal detector.
1907 Fessenden patented heterodyne detection — well before its time.
1907 Lee de Forest patented triode valve.
1910 Lieben built valve amplifier.
1913 First USA musical transmission.
1922 British Broadcasting Company came into existence.
1930 Loewe's integrated circuit valve receiver.
1930s CRO introduced as measuring instrument.
1950s Transistors begin to replace valves.
1960s Integrated circuits replace discrete component circuits.
1970s Electronics firmly in systems, building block, state.



◀ *Would you believe that there is a ten transistor radio on this tiny 1 mm square chip. The device is the Ferranti ZN414 radio IC.*

10 Elements of transistor amplifiers

AN AMPLIFIER, whether electronic, mechanical, acoustic or optical, is a system building block. It allows the amplitude of an input signal to control a secondary source of power such that the amplifier output is of larger power (or voltage or current etc.) than the input signal. This concept is shown as a block diagram in Fig.1. In its simplest form, an electronic amplifier has one input, one output and source of power. The common line is usually not shown in block diagrams, being there by inference. Actual circuits always require a common line which is variously referred to as earth, ground or negative rail.

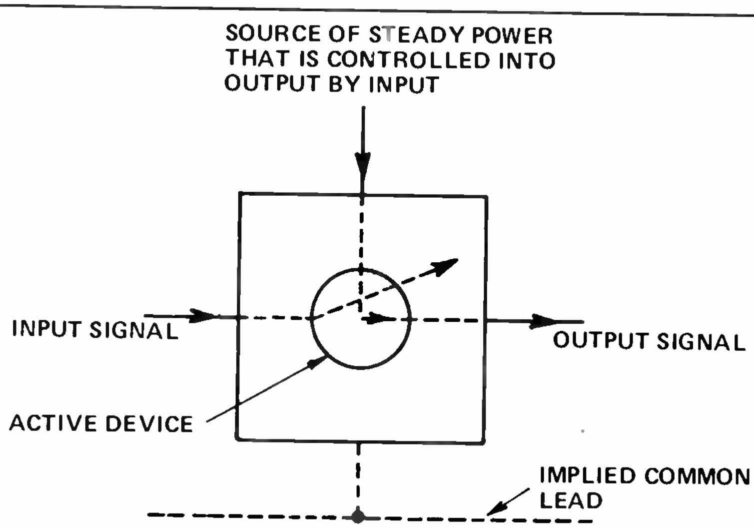


Fig. 1. Schematic diagram illustrating the Amplifier function.

ALL SHAPES, SIZES AND PURPOSES

Although the basic electronic building blocks now available are extremely versatile, there is still no single magic box that can perform all amplifier tasks at the best price *and* performance. Consequently, we make do with many different forms of amplifier to suit an even greater number of applications.

Most amplifiers increase signal voltage amplitude; others, more unexpectedly may reduce it. In both cases we say the amplifier has a gain eg. a gain of 10 – or a gain of 0.1.

The most common need to amplify the *voltage* at the input, but often we may need to increase the current or power level. Yet another need might be to accept a current input and provide a voltage output. The purpose of the amplifier must be clearly understood, for the design and trouble-shooting procedures will differ for each case.

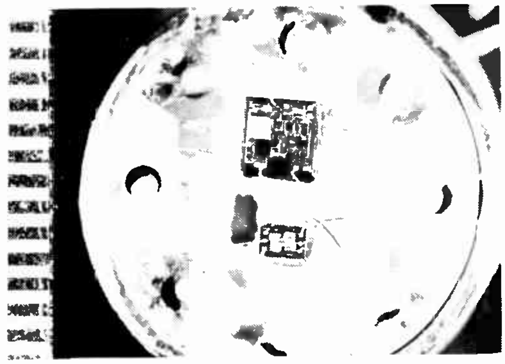
Newcomers to electronics may think that an amplifier must alter the signal/amplitude-level linearly without affecting its time or frequency characteristics, that is, it should amplify with fidelity. This is certainly so with hi-fi audio-frequency amplifiers and with very sensitive transducer amplifiers, but again some amplifiers are designed to distort the signal in some ways to suit a particular purpose. More about these later.

AMPLIFIER JARGON

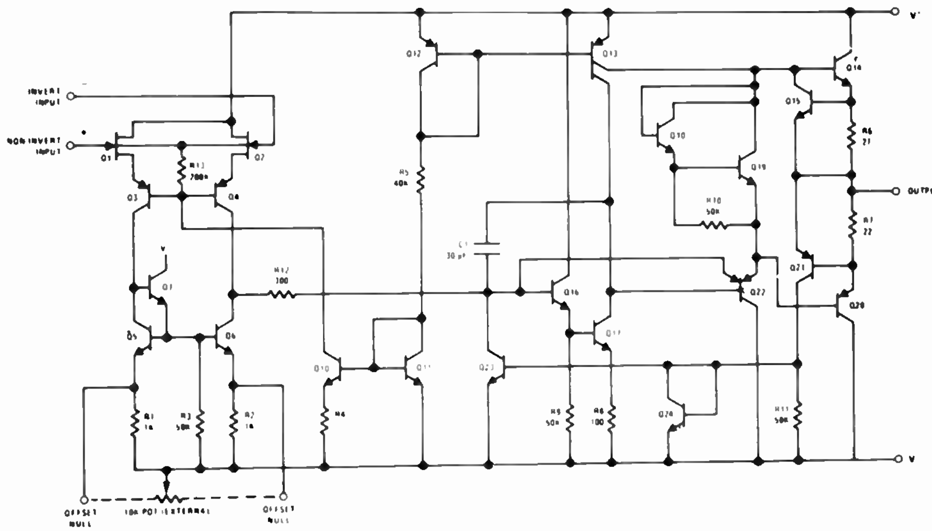
The role of an amplifier is denoted, to some extent, by a prefix. For example a *pre-amplifier* may precede a main amplifier. It amplifies low-level signals (micro-amperes, microvolts and microwatts). Figure 2 shows a string of amplifiers in a typical system.

A *power amplifier* increases the power level of signals in order to drive the output device of the electronic system e.g. the loudspeaker in a hi-fi system; the display tube in an electronic counter. What constitutes a power amplifier and what constitutes a small-signal amplifier is quite arbitrary in absolute terms – the power stage of a digital pocket calculator needs to drive devices rated in milliwatts, but a rolling-mill control may need tens-of-kilowatts capability.

Amplifiers have other applications apart from providing gain. You will



Hybrid FET-input, operational amplifier IC. The small chip contains two FET transistors, the large chip the remaining bipolar transistors. The circuit contained in these two tiny chips is shown on the right. Each division of the scale on the left is ½ mm.



remember in an earlier section, we discussed how connecting a low impedance meter to a high impedance circuit could affect, or even damage, the circuit. This effect, the loading of one stage by another, may be overcome by using an amplifier as a "buffer" between the stages.

Buffer-amplifiers usually have a voltage gain of less than one. However, they do have a power gain and their usefullness is mainly in that their input resistance is considerably greater than their output resistance. Thus the output of a buffer stage can be loaded heavily with little effect on the input. They are, in effect, impedance converters.

Another amplifier characteristic of interest is whether it can handle direct-coupled signals or not. If the signal is coupled to the input via a capacitor, dc signals cannot pass, and such an amplifier is known as an *ac amplifier*. This is not necessarily a disadvantage for, in many systems, only ac signals are of interest.

Another type of amplifier that will often be encountered is the so-called *operational amplifier*. In the early days of electronics, dc amplifiers were difficult and expensive to build because any drift of component values or gain resulted in an unwanted output change. Thus special design procedures had to be used for dc amplifiers, making them very expensive. Nevertheless, they were used extensively in early analogue-computer systems to perform basic arithmetical operations — adding, subtracting, sign inversion and integration — hence their name. (This will be expanded later in the series). Today the operational amplifier can be manufactured inexpensively in integrated circuit form.

In fact, the tables are now turned; the modern operational amplifier is even challenging the single transistor in price, and has tremendous advantages in stability and flexibility, over discrete transistor stages. Indeed these new basic building blocks come close to providing an all-purpose basic amplifier unit.

FREQUENCY RESPONSE

A very small change in the dc level at the input of a dc amplifier will produce a corresponding dc output-level change. The ratio of output to input-level change is called *dc gain*. In an ac amplifier this change is virtually zero because dc signals are not recognised. This does not, however, mean that there is zero dc level at the output, merely that it is unchanged by very-low frequency signals.

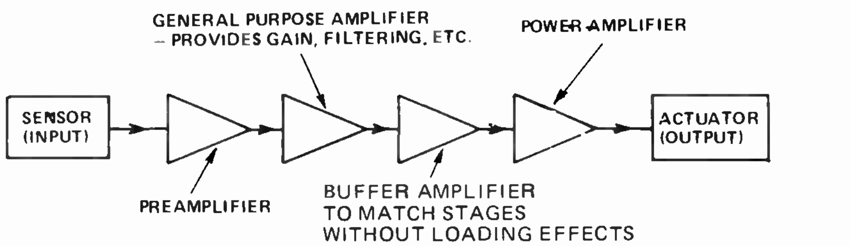


Fig.2. Amplifiers having different functions are often combined in a series chain to achieve an overall purpose.

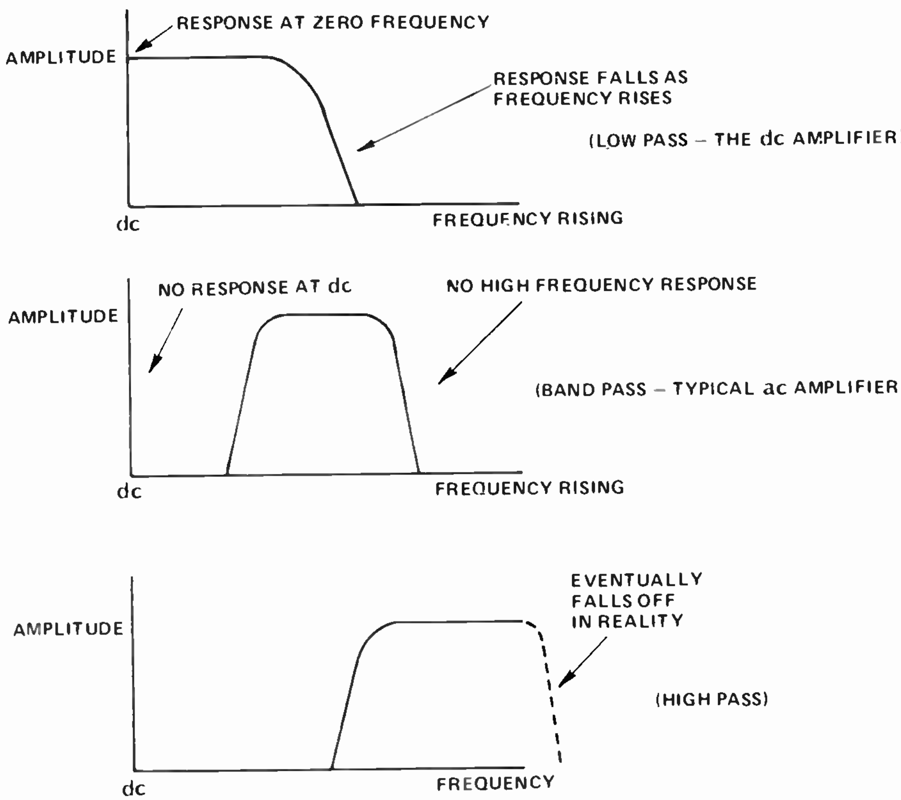


Fig.3. Response curves of amplifiers having three different amplitude/frequency characteristics.

ELECTRONICS –it's easy!

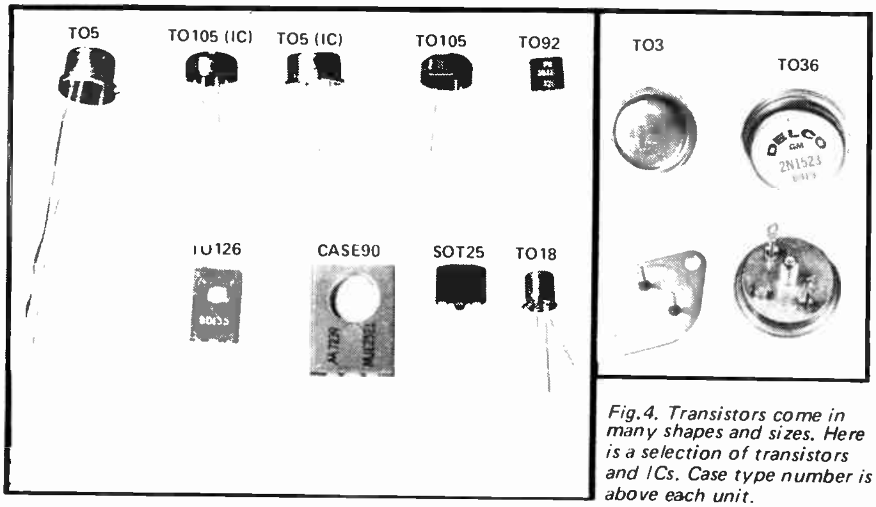


Fig.4. Transistors come in many shapes and sizes. Here is a selection of transistors and ICs. Case type number is above each unit.

The frequency performance of all amplifiers can be shown by two graphs – amplitude versus frequency and phase versus frequency. The first is more commonly encountered. There are other things a designer needs to know, such as time-response to a step-change input, but for the moment we will restrict ourselves to the amplitude versus frequency characteristics.

Physical factors make amplification very difficult at high frequencies. Thus all amplifiers cease to be effective at some upper frequency, but in practice, it is the attainable relative-frequency limit that matters. For example, if the signal to be amplified has no content beyond 20 kHz – as in hi-fi sound systems – there is little point in using a unit with 200 MHz capabilities. This would be more expensive to build and, therefore, a waste of effort.

We use several descriptive terms that denote an amplifier's type of frequency response. Figure 3 shows three main classes – Low Pass (passes only frequencies below a selected cutoff point) Band Pass (passes only frequencies between upper and lower cutoff points), and High Pass (passes only frequencies higher than a selected cutoff point).

Note that the high-pass amplifier still has some upper frequency limit beyond which its response will drop off. The same terms apply to filter circuits – indeed amplifiers can be regarded as filters capable of providing gain.

The frequency response of an amplifier is primarily limited by the active device itself (transistors etc) and secondly by the passive components around the active device which modify its performance. Some amplifying elements will work at megahertz frequencies, some only at kilohertz frequencies. Each have their uses.

PRACTICAL LIMITATIONS

The first active electronic-amplifier element was the triode thermionic valve (briefly described in the last section). This has now been replaced in most applications by the transistor. The transistor does the same job but with less power loss, smaller space requirements and much reduced cost. Several packaged forms of transistor are shown in Fig.4.

The system designer would ideally like amplifiers that accept any polarity of input signal (be it negative or positive with respect to the common lines) and amplifiy it without changing the polarity, or distorting the wave shape in time or amplitude.

Unfortunately neither the thermionic valve, or the transistor, can provide these facilities unless they are used in special ways along with passive elements. Both devices individually will only operate with one polarity of input signal – see Fig.5. If the signal swings to the other polarity, the output disappears: they become rectifiers. Transistors may be constructed to operate with either polarity dc signal, but not both polarities with the same device. That is, they may be constructed as *complementary* units, valves cannot.

Another practical limitation is that these basic devices can only tolerate certain maximum-magnitude signals; as the input signal is increased, a point is reached at which the output signal ceases to increase in amplitude (it gets clipped). If exceeded still further the device may fail altogether. These two effects are the main shortcomings of both valve and transistor, and are illustrated diagramatically in Fig.6.

Eventually an active element may be discovered that does not suffer from these shortcomings; until then we must modify the characteristics of existing active elements in order to obtain the characteristics we need.

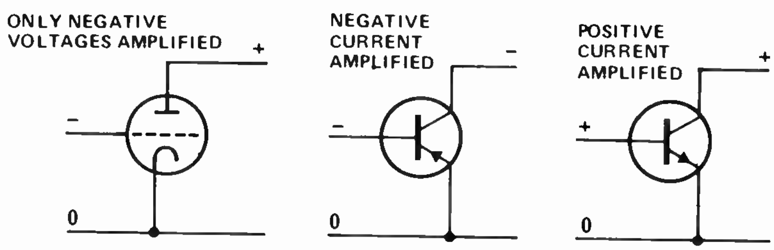


Fig.5. Valves and transistors, when used above, can only handle one polarity of signal. Any other polarity signal is clipped as in a rectifier.

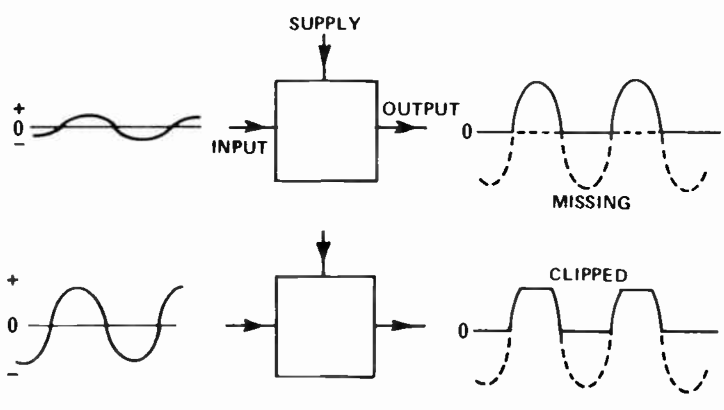
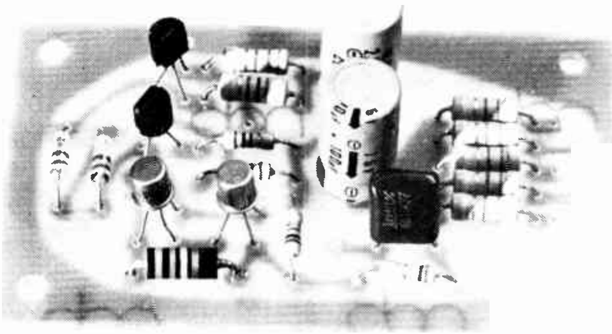
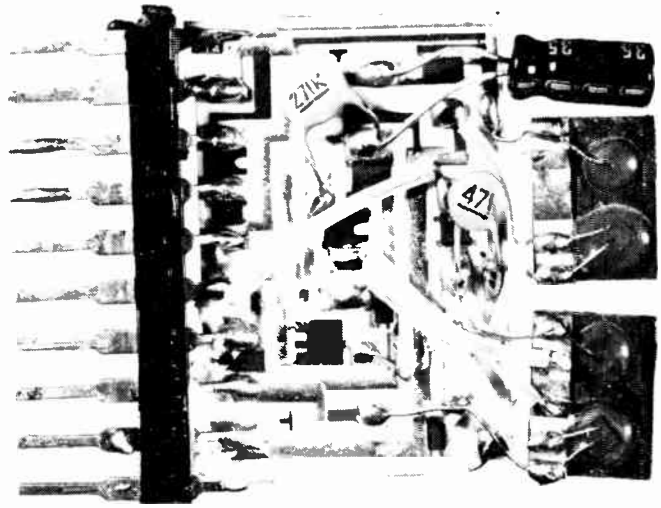


Fig.6a. Effect of feeding a smaller bipolar signal into a transistor. One polarity of half cycle is clipped. (6b). If the input signal is increased sufficiently the tops of the waveform will also be clipped.

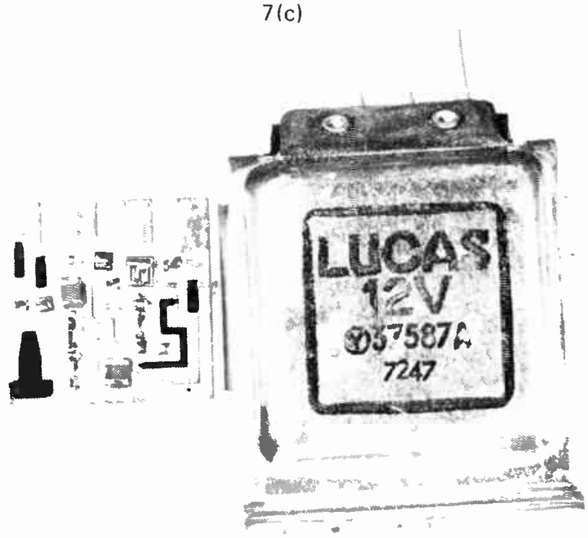


7(a)

7(b)



7(c)



7(d)

Fig.7. Typical amplifiers using devices having differing levels of integration. (a) Typical discreet transistor stage. (b) 25 watt amplifier using hybrid IC module. (c) Internal view of Sanken 10 watt power amplifier of hybrid design. Note power transistors at top of module. (d) Voltage regulator for cars (from Lucas). It contains the thick-film hybrid IC on left which has three transistors, two diodes, two capacitors and five resistors assembled onto a 25 mm square ceramic substrate. See if you can pick the individual components.

This is done by using the device in combination with other active and passive elements to form complete circuit combinations that become our required basic amplifier blocks. Such circuits are either built from individual components – the discrete circuit; or alternatively they are purchased ready designed and manufactured as hybrids – a discrete circuit packaged into one unit. A third alternative is the integrated circuit (the IC) in which all active and passive elements are fabricated on a common substrate. Figure 7 shows several modern amplifiers based on the transistor amplifying element.

AMPLIFIER CHARACTERISTIC CURVES

The various types of individual amplifier elements behave differently, have different signal-level handling ability and have different input-to-output signal ratios (gain). Furthermore, the gain may depend upon the amplitude of the input signal and on what is connected to the output.

The information, needed by a designer on device characteristics is commonly provided by graphs known as characteristic curves. We met the simplest form of curve when we discussed the light-dependent resistor in Part 2 of this course. In that case there was only one relationship – that of resistance versus light level.

The problem of presenting characteristic curves for amplifiers is more complex than for the light-dependent resistor, for there are an infinite number of describing curves. To understand this, consider the relationship between the supply current (I) flowing into an active element (Fig. 6) and the voltage developed at the output (V out). It is not possible to draw a unique single

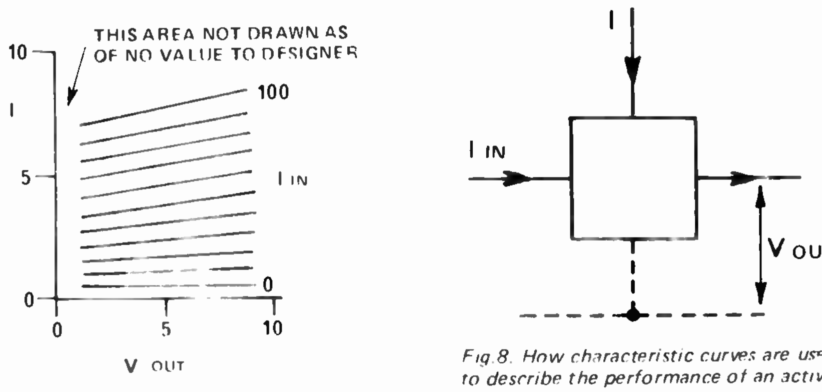


Fig.8. How characteristic curves are used to describe the performance of an active device.

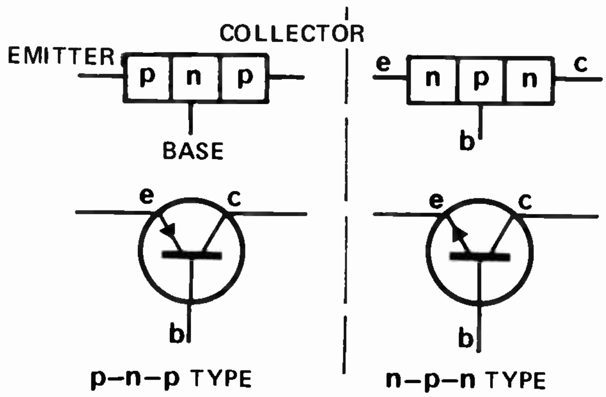
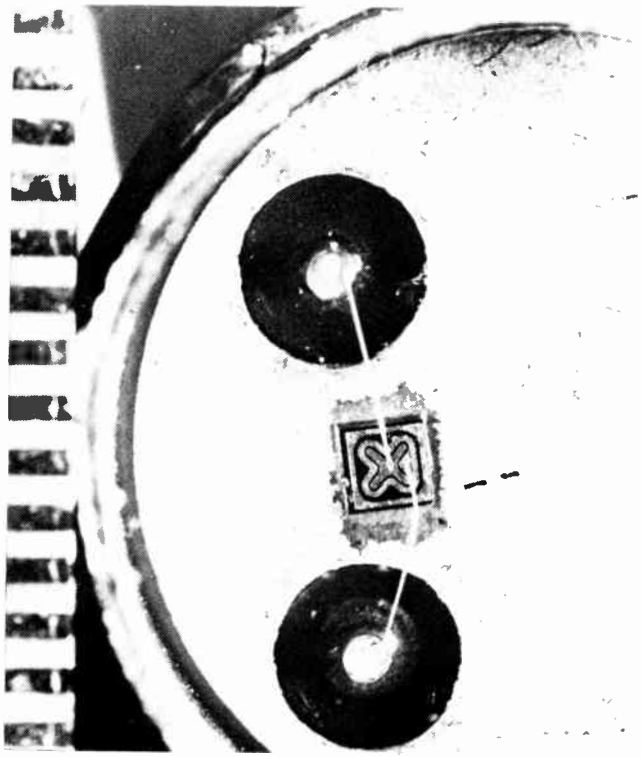


Fig.9. The basic structure and symbols for the two elementary transistor types.



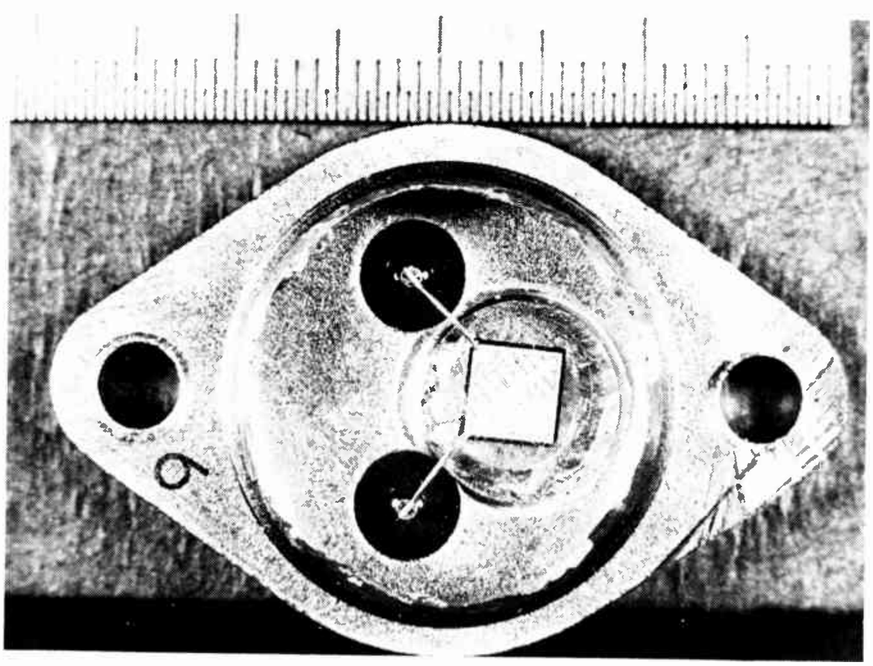


Fig.10. The actual transistor chip is indeed tiny – as these enlargements show. The scales at left are half millimetre divisions. The smaller chip is a high-frequency small-signal transistor, mounted in a TO5 case, the other a power transistor mounted in a TO3 case.

graph, as the relationship depends upon the signal current into the input terminals – call it I_{in}. For each value if I_{in} there will be a specific graph of I versus V (out).

A convenient way of representing what happens is for us to draw individual curves at evenly-spaced, realistic values of I_{in}. The result is a family of curves as depicted in Fig.8.

A little thought shows that other families can be constructed also, output-voltage versus input-voltage for various values of input current is one. Furthermore the fixed parameter – could be input voltage instead of current – as is the case for valves.

The characteristics of both valve and transistor devices can be visualised this way (as indeed can any type of three terminal amplifier) and these curves are of great value to designers.

Most people engaged in electronics do not need to measure the characteristic curves for themselves; they are provided in manufacturers' data sheets. It is important for us to understand these curves, for they help explain how the non-ideal characteristics of active elements (discussed above) are overcome in practical circuits. Before discussing how this is done we need to know more about the transistor itself.

THE TRANSISTOR

Transistors are made from two basic materials – germanium or silicon. These two materials are known as semiconductors because they are neither good insulators, nor good conductors. That is, they are somewhere in between.

Germanium was used for early transistors, but has largely (although not entirely) been replaced by silicon in modern devices. Although there are some important differences between transistors constructed from these two materials, the basic theory, as follows, is the same.

The basic pure material is modified by adding a controlled amount of impurities called dopants, to form two new materials, one (called P type) having a deficiency of electrons and one (called N type) having a surplus of electrons.

If two pieces of these differently doped materials are intimately joined we have what is called a PN junction. Such a junction of P and N materials will conduct current more readily in one direction than in the other – it is in fact a rectifier, or in other words, a semiconductor diode.

Current flow occurs when the P type material is made more positive than the N type material. The physics involved in this phenomenon are complex, but of little interest at this stage. We are only interested in the fact that it happens.

To make a transistor we add a third layer of material to form a three-layer sandwich in either NPN or PNP format. We refer to the transistors in this way — as a silicon NPN or PNP type etc. The symbols for the two types are shown in Fig.9. Each terminal is given the name as shown, the base being the centre connection, the emitter the one marked by an arrow and the collector unmarked. Note particularly that the direction of the emitter arrow denotes whether the transistor is a PNP or NPN type also that the symbol is the same for both germanium and silicon devices.

In actual manufacturing processes the three layers are formed by selectively growing N and P crystal layers, or by diffusing P and N impurities into the opposite sides of a pure, silicon or germanium crystal.

The actual transistor chip may be extremely small, often pin-head size and is generally a tiny fraction of the total packaged volume of the device. This is illustrated in Fig.10 which shows the inner construction of different types of transistor. From this we see that although small, a conventionally packaged transistor wastes a relatively enormous amount of space. Integrated circuits, where both active and passive components are made and connected by layering and diffusion processes, are logical developments from transistor technology — it is just as easy to fit 20 or 100 transistors in a T05 case (Fig. 4) as it is to fit one.

The main problems in integration are in limiting power dissipation within a given chip or case, and in fabricating resistors and capacitors.

SYMBOLS

As we go further in electronics we must use shorthand methods of expressing things — otherwise explanations tend to become unwieldy. For example, in our discussion of transistor parameters we will be considering the currents, voltages and impedances etc. associated with each lead of the device. To avoid having to write for example, "current in the collector lead" we simply write I_C. The main symbol I tells us we are concerned with current and the subscript 'C' tells us that it is the collector lead we are talking about.

Thus E_b = base voltage
E_c = collector voltage
I_b = base current
E_{ce} = voltage between collector and emitter.

Now that we have established our shorthand we are in a position to examine the practical characteristics of transistors.

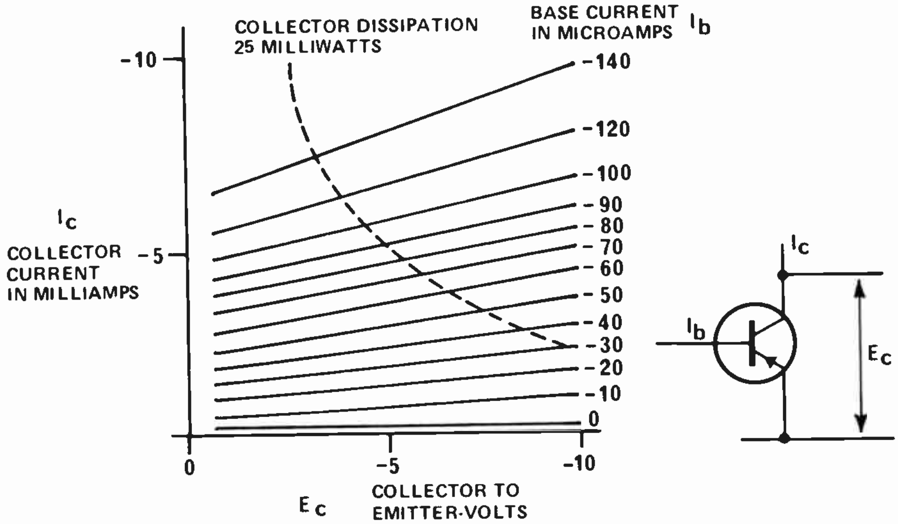
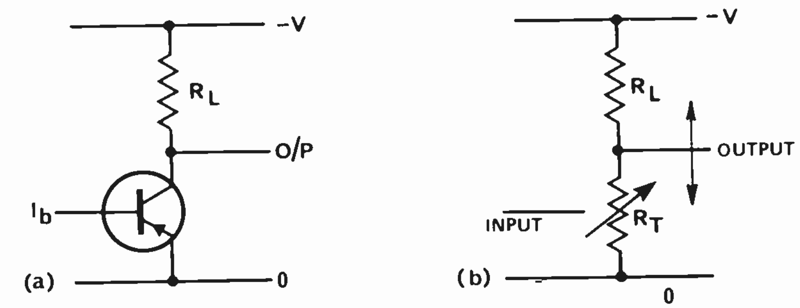
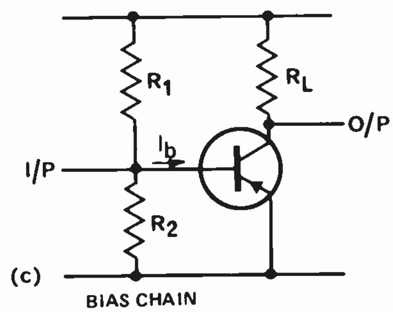


Fig.11. Typical characteristic curves for a small signal PNP-transistor.





$\frac{R_1}{R_2} \approx 4$ $\frac{R_1}{R_L} \approx 10$

Fig.12(a). To use a transistor as an amplifier we first need to insert a load resistor RL. (b) The load resistor together with the resistance of the transistor become a voltage divider. (c) To amplify ac signals we must use a bias current to offset the operating point.

CHARACTERISTIC CURVES

Let us examine what happens if we hold the collector-to-emitter voltage, E_{ce}, constant at -5 volts and then vary the base current, I_b from 50 to 60μA we find that we have a corresponding I_C change of 500μA (0.5 mA). Thus we have a gain, β of 500/10 = 50.

Note that corresponding changes in I_b at other points (e.g. 90 to 100μA) does not result in the same gain. In fact, there is non-linearity at extremes of I_b which would result in distortion of the signal.

In practice it is not necessary to perform these calculations, the manufacturer tells us the gain in his data sheet. This is referred to as β or Hfe (don't worry about interpretation of this latter symbol) and is the ratio of the change in collector current resulting from a small change in base current.

That is $\beta = \frac{\Delta I_c}{\Delta I_b}$ (Δ means small change in)

Values of β range from 5 or so for early transistors to several hundred, or even thousands in modern components. Manufacturing tolerances don't allow all transistors of any type to have the same β' and the manufacturer usually specifies the limits within which the device current-gain will fall.

For example the BC108 is a popular audio transistor specified as having Hfe (β) greater than 125 but less than 900 at I_c = 2 mA and Vce = 5 volts.

Referring back to Fig.11, we find a dotted line across the curves which represents the maximum permissible

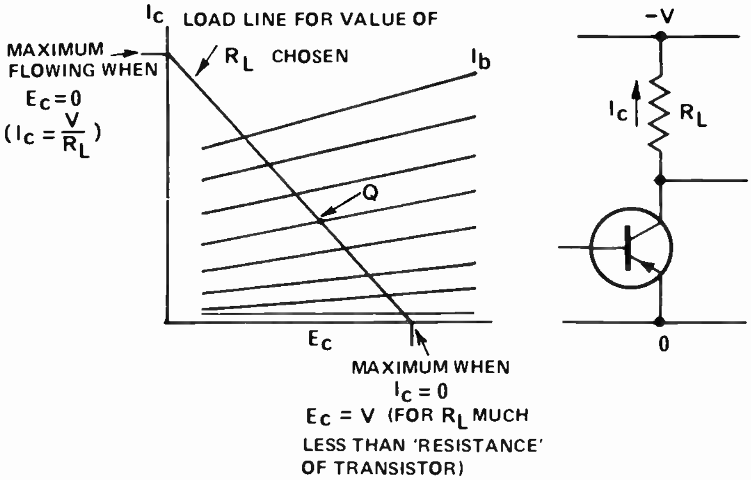


Fig. 13. Output impedance of the stage is equal to the value of R_L. R_L is generally chosen to be about one-tenth of the input impedance of the following stage provided that $(\frac{V}{2})^2 R$ does not exceed the rated device dissipation. The load line is then drawn and operating point Q determined as detailed in the text.

midway (this is called the Q point) along the load line and both half cycles of our input signal will be amplified linearly.

The degree of distortion is decided by the extent to which the input signal varies I_b up and down about the Q-point. Small signal changes will be undistorted but not large ones. One cause of this is that the gain I_c/I_b will change at the limits.

Secondly, if the input signal increases still further, the peaks of the sine wave will be clipped, at one end by the base current reaching zero, and at the other by the collector voltage being driven to zero (this latter condition is called saturation). Hence it can readily be seen, from the characteristic curves and load line, what maximum input signal can be applied without distortion occurring.

Note that the load line must always lie below the maximum power dissipation curve.

power dissipation for the device. This is determined by the maximum heat that can be dissipated without the chip being destroyed, unless the device is cooled by a heat sink or with forced air circulation.

For example if the transistor of Fig.11 has an E_{ce} of 5 V, then the collector current must not exceed 5 mA if the dissipation is to be less than 25 mW. Thus the user must check his design to ensure that under worst case conditions (component tolerances, power supply voltage etc.) this dissipation is not exceeded. The device must never be operated at any point above and to the right of the dissipation curve.

Thus we see that much information can be extracted from the characteristic curves.

THE BASIC AC AMPLIFIER

Used alone, the transistor cannot amplify ac waveforms. The two main limitations are its inherent rectifying action and an effect known as thermal runaway. In addition we must devise a way of taking an output from the transistor.

The transistor may be considered as a resistor whose value is varied by the input base current. Hence, if we place a resistor in series with the collector lead of the transistor, we will have a voltage divider as shown in Fig. 12. The collector current, as it changes in response to changes in base current, will produce an output voltage across the series resistor. This series resistor is called the 'load' resistor and is denoted by the symbol R_L.

Note that to drive more current into the base we must raise E_b towards the collector supply voltage. The resulting increase in I_C will cause the voltage at the collector, E_C, to fall. Thus the output voltage will be the inverse of the input. In other words, the transistor connected in this fashion, changes the phase of the input voltage by 180°.

BIASING

If a sine wave were to be applied to the base of the transistor in Fig. 12a, the negative half cycles would be clipped off — the waveform would be rectified as previously explained. We can overcome this by applying a dc 'bias' current to the base such that the input signal either adds or subtracts from this current but *never* drives the base current to zero.

Hence the collector current will also be biased away from zero and will follow the variations in base current. In practical circuits it is not feasible to have a separate battery or power supply to provide bias, so it is usually derived from the collector supply. The most common method is by using a voltage divider as illustrated in Fig. 12c.

Biasing can also be illustrated using characteristic curves. For any chosen R_L value, there will be corresponding pairs of I_C and E_C values — Ohms law again. This means for any value of R_L we can draw a line — called the load line — across the characteristic curve as in Fig. 13. The importance of this curve is that the input signal, I_b moves up and down this line. If we do not add a bias current to I_b we would be operating at the bottom end, where I_b = 0, and only negative swings of I_b would be amplified. By adding a *quiescent* bias current we put the mean operating point at a place

THERMAL RUNAWAY

As well as the currents I_b and I_c that are designed to flow in the transistor there is leakage current through the normally reverse-biased, collector-base junction. Some of this current will flow through the base-emitter junction (actually all of it if the base is not connected) appearing as a normal signal. The apparent signal current will be amplified causing an I_c of βI_b.

Now here is the danger — the leakage current is proportional to temperature. So the increased I_c heats the transistor, the leakage current increases, I_c increases still further — and the process may continue until the transistor destroys itself.

The actual process is more involved than we have described but the explanation suffices for our purposes.

With silicon transistors leakage current is very small and of little importance but silicon has another temperature effect that produces similar, although not as serious, thermal runaway. This is that the E_{be}

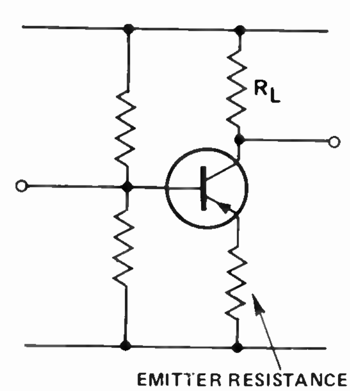


Fig.14. The amplifier is stabilized against the effects of thermal runaway by adding an emitter resistor.

ELECTRONICS -it's easy!

of a silicon transistor, required for a certain collector current, falls with temperature. Hence with a fixed input voltage the resultant I_c causes a rise in temperature, which causes a decrease in E_{be} required, and hence a further rise in I_c — result thermal runaway.

Silicon transistors can be used over a much wider range than can germanium but thermal runaway must be compensated for with both types.

Fortunately this potentially damaging effect is easily overcome, in both cases, by adding a resistor into the emitter path as shown in Fig. 14. Its effect is as follows.

As the collector current rises (due to leakage current) the voltage dropped across the emitter resistor, Re, increases thus reducing the base-emitter voltage V_{be}. This reduces the base current, and almost restores the collector current to its original value. Mathematics tells us the ideal conditions necessary to achieve stability — an emitter resistance roughly one sixth of the collector load resistance is about right. The bias chain values must be readjusted for this and again there are complex mathematical expressions for optimizing the values. In practice a good choice is for the chain to have values in the same ratio as the collector chain but about ten times larger.

BYPASSING

Having overcome thermal runaway conditions we now find the amplifier is nicely stable but lacks gain. This is because the same collector current flows through the emitter resistor as through the load resistor. Hence the gain can only be equal to the ratio of R_L to R_e, that is, in our case 6. And this is completely independant of β. We can restore our gain by adding one more component— a capacitor across R_e.

Thermal effects occur slowly by comparison with ac signals (10 Hz and above) so a capacitor connected across the emitter resistor will act as a low impedance to ac signals (thus restoring ac gain) but as a non-existant component to dc. Hence we get the best of both worlds — thermal runaway is eliminated and ac gain is maintained. The capacitor is chosen such that its reactance is about one tenth the value of R_e at the lowest frequency of interest.

REFERENCES

"Solid State Basics" and "A Course in Radio Fundamentals" both published by the American Radio Relay League; available from most technical bookshops and also from Electronics Today International, 15 Boundary Street, Rushcutters Bay, NSW.

ELECTRONICS – in practice

THE CIRCUIT of a typical ac amplifier, for audio frequencies, is given in Fig. 15. The input signal is coupled in via a capacitor that provides dc isolation between the preceding stage and the bias network.

As the capacitor needs to be fairly large (X_c less than one tenth the resistance from base to ground at lowest frequency) it is usually an electrolytic. An electrolytic may be used as long as the positive terminal is connected to the most positive dc potential.

The circuit uses a readily available, inexpensive transistor and may be put to work (and tested) by adding the components as shown in Fig. 16.

In effect we now have a light intensity meter which can be used to monitor the modulated content of the radiation from a fluorescent-light tube. Note that it does not measure the steady-state light radiation.

The light-dependent resistor, type ORP12, provides a small amplitude 100 Hz signal when excited by the light from a fluorescent tube. The amplifier increases the signal amplitude by about forty times. The output from the amplifier may then be half-wave rectified to provide a dc output proportional to the level of the 100 Hz light signal. This may be measured by a normal multimeter, or alternatively, the ac signal may be fed directly to high impedance headphones. You will then hear the 100 Hz tone from the light radiation.

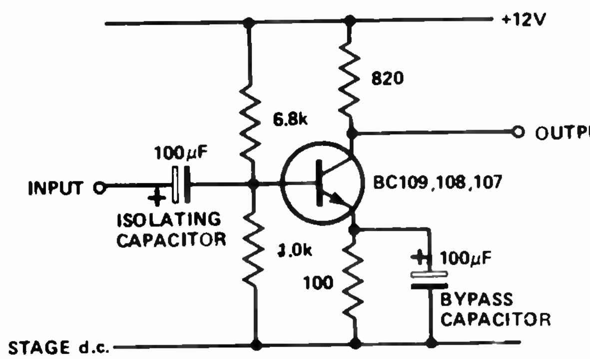


Fig. 15. The addition of a bypass capacitor restores ac gain and an input capacitor provides isolation for the input. We now have a complete ac amplifier. Two or more stages can be used in series if more gain is required.

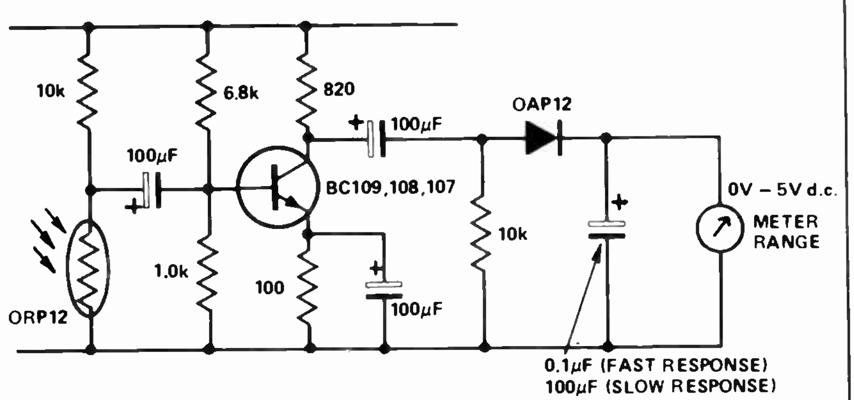


Fig. 16. The modulated light output from a fluorescent lamp may be amplified and measured by adding just a few more components. This also provides a simple test of correct amplifier operation.

11 Emitter followers and dc amplifiers

IN OUR GENERAL discussion of the ac amplifier we have shown how this common system component can be used to increase the amplitude of ac signals. The frequency range of a typical ac amplifier may extend from a few hertz to many megahertz, the upper limit being limited mainly by the performance of the active device (transistor etc) used.

The type of amplifier illustrated, in Fig.15 of the previous section, is only one of several different basic design concepts that we may use. Each different approach has particular advantages and disadvantages that suit it more to one application than another. For example a different approach must be used where the amplifier is to be loaded by a low impedance or where dc signals must be amplified.

We will now discuss the three basic amplifier configurations, their fundamental properties and their typical uses. This will then equip us with the information needed to understand basic dc amplifiers and, in addition, modern integrated-circuit amplifiers.

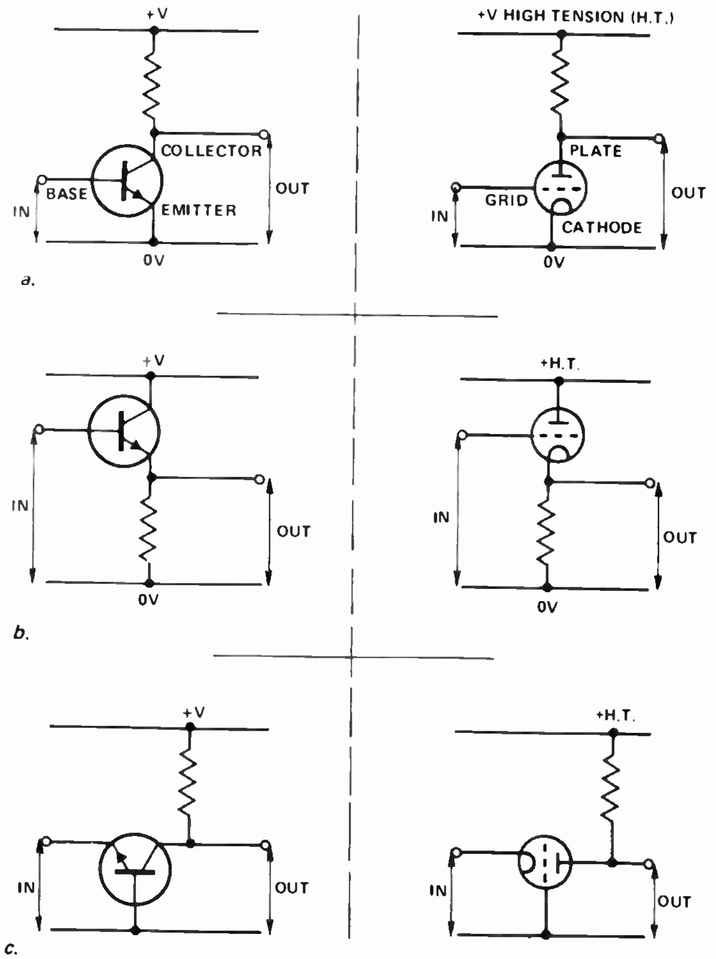


Fig. 1. Basic configurations of transistor (left) and valve (right) amplifiers. (a) Grounded emitter, grounded grid. (b) Grounded collector more commonly called emitter follower, and grounded plate more commonly called cathode followers. (c) Grounded base and grounded grid.

In our previously described amplifier you will remember that we developed the circuit from a basic configuration of the transistor where the emitter is connected to the negative rail. This method of connection, naturally enough, is known as "grounded-emitter" and is illustrated in Fig.1a (together with the equivalent valve circuit). This, and the other methods of connection were first devised for use with valve circuitry and then conveniently passed on to transistor technology. There is one major distinction between the two amplifier elements – valves operate as voltage devices, transistors as current devices. In the 50's and early 60's, transistor technique was taught by using analogues with the then established and widely known valve practices. Today, valves play only a limited part in electronics, but we have included valve counterparts alongside the transistor circuits to assist those previously trained in valve technology and the newcomer too will be acquainted with components that are still used in some special applications.

THE EMITTER FOLLOWER

Another valuable configuration is that given in Fig. 1b – the grounded collector circuit which is more commonly called an emitter-follower.

In this case it is the collector that is connected directly to the supply rail, not the emitter. The term 'grounded' may appear incorrect but, when it is remembered that a perfect voltage supply has zero resistance, it can be seen that the collector is effectively connected directly to the ground line. As it is much a case of where the essential load resistor is placed it might be easier to remember that this configuration places the load resistor in the emitter lead, *not* the collector lead. The transistor is wired into the circuit with the same polarities at each connection as for the grounded emitter.

In the development of a satisfactory grounded-emitter circuit we saw how the addition of an emitter resistor provided thermal stability. We also saw how this resistor reduced the dc gain of the circuit. This is because the collector current through the resistor produces a voltage drop which opposes the original drive voltage applied to the base-emitter circuit.

The emitter follower uses this effect to provide impedance buffering between two stages. The output voltage developed across the resistor is closely equal to (but not quite the same because of the voltage drop across tne base-emitter junction) that applied to the input.

This may seem a futile process for the voltage level of the signal cannot be amplified. In fact, however, it is the signal current that is amplified. The emitter-follower, therefore complements the operation of the grounded-emitter circuit. It is invaluable as a means to raise the current level of signals without altering the voltage level.

Although current gain is very important in some applications (discussed later), in small signal situations we usually regard the emitter follower as an impedance — conversion stage. This will become more obvious after we examine emitter-follower characteristics.

The voltage drop across the forward biased base-emitter junction is a constant (almost) 600 mV for a silicon transistor (400 mV for germanium). Thus the voltage at the emitter closely follows the signal at the base, but with a 600 mV lower mean dc level.

Hence the voltage gain of the emitter follower is always slightly less than unity.

$$\text{gain } A \simeq \frac{Ze}{Ze + (\frac{1}{gm} + \frac{Zs}{\beta})}$$

Where

Ze = impedance in emitter

$\frac{1}{gm}$ = a factor dependant on resistances within the transistor but typically 50 ohms at 1 mA for small transistors (falls with increasing current).

Z_S = source impedance

β = transistor current gain.

Thus if an emitter resistor of 1k is used with a transistor having a β of 100 and the impedance of the source is 2k.

Voltage gain A =

$$\frac{1000}{1000 + (50 + \frac{2000}{100})} = \frac{1000}{1070} = 0.93$$

Input Impedance

$$Zin \simeq \beta\,(Ze + \frac{1}{gm})$$

Thus for our example

$$Zin = 100\,(1000 + 50)$$

$$= 105\text{ k}$$

Output Impedance

$$Zout \simeq \frac{Zs}{\beta} + \frac{1}{gm}$$

$$= \frac{2000}{100} + 50$$

$$= 70\text{ ohms}$$

The actual output impedance is the value as calculated above in parallel with the emitter resistor. That is, in our case —

70//1000 = 65 ohms

Thus we can see that the input impedance of 105 k will not appreciably load the 2 k source and the full signal voltage (0.93 gain) appears at the emitter across an impedance of 65 ohms. Thus we see how the impedance conversion has taken place. By choosing the correct values this impedance may be adjusted to a desired value, eg 50 ohms.

The emitter follower, therefore, can be used to connect a low input impedance stage to a preceding high-output impedance stage without introducing serious attenuation due to loading.

For example, if a stage with an output impedance of, say 10 k is to drive a stage with 1 k input as shown in Fig.2 a direct connection would load the first stage so much that its signal voltage output level would be reduced to roughly one tenth of its original magnitude. A single emitter-follower stage can be designed to have 100 k input and 50 ohm output which will enable the original two stages to be joined with little attenuation of the signal level. Figure 3 shows a typical circuit in which voltage gain is obtained by a grounded-emitter stage followed by buffering with an emitter follower.

The amount of impedance reduction

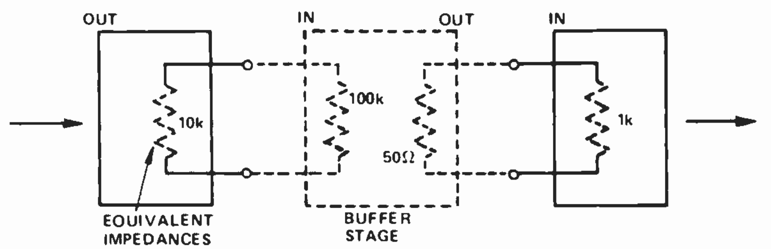
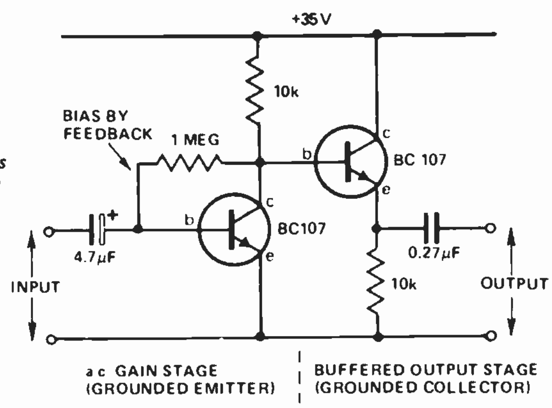


Fig. 2. The emitter follower, used as a buffer stage, allows two stages to be coupled without introducing loading effects.

Fig. 3. A typical grounded emitter stage buffered by an emitter follower. Note the bias arrangement for the first stage and that base of the emitter follower is directly coupled.



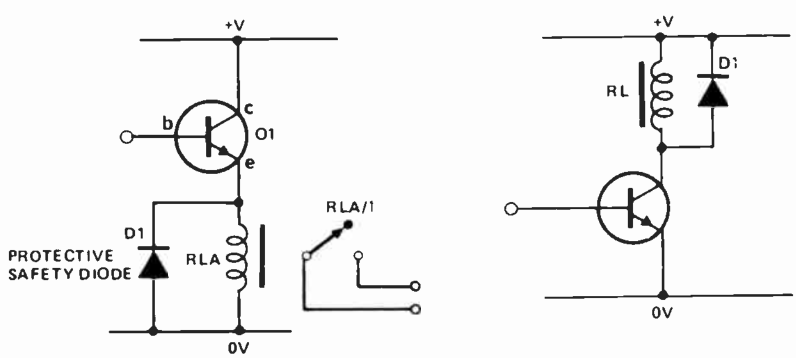


Fig. 4 (a). In output stages the load becomes the resistance in the emitter of the follower circuit. (b). The alternative grounded emitter circuit note that diode D1 absorbs high voltages induced in the relay coil when the transistor turns off.

	GROUNDED BASE	GROUNDED EMITTER	GROUNDED COLLECTOR
Current Gain	<1 (.98)	High (200)	High (200)
Voltage Gain	High	High	<1
Input Impedance	Low (40)	Medium (2 k)	High (100 k)
Output Impedance	High (1 Megohm)	Medium (30 k)	Low (1 k)
Power Gain	Medium (30 dB)	High (40 dB)	Low (16 dB)
Cut-off Frequency	High	Low	depends on R_L
Voltage Phase Shift (L.F.)	Zero	180°	Zero

Fig. 5. Comparison table of characteristics for alternative connection modes. Note that these are typical values only.

attainable depends largely upon the β value of the transistor. Where greater than tenfold reduction is needed the designer can resort to cascading two or more emitter-follower stages or use can be made of special semi-conducting active devices (eg the field-effect transistor) that have high input impedances.

When the buffer stage is also the final output stage, and is required to drive an actuator such as a loudspeaker or relay coil, the actuator itself may be used as the emitter resistor, being wired into circuit as shown in Fig.4. In this case there is no need to provide a separate resistor.

The emitter follower does not change the phase of the signal. This contrasts with the grounded emitter amplifier where a positive-going signal becomes a negative going output. That is, ac signals are phase shifted by 180° or one half-cycle.

The emitter-follower is a robust stage and is less likely to be damaged than the grounded-emitter circuit. The main point to watch is that the emitter load impedance (the resistance value added in parallel with the input resistance value of the next stage) is not so small that the collector current I_C exceeds the manufacturer's stated safe maximum value.

The input base connection of the emitter-follower stage is usually coupled directly to the output (collector) connection of the preceding stage. There is no need for thermal runaway compensation or for a bias network. The emitter-follower is a very simple stage but nevertheless a very important one.

GROUNDED BASE CONFIGURATION

Having grounded, firstly, the emitter then, secondly, the collector the next obvious stage design is to ground the base.

The schematic of a grounded-base stage is given in Fig.1c. This design is seldom used but there are circumstances where its peculiar characteristics render it useful. It can provide voltage gain roughly equivalent to the β of the transistor but it cannot provide more than unit current gain. Its usefulness arises from its ability to couple low-impedance input transducers – microphone transformers for example – to normal grounded emitter gain stages with optimum power transfer. The input impedance of the grounded-base stage is in the region of tens of ohms and the output impedance is near a megohm.

The characteristics of the three configurations for a transistor are tabulated in Fig. 5. Such a table can only be used as a guide, for the actual values of each circuit depend largely upon the β of the device and upon the passive components wired in to form the practical working stage – all remarks made apply to both p-n-p and n-p-n transistors alike; only the polarity of the supply needs to be changed.

THE COMPLEMENTARY TRANSISTOR

The foregoing explanations implicitly suggest that the emitter of a p-n-p device must always be connected to

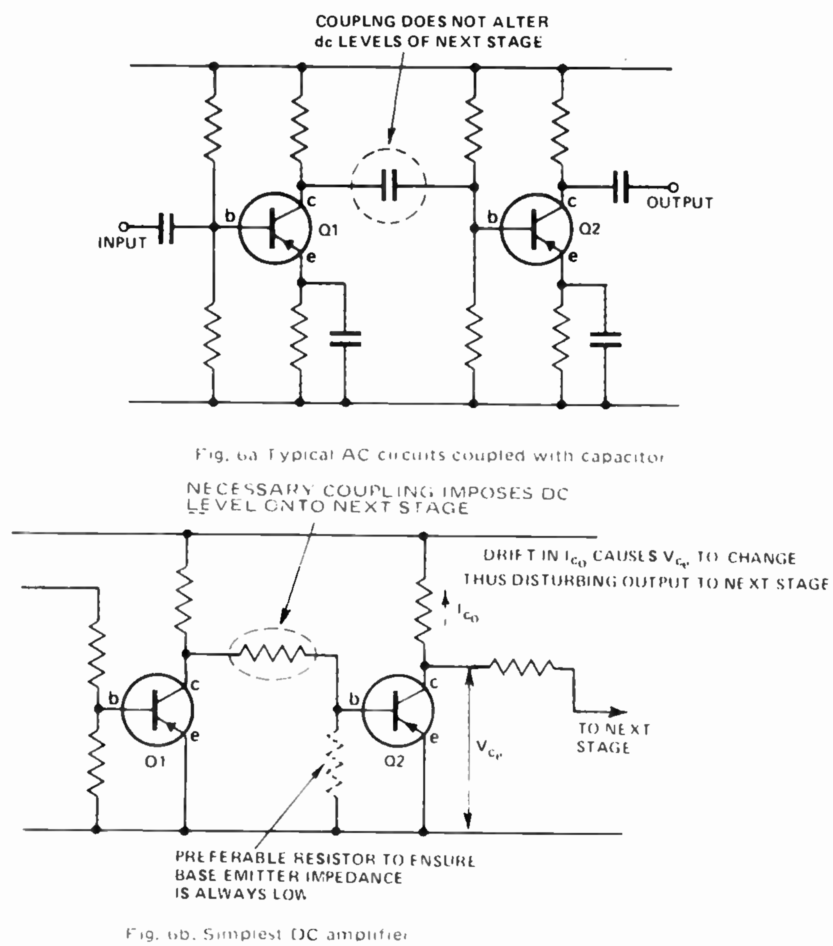


Fig. 6 (a) Coupling between ac stages by a capacitor does not affect levels of either stage.
(b) Direct coupling between dc stages creates problems because of imposed dc levels.

the positive polarity and that of a n-p-n to a negative polarity. This is usually the case in practice — but not an absolute rule.

Remembering that the transistor is a three-layer device we can see that it is, in principle, symmetrical. The p's of a p-n-p device could, in principle, be either the emitter or the collector, implying that it could be connected either way into a circuit. In practice, the junctions are made in such a way that operation is optimized for the connections stated by the manufacturer. It is, however, possible to procure special transistors that are made to exhibit similar characteristics for both possible connections of the collector and emitter, but one seldom meets the need for this in electronic circuits.

DC AMPLIFIERS

We have seen how it is necessary to add passive components to a basic active element to construct a practical ac amplifier. The same applies to constructing a practical dc amplifier.

To better understand what is required let us examine the different requirements of ac and dc amplifiers.

In the ac amplifier two different design conditions exist together, the bias and other steady state conditions and, the ac coupling which allows the signal to cause variations around these steady state conditions. This is necessary so that both polarities of the ac waveform may be amplified. Thus each stage in a chain of ac amplifiers is self-contained that is, the dc levels of one stage are not imposed on the next. This is illustrated in Fig. 6a.

If the signal to be amplified is a dc level (including also signals below 5 Hz) it is not possible to isolate the steady-state conditions of successive stages and some means of direct connection must be used.

Figure 6b illustrates a basic method of interconnecting dc amplifiers by means of a resistor. It is obvious that the dc level at the collector of Q1 will cause current flow into the base of Q2 and a corresponding collector current in Q2. This implies that with no signal to the base of Q1 the output voltage from Q2 will not be zero, and its level will depend on the conditions in the previous stage.

From this we see that the first important requirement is to *carefully* select resistor values such that the following stage is not driven into saturation. The series coupling resistor is thus chosen to limit base current into Q2. It must not be too high, however, because the dc signal will be attenuated by the ratio of this resistor to the base-emitter resistance of the following transistor.

A further problem is that when the input to the base is zero, the collector output is not zero but at the supply voltage. So that even with no input to the first stage, the second may well be saturated.

Thus this particular approach, whilst capable of providing some dc amplification, is not very practical. Resistors must be chosen to suit actual betas of the transistors used as a change in gain means a change in output current and in bias to the next stage.

Assuming we managed to establish a workable set of values the next problem is that the values of the components and the gains of the transistors may (and do!) change with age and, more spectacularly, with temperature. Thus if the output (dc) voltage is somehow set so that it is zero with zero level input, it subsequently will drift in time and with temperature. If the overall gain is 100 000 (typical value) it does not require much drift at the first stage to fully saturate the last stage! (A 10 V swing is produced by a tenth of a millivolt change).

Finally, to add to the problems to be faced by the designer we have not overcome the problem of amplifying both positive and negative polarity signals; the schematic arrangement of Fig. 6b can only handle negative signals. Positive signals merely bias the input stage into a totally non-conducting stage. An n-p-n equivalent (of Fig. 6b) handles positive signals but not negative.

TURNING THE TRANSISTOR ELEMENT INTO A WORKING DC AMPLIFIER

A decade ago the electronic system builder had to design and build his own dc amplifiers Commercial units were available but were very expensive. The dc amplifier was regarded as a system block best avoided if possible! Numerous designs were investigated in an attempt to overcome the problems in a satisfactory way but it was not until 1936 that the first successful high-gain dc amplifier was built (in Sweden) by Buchtal and Nielson. Since then many

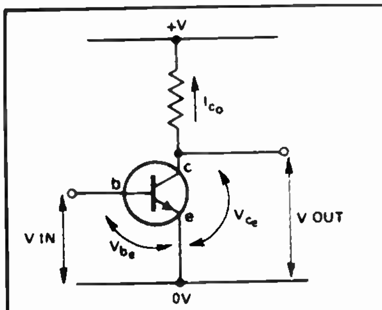


Fig. 7. Basic grounded-emitter stage of a dc amplifier is rarely satisfactory due to drift in V_{OUT} caused by changes in leakage current I_{CO}.

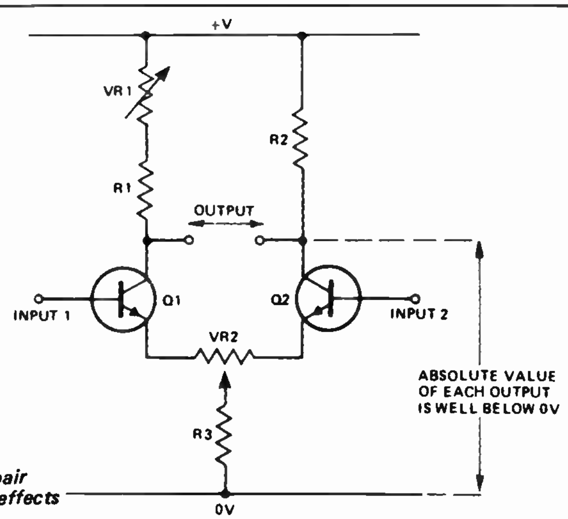


Fig. 8. Schematic of the differential pair arrangement which is used to reduce effects of thermal drift.

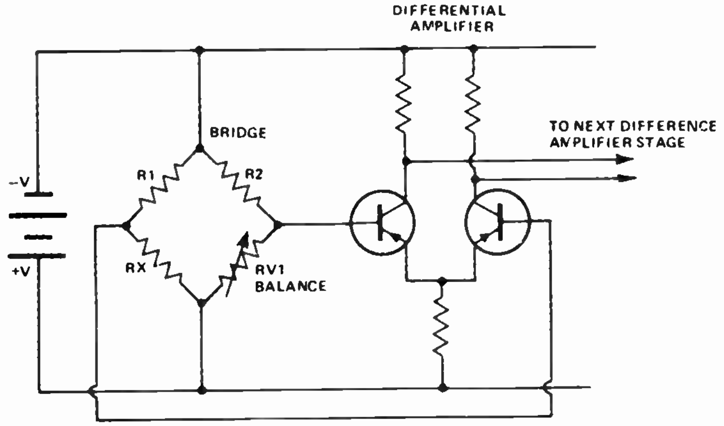


Fig. 9. Differential dc amplifiers are ideal for use with Wheatstone bridge arrangements and sensors of various kinds. The circuit provides high common-mode rejection. That is, it rejects noise, hum etc and only amplifies the difference signal.

intriguing circuit 'tricks' have been devised to overcome the drift problem that is still encountered with this and subsequent designs.

Today the situation has, quite suddenly, been reversed and we more often than not use a dc amplifier to provide the relatively simple ac amplification than build a special-purpose ac amplifier.

This revolution has come about with the use of integrated circuit manufacturing methods whereby numerous elements — (typically, 20 transistors, half as many resistors and a capacitor or two) are formed into a dc amplifier that, now markets for \$1.00 or less and, may be mounted in a space about 5 mm square.

The circuit requirements of a dc amplifier stage have not been eased; in fact a modern amplifier in integrated form contains more elements than its earlier discrete predecessor. Now, a few highly specialised designers devise the IC circuit which, after extremely thorough testing, is made as a one piece package that the electronic-system designer then uses as a basic building block.

The low price of such amplifiers means that, despite their internal complexity they can be used as freely as transistors were a few years ago.

Before we discuss how to use these amplifiers, let us consider some basic circuit techniques that are used to create the general purpose dc amplifier.

THE DIFFERENTIAL PAIR

In the basic grounded emitter circuit shown in Fig.7 V out will be roughly β times V in. However an unwanted leakage current, termed I_{co}, also flows through the device and resistor and, produces voltage drops across them. Thus the V_{out} value may alter even though V_{in} remains the same. When several stages are cascaded, to provide a gain approaching a million, the temperature dependency of I_{co} is large enough to produce a considerable swing in output voltage. Obviously such a system is unworkable, more a thermometer than a useable dc amplifier!

One remedy is to control the temperature of the element and this was standard practice in early units. Today internal electronic compensation will overcome this problem except in the most stringent cases.

There is, however, a more powerful method of eliminating the temperature effect. It uses two transistors to form what is known as a *differential pair* — as shown in Fig. 8.

When used as a single input dc amplifier, input 1 (or 2) is connected to the bottom rail with the signal to be amplified being fed into the other input. (The emitter resistor provides further temperature compensation). When the working input is also connected to the bottom rail both transistors are connected in an identical manner. Thus the two collector resistances are equal, and if the two transistors have similar leakage currents, the voltages developed at each collector will be closely identical and will 'track' each other with temperature changes.

The output is taken to be that between the two collector voltages, not from one of the collectors to ground. When the two inputs are identical (no difference input signal) the output will be zero. (If not, VR1 is trimmed to make it so.) If one input rises above the other in magnitude, the output between the collectors will swing accordingly, but with the opposite sense and larger amplitude.

In this way the differential pair handles bi-polar (positive or negative going) signal swings, and provides significant temperature compensation.

A further advantage of the differential method is that any noise (such as mains interference or hum) is common to both transistors and, therefore, does not appear at the output. This is called common-mode rejection.

A similar differential circuit can be constructed using a pair of emitter-followers. In this case current gain is obtained instead of voltage gain.

To obtain more gain such a stage can be connected to the two inputs of a following differential pair. Note particularly that the output has no connection with the common lower rail and any attempt to make such a connection prevents correct operation of this compensating method.

In many cases where dc amplification is needed, the input already exists as two leads which cannot be connected to earth — Fig.9 shows the commonly encountered Wheatstone bridge used in measurement. A small change in RX causes the bridge to go out of balance providing either a negative or positive output signal to the differential amplifier. In practice RX might be a temperature-sensitive resistor (thermistor) a strain-sensitive resistance grid (strain/gauge) or a light dependent resistor (LDR), to name just a few uses of the bridge.

Thus it can be seen that the differential pair concept is invaluable in the creation of a workable dc amplifier. In discrete designs the transistors must be carefully matched for best results. In IC designs however this close matching of both characteristics, and the temperature of the devices is almost automatically achieved.

THE COMPLEMENTARY PAIR

The differential pair can handle a bi-polar signal swing but has two major disadvantages. Both output leads must be isolated from ground and the method is wasteful of both power and transistors. For these reasons dc or ac power output stages (where power lost as heat is expensive)often use what is called a *complementary-pair* circuit — shown schematically in Fig.10.

Here the load is connected between the two joined emitters of p-n-p and n-p-n transistors and the 0 volts rail, and the two bases connected together. If the input signal swings positive the upper transistor begins to conduct, increasing the positive voltage applied to the load, and the other transistor is

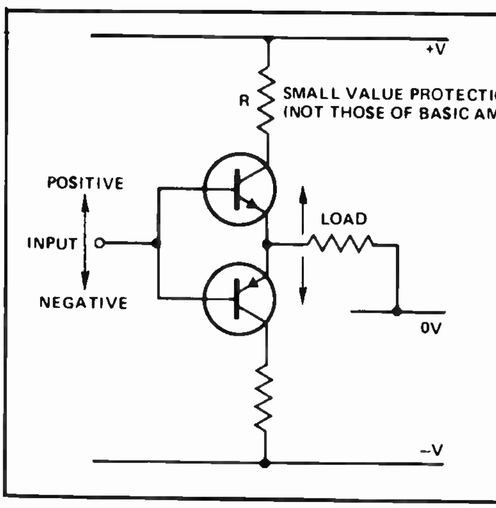


Fig. 10. The complementary pair provides bipolar operation in a very efficient way.

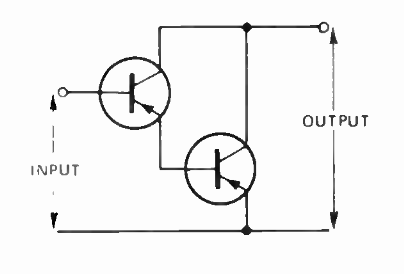


Fig. 11. Two transistors connected as a "Darlington pair" provide a super transistor and perform better than two emitter followers cascaded.

biased into a safe 'off' state. In the reverse direction the opposite applies. As the complementary pair uses emitter followers it is inherently stable. However transients or other effects could possibly switch the off-state transistor to an on-state in which case the transistors would rapidly be destroyed. Addition of small value resistors in each collector helps to reduce this risk.

THE DARLINGTON PAIR

When the need arises for an amplifier with high input impedance the initial stage could be an emitter follower. If still higher input impedance is needed it is better to use the *Darlington-pair* circuit shown in Fig.11 than to cascade emitter followers. Although not immediately obvious, this circuit does consist of two cascaded emitter followers in which the emitter load for the first transistor is the base-emitter junction of the second. With the Darlington pair it is relatively easy to obtain input impedances of greater than 1 megohm. For still higher values the designer would normally use the field-effect transistor (FET). This will be explained later in the series. Darlington pairs are, in effect, a super-transistor for the combined unit still has three terminals, has far greater input impedance and a typical combined gain of 30 000. The pair is available as a single packaged unit.

THE INTEGRATED CIRCUIT LINEAR AMPLIFIER

Having covered the main (but by no means all) circuit concepts used to build high performance dc amplifiers, we are better able to look a little closer at the IC operational amplifier. This circuit block is now used as the general purpose amplifier for both dc and ac analogue signals.

The ideal amplifier should be extremely stable to temperature changes, should not drift over long periods of time, should have relatively high input impedance, very low output impedance, wide tolerance to voltage supply variations, not be damaged by accidental short circuits of the output and be standardised in mounting methods and supply voltage.

Before IC devices were made, numerous manufacturers provided dc amplifiers in even more numerous packages and forms. This did not lead to the drastic price reductions realized by IC manufacturing, additionally their high cost did not guarantee that the units were as good as their makers claimed.

Today there are many makers of integrated circuit components. All offer dc amplifiers that provide a performance so good that we rarely even remember that dc amplifier design is very difficult. We just wire them in and forget them.

Figure 12 shows the basic circuit schematic of the µA 709 operational amplifier, once in common use. This unit required the addition of several components before becoming a standard IC offered by numerous manufacturers. A later design that is commonly available, and inexpensive, is the µA 741 (Fig. 13). This unit needs fewer components to complete the amplifier circuit.

Integrated circuit amplifiers are marketed in single units mounted in round cans and flat packs. One form is also sold with *four* dc amplifiers on a normal dual-in-line flat pack. Figure 14 is a useful reference chart of the

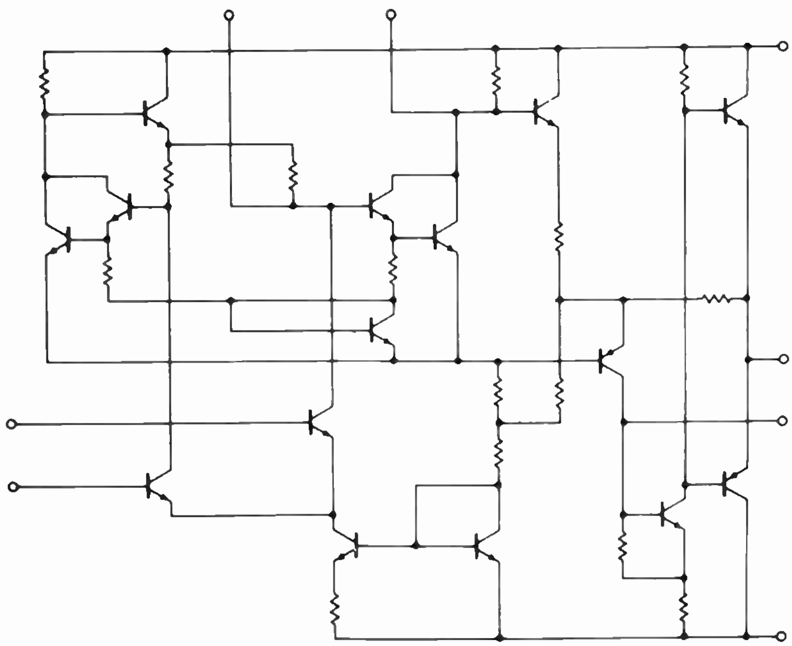


Fig. 12. Schematic diagram of the µA709 high performance, integrated circuit operational amplifier as marketed by Fairchild. Can you pick the Darlington pair, a complementary pair and a differential pair?

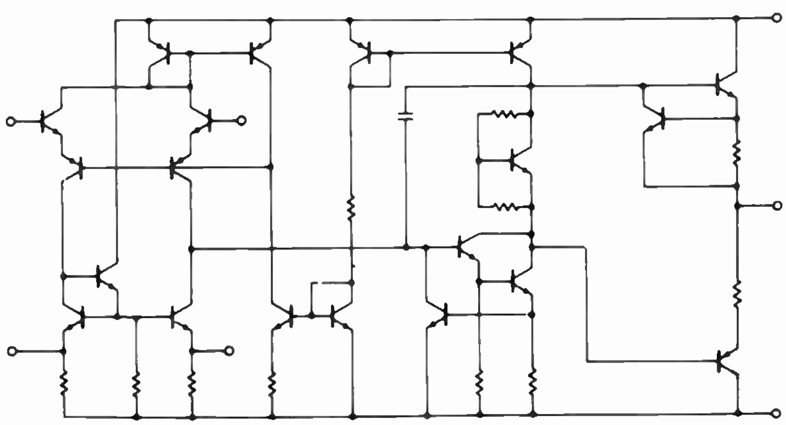


Fig. 13. The µA741 IC amplifier fills a similar role to the 709 but, as this schematic shows has a more complex circuit and includes the frequency compensation capacitor on the chip. This IC is therefore simpler to use and gives superior performance.

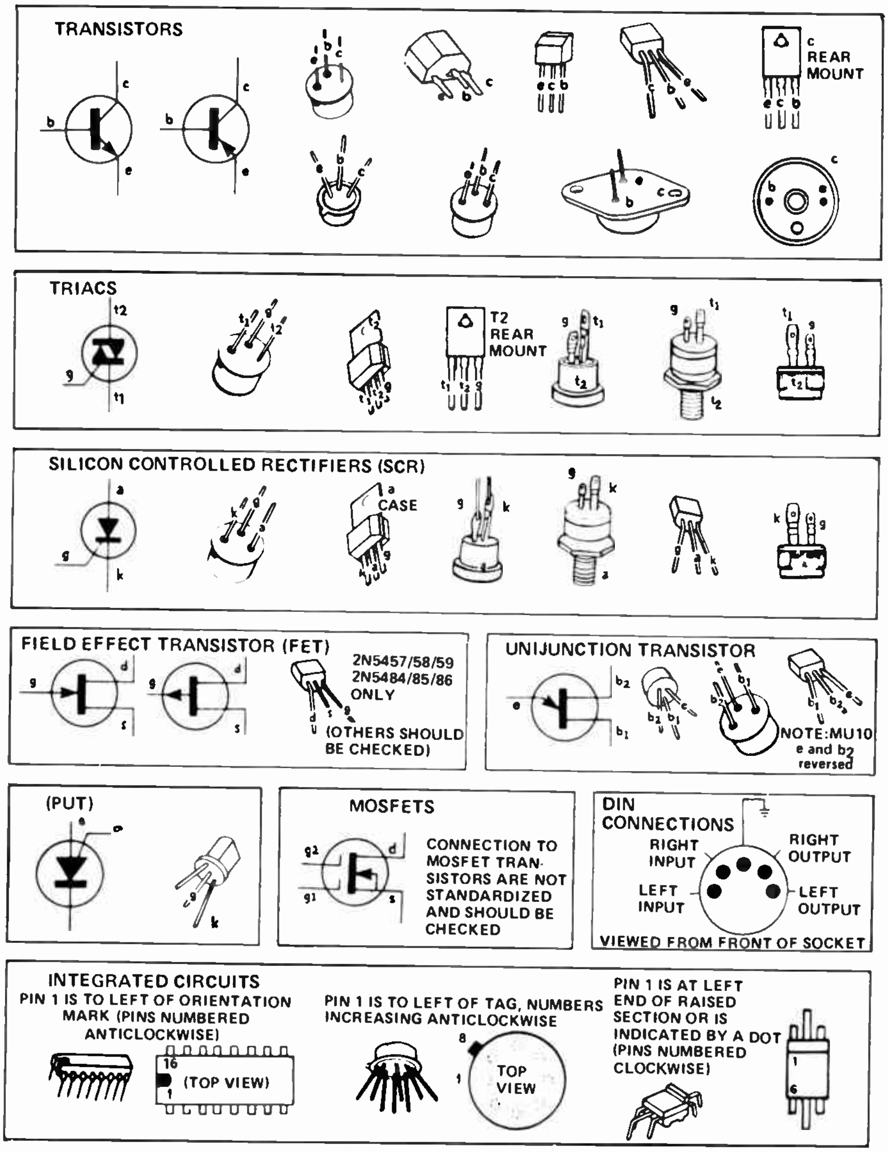


Fig. 14. Symbols, case outlines and connections for the commonly encountered active devices.

main semi-conductor packages (including some yet to be discussed) and their connections.

PRODUCING AN IC

Integrated circuits (ICs) contain many electronic components formed together within a single piece of semiconductor material, commonly silicon. Like transistors, the heat losses are small so the unit does not become overheated unless over-loaded. The actual space needed to perform the electronic function is usually very small. Integration enables drastic size reductions to be obtained. By using stringent manufacturing controls and mass replication of a single device to suit large volume markets, it has been made possible to drastically reduce the per unit price. It is now commonplace to provide 2500 components with a 2 mm x 2 mm area for the price of a dollar or two . . . provided millions are made of the same IC. The first stage in making an IC is development of the required circuit design – perhaps a dc amplifier, a radio receiver, an audio amplifier for a record player or a digital computer sub-system.

The initial design is first made up using discrete components. It is then converted into IC form by preparing masks that when overlaid and used in special ways, selectively form the silicon material. This stage requires extensive experience as the shapes greatly decide the characteristics of the components produced.

Next comes microphotography. Drawings of the patterns are prepared on Mylar as large as a metre square. The Mylar is masked to form the pattern needed. The drawings are photo-reduced some forty times onto a 50 mm x 50 mm glass slide. The slide image is then reduced a further ten times and repeated side by side to fill a 50 mm x 50 mm area. Each IC will have from three to six masks, used in sequence selectively to alter the substrate material. An ingot of extremely pure silicon is grown as a 50 mm diameter cylinder. This is cut into wafers 40 mm thick, each polished to a high degree. 'Epitaxy' is then used to grow controlled semiconductor material on the surface. These wafers are processed to form a resist mask image on the surface. As each mask is added, the wafer is placed in a diffusion furnace which alters the semiconductor wafer where resist does not exist. This is repeated for each mask in the correct order. A final mask adds a metal pattern that provides connections as needed. As many as 50,000 ICs are processed at one time.

Next comes testing. Computer controlled probes test each circuit in turn to see if it performs properly. If not, it is marked with a blob of ink to identify it as useless. A good yield at this stage would be 50 per cent useful units.

The wafer is then cut into its individual chips each containing one IC system. These are assembled into a package, together with leads, fixing points, heat sinks etc. The IC is then finally tested and the package marked before the unit is ready to use. Some ICs may be 'burned in' for many hours to eliminate the few failures that may occur. Once these have been eliminated, the remainder will exhibit good reliability.

(This account is a precis of 'The Inside Story of Integrated Circuits' by the National Semiconductor Corporation).

REFERENCES

Semiconductor manufacturers are pleased to provide data sheets and application notes for their products. The electronics engineer uses these for design data and basic system ideas. Well known companies are:–

National Semiconductor, Fairchild, General Electric, Motorola, Philips, RCA, STC, Plessey, Sinclair, Texas Instruments, Hewlett Packard, Sprague, A.W.A. Data is provided in several forms.

(a) Books and binder catalogues that are expanded with time (not always a free service).

(b) Application notes which often run to 100 page unbound books.

(c) Individual data sheets giving vastly more data and detail than the normal user needs (these may tend to confuse when first encountered).

For those who have a mathematical bent, and an urge to know the intricate whys and wherefores of amplifier design, an inexpensive useful text is "Transistor Manual" by General Electric. It gives the theory of operation of transistors and plenty of generalised practice.

ELECTRONICS – in practice

OFTEN THE NEED arises to drive an output device, such as a relay, with a smaller input signal. The gain needed can easily be obtained with a single dc connected transistor stage which switches the power to the relay. This requires the transistor to operate in what is called switching mode (rather than linear amplification). Although switching circuits have not yet been covered this exercise is useful, helps to build more confidence in the use of transistor devices and illustrates more problems of dc circuitry.

The basic circuit is given in Fig.15. With the switch in the off position the transistor base current is at minimum and V_{ce} rises to nearly 12 V (the transistor 'resistance' changes to be very much greater than the resistance of the coil of the relay). The coil is then de-energised. (The switch, in practice, could be replaced by any other device which provides sufficient voltage change to switch the transistor, eg another transistor or LDR etc).

When the switch is put to the on position, base current flows and is limited by R_b. This is chosen to ensure that the transistor is fully conducting and, therefore, providing a very low resistance compared with the relay coil. The coil is then virtually connected across the supply and the relay operates.

A diode is connected across the relay to absorb the large voltage spikes that are induced across the relay as the current through it collapes when the transistor is turned off. Without it the transistor would be pulsed and probably destroyed.

The design steps are as follows:

(a) To fully energise the relay 12 V/185 Ω of current (65 mA) must flow through the transistor.

(b) From data charts we select BC108 (or BC107, 109) as being capable of withstanding the 12 V supply and the 65 mA needed. The actual limits of the BC107 n-p-n series are shown in Table 1.

(c) The minimum gain (β) of a BC 108 is around 100, so the base current of the on-state must be at least 65/100 = 0.65 mA. To be on the safe side double this, for it is essential that the transistor be properly saturated. A base current of, say, 2 mA is, therefore, needed. Further delving into the data sheet values and curves establishes that this value of I_b is well within safe limits.

(d) Choose R_b to give 2 mA from a 12 V supply which leads to a value of 6 k for which the nearest preferred value is 5.6 k.

These simple steps can be greatly elaborated upon by the trained expert but for general usage they are adequate.

TABLE I

	BC 107	BC 108	BC 109
V_{ce} max. (Volts) (base shunted to emitter)	50	30	30
V_{ce} max. (Volts) (base lead open)	45	20	20
I_{cm} (mA)	200	200	200
P total max (mW) (maximum dissipated power)	300	300	300
hfe (β)	125-500	125-900	240-900
ft (max. useful frequency)	300 MHz	300 MHz	300 MHz

The circuit, therefore, operates the 65 mA relay coil with a control current of only 2 mA – a considerable reduction. For further sensitivity another stage can be added as shown in Fig. 16.

Now the design criteria is that the extra transistor provides 2 mA when switched off – which also means it must pass about 4 mA when on, for the collector resistor decides the flow into the following stage as well as into the transistor. The base current of Q1 is, therefore, now 50μA or less. Note also that the phase is now changed – with the switch up the relay is 'off' instead of 'on' as in Fig.15.

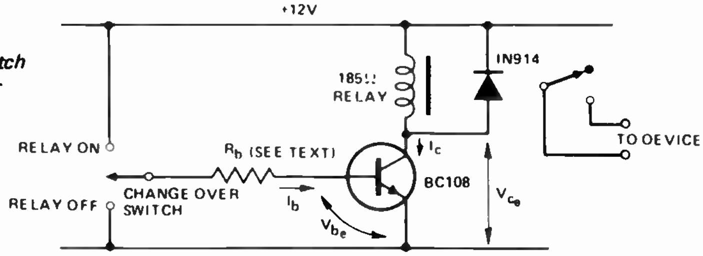
Switching circuits are the easiest dc circuits to design, for thermal effects are not so rampant. This is because the currents needed can be over-driven to ensure that changes in leakage and gain do not bring the transistor out of the 'off' or 'on' state.

It is good practice to clamp the base to earth when the transistor is in the off state. Should the base be left open-circuit, thermal runaway may occur. This is because the leakage current, dependant on temperature, causes V_{be} to rise, collector current to increase and further temperature rise to occur – thermal runaway.

In the two-stage switch V_{ce} of the first transistor never quite goes to zero so a small base current is injected into the next stage. Thus it is by no means certain that the second transistor will turn off.

This can be overcome by using a small resistor or a forward-biased diode (which produces a 600 mV voltage drop) in the emitter of the relay driver transistor. This raises the turn-off voltage at the base of Q2 well above the saturation-voltage at the collector of the preceding transistor ensuring reliable turn-off of Q2, and little chance of spurious turn-on due to transients.

Fig. 15. The use of a transistor as a dc switch provides greater sensitivity for a relay.



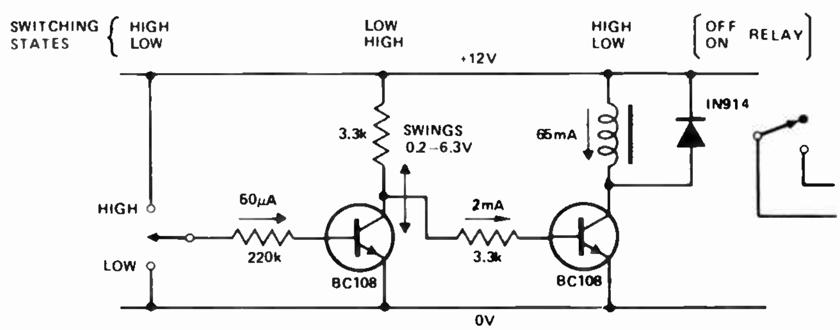


Fig. 16. Two stages of dc amplification further increase the sensitivity, but are more difficult to design (see text).

12 Basic operational amplifiers

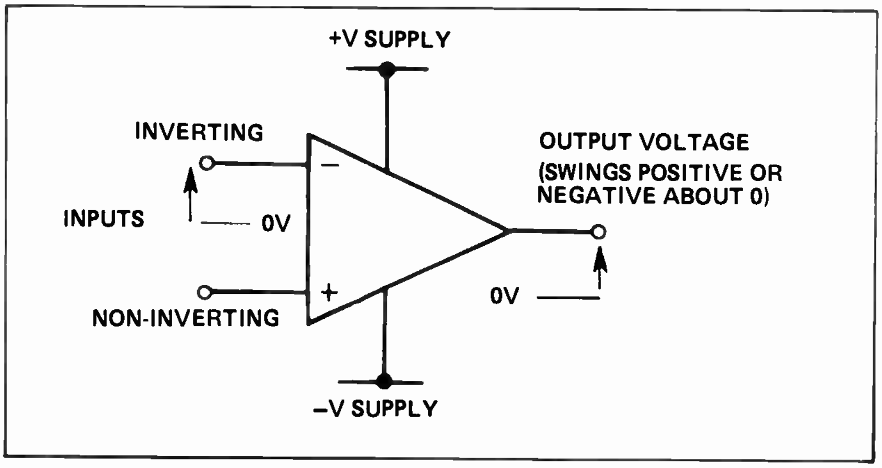


Fig. 1. The basic symbol and connections for an operational amplifier.

IN THE TWO previous sections we explored the role of basic amplifier circuits and investigated how adding extra passive components converts the basic active device into a practical amplifier building block.

These days the cheapest and most straightforward method of amplifying signals is to use one of the many, readily available integrated-circuit operational amplifiers (the op-amp). The methods of using op-amps are universal even though various types may differ in details such as stability and cost etc.

We will see that, provided the basic operational amplifier has a dc gain of 10 000 or more, and draws very little input current, the internal design is of little consequence.

The operation of the complete amplifier system (whether it be based on transistors, ICs or even valves) is determined primarily by the way components are connected around it. That is the basic op-amp unit can be made to perform literally hundreds of different functions by adding appropriate external circuitry. It is this extreme degree of versatility, plus the extraordinarily low price of IC devices that make op-amp techniques so attractive.

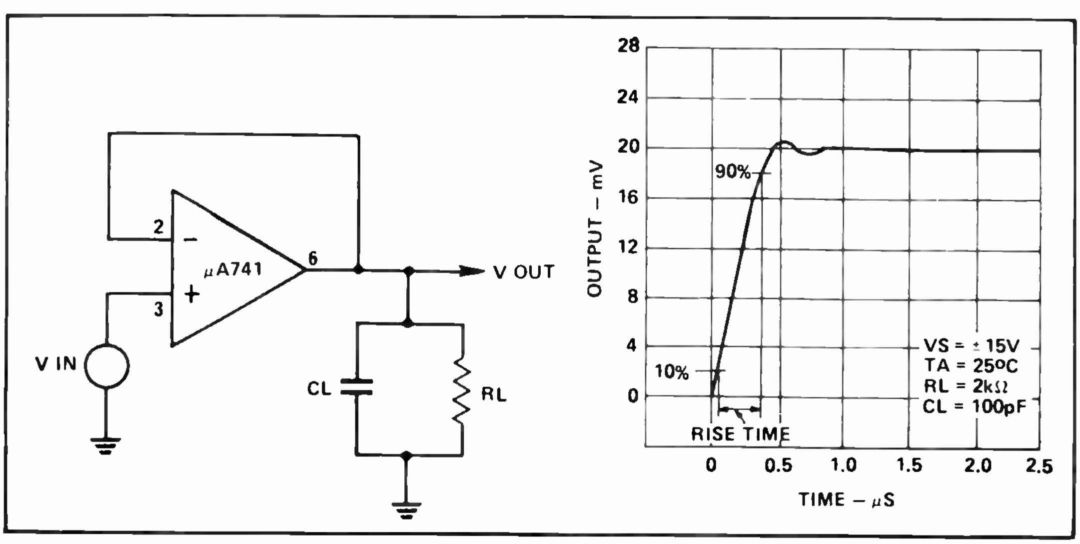
Before we move on to see how such versatility is achieved, we must study the terms used to describe the characteristics and performance of operational amplifiers.

LINEAR VERSUS DIGITAL

We have already described how a transistor stage may be used to amplify, with low distortion, continuously varying voltage or current signals. Circuits that perform this way — they increase the level of complex waveforms without changing them in any other way — are said to be LINEAR systems. By contrast, it is also possible to use the same basic active element so that it is either fully 'on' or fully 'off', depending on whether the input signal is above or below a preset level. The device actually moves through the linear region so quickly that it is no longer a linear device but a switch. There are many kinds of switching circuits and the entire range of such devices and circuits is loosely classified as DIGITAL (digital meaning ON/OFF or step by step operation).

Integrated circuits, therefore, are catalogued by the makers as either linear or digital devices. The op-amp

Fig. 2. Transient response of a 741 type op-amp to a step input change. The test is performed using the circuit shown at left — basically a voltage follower circuit.



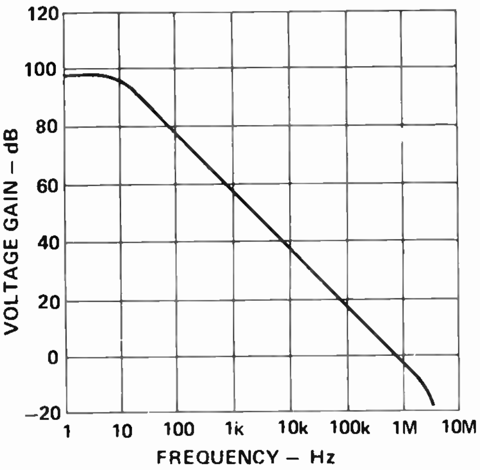


Fig. 3. The open loop (no feedback) frequency response of a 741 op-amp. The constant roll off of 6 dB/octave is built in to ensure stability.

selection form a sub-group of the linear range – others being voltage regulators, oscillators and special purpose units such as timing circuits. It is worth knowing that it is often possible to make a linear circuit perform a digital function but usually the reverse does not hold. In principle at least, the op-amp can be made to fulfill just about all signal processing black-box requirements but to conserve space and power, and to keep costs down, it is usually better to use special-purpose ICs for many applications. Selecting the right component is largely a matter of comparing the cost of various alternatives for the particular job. It may well be better to use a modified op-amp, or even a discrete circuit, to fill a special task rather than await delivery of an exactly right, but harder to procure, special IC.

COMMON LINEAR TERMS

These terms tie in with the general schematic for an op-amp, given in Fig 1. The amplifier itself does not necessarily require a zero/volt connection, it amplifies the *difference* between voltages at the two input terminals.

Large Signal Voltage Gain

This is the ratio of the maximum output voltage swing (under appropriate loading conditions) to the change in input required to drive the output from zero to this voltage. A typical value of gain is 200 000 with an output swing of ±10 V. The input change, therefore, needed to provide full output swing is a mere 50μV. This may seem alarmingly small – a copper to solder terminal connection (forming a thermocouple) will generate signals of the order of 5 to 10μV with small temperature changes! In practice, however, it is rare to use the full gain capability. Gains approaching infinity are necessary, however, so that the performance of the amplifier is entirely dependant on the input and feedback networks – not on the device itself.

Input Offset Voltage

A differential voltage of only 50 microvolts is necessary to provide full output swing. However due to manufacturing tolerances the matching of the input transistors may not be exact and a small offset voltage may be required at the input to balance the amplifier under no-signal conditions.

This voltage is normally less than 6 mV, but could be troublesome in a low level dc amplifier. Therefore provision is made on most op amps for connecting a potentiometer to null out this voltage, thus making the output zero under no-signal conditions.

Input Bias Current

All operational amplifiers (and also transistor amplifiers) require a small steady-state input current called the input bias current.

Input Offset Current

The difference between the two bias currents in a differential op amp is known as the input-offset current, and is specified at a particular temperature. With equal resistances in series with the two input terminals, it is only this difference in bias currents which produces an offset error. When the input source impedances are high the effect of input-offset voltage is far less than that of bias and off-set currents.

A typical value of input offset current would be 50 nanoamps (50 x 10^{-9} amps), but may be much lower in more expensive op-amps.

Input Resistance

With one of the two inputs grounded, the input resistance is that seen looking into the other input. A general purpose op-amp exhibits around 0.5 megohm input resistance. Some better quality amps go higher and the ideal, as we shall see later, is an infinite input resistance.

Feedback, when applied, modifies this value considerably, may reduce it to zero (inverting amp) such that the input impedance is the value of the resistance in series with the input, or may increase it to several megohms (non-inverting amplifier).

Output Voltage Swing

This tells us how far the output voltage can change in both positive and negative directions. It will always be a little less than the supply voltage. For a supply of ±15 V a typical swing (without distortion occuring) would

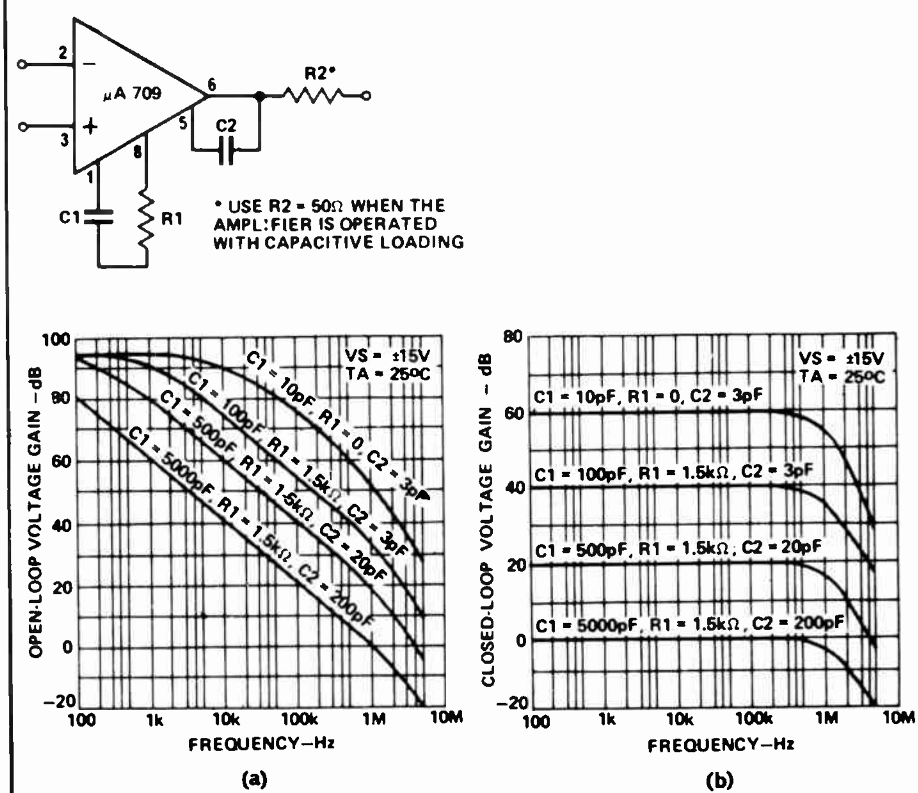


Fig. 4a. Some op-amps allow frequency response to be tailored by choosing external components. This set of curves is for the 709 op-amp.

Fig. 4b. The choice of components also determines the frequency response in the closed loop mode.

ELECTRONICS -it's easy!

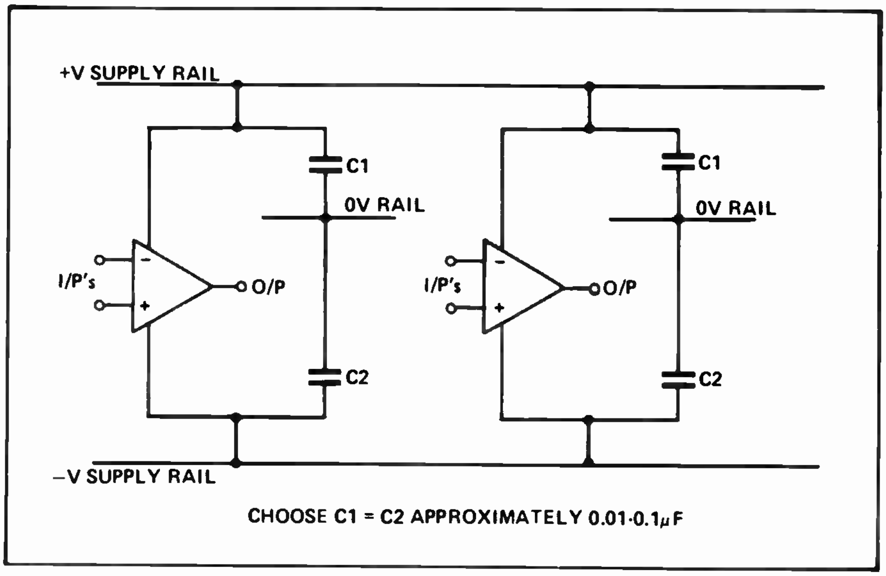


Fig. 5. Each op-amp should have decoupling capacitors across power supply lines to ensure that transients are not coupled from one amplifier to the other through the power rails.

be ±12 to ±14 V. Op-amps work satisfactorily over a wide range of supply voltages — voltages less than the maximum specified may be used.

Input Voltage Swing

This value must not be exceeded if the amplifier is not to be damaged. In most modern op-amps, such as the LM301, 741, etc, the inputs can be taken to the supply rail together and in some amplifiers may be taken to opposite supply rails (that is, 30 volts between them) without damage. However some older types such as the 709 may only withstand a common mode voltage of ±8 volts and a maximum differential voltage of ±5 volts.

Input common mode rejection ratio

This is the ratio of the input voltage range, to the maximum change in input offset voltage over this range. It is quoted in decibels being typically 80-90 dB i.e. if the inputs are moved by 10 V the offset voltage could change by 1 mV.

Output resistance

A typical value of output resistance measured into the output terminal with the output near zero volts) is around 100 ohms. This measurement is made with a small signal level and at approximately 400 Hz to avoid dc drift problems. This however is the open loop (no feedback) output impedance and is substantially reduced when feedback is applied. The maximum load which can be connected to an operational amplifier is not determined by the output impedance but by the current that the op-amp can supply (typically 10 mA).

Output power

The normal op-amp is usually designed for low power output only. If power is required a power stage of discrete transistors (or special power ICs) is added after the op-amp.

Supply Voltage Rejection Ratio

This relates the change in input offset voltage to the corresponding change in supply voltage producing it. It expresses how well the circuit ignores voltage supply variations due to mains fluctuations etc.

Typical values lie around 100 μV change per volt of supply change. In critical dc amplifiers the power supply, therefore, must be stabilised (that is, the voltage must not change with respect to zero). For example, if the design can tolerate only 10μV change in input offset voltage the supply must be stable to within 100 mV of its magnitude.

However for general applications rejection of supply voltage changes is sufficiently good that close regulation is not required.

Power consumption

Even when an op-amp is not providing an output, that is when the output voltage is at zero, the circuit still consumes power. This value is usually quoted for zero output conditions (the greatest internal power loss across them) and is in the region of 100 mW.

Peak output current

The output current must not exceed the stated value or permanent internal damage may occur. Many designs, however, now incorporate protective circuitry that enables the output to be short-circuited without damage.

Dynamic response of operational amplifiers

A dc amplifier, by definition, apparently has no need to handle fast changing signals. In fact, it is quite normal for it to need a good frequency response extending to tens of megahertz. Examples of this are the wide band ac amplifier having a response down to dc and a dc amplifier which will faithfully reproduce a sudden voltage change in a control system. In such cases the system designer needs to know more than just the dc large-signal gain of the op-amp — he needs to know the gain at all frequencies.

The response of the op-amp to a step input voltage is called its transient

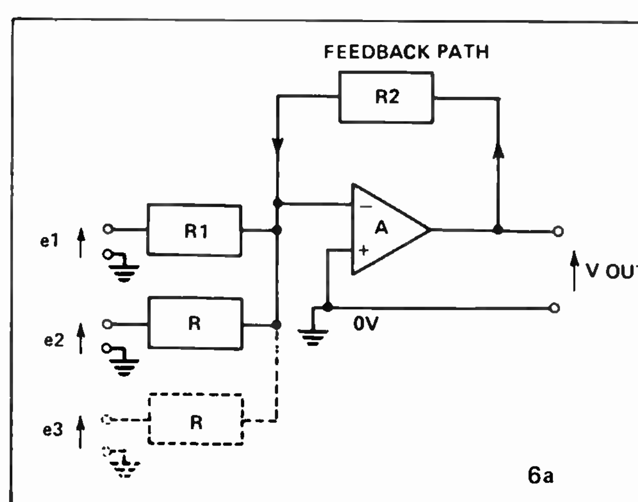


Fig. 6a. General circuit configuration for an inverting amplifier.

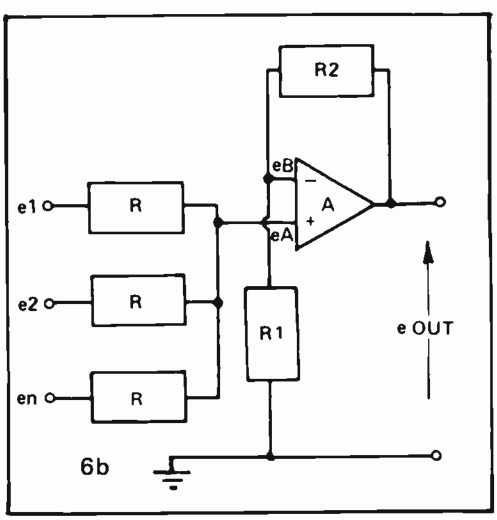
Fig. 6b. General circuit configuration for a non-inverting amplifier.

response. That is, this parameter defines how the output of the op-amp will follow an input change with time under closed-loop (with feedback) conditions. The usual way that this is quoted is by an amplitude/time graph — as given in Fig. 2 — which shows how the output changes when a 'perfect' step-rise in input voltage is applied. Note also that the diagram indicates the load resistance value, the capacitance of the load, the supply voltage and the device temperature: each of these will alter the shape.

A second dynamic characteristic is the frequency response. Figure 3 shows a typical response curve (note that such curves vary greatly with different amplifiers) for an op-amp without feedback (called open-loop operation). In general such curves always have the same basic shape; flat to begin with and then falling off at the same rate of 6 dB per octave (20 dB decade). There is a good reason for such a characteristic — it ensures stability in closed-loop working.

If the slope were increased the amplifier could introduce excessive phase shift. Thus the feedback could become positive rather than negative and the amplifier may oscillate. Some op-amps have facilities for the circuit designer to provide external compensation to the IC. This usually consists of an RC network or a single capacitor, the values being selected to suit the application. Figure 4 shows how these values alter the frequency response of the popular 709 type of op-amp.

A third important dynamic term is Slewing Rate. A typical value is stated at 0.5 V/μs, meaning that the output can change no faster than half a volt in each microsecond. The value is quoted for a feedback connection of unity gain — at other values of gain the rate will be different. Thus although the amplifier may well handle a small signal at a given frequency a large



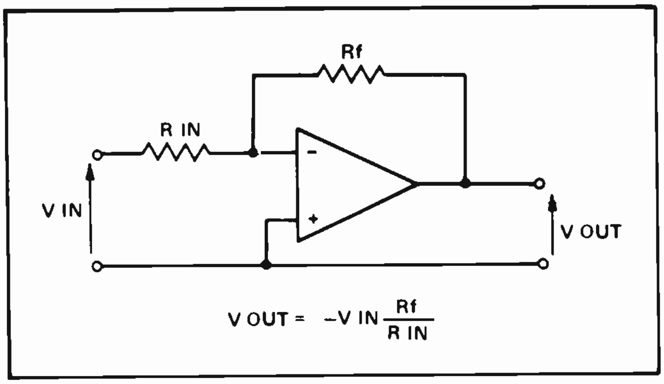


Fig. 7. In this configuration the stage gain is determined by the ratio of Rf to Rin — not by the device gain.

signal at the same frequency may well be distorted because of the slew-rate limitation.

The above terms are those commonly encountered. Other more obvious parameters are given — temperature range of operation, and lead temperature when soldering. The manufacturers of ICs also give a variety of curves for various parameters — voltage gain versus supply voltage, power dissipation versus temperature and many more. These are all helpful from time to time but, in general, the casual user will not need to explore them in depth.

THE MAGIC OF FEEDBACK

The basic op-amp will only accept very small input signals because of its enormous gain. At first sight this seems to be a peculiar way to go about things for surely the optimum would be to design the internal circuitry to give the gain needed and no more. We will soon see that there is a better, and more versatile way of obtaining any required gain (by adding a simple network to the amplifier). For this concept to work correctly, the amplifier must have a very high gain.

The two basic amplifier circuits are illustrated in Fig 6. The circuit of Fig 6a is an inverting amplifier and that of Fig 6b a non-inverting amplifier.

For the purposes of our discussion we must assume that the amplifier has an infinite input resistance and infinite gain. The input signal (in Fig 6a) is applied via R1 and the output is fed back to the input via R2.

Thus, as the ideal amplifier draws no current, the current in R1 is

$$I_{R1} = \frac{e1}{R1}$$

and the current in $I_{R2} = \frac{eo}{R2}$

A theorem not yet covered (called Kirchoff's Laws) states that the sum of the currents at any point in a circuit must be zero. Therefore $I_{R2} = I_{R1}$

By Ohm's Law: $\frac{e1}{R1} = \frac{eo}{R2}$

Now the gain $A = \frac{eo}{e1}$

$$\therefore \quad A = \frac{R2}{R1}$$

How convenient! The gain of the amplifier may be set by adjusting the ratio of R2 to R1. In a practical amplifier there is some error because the input impedance is not zero and the gain is not infinite. But providing the amplifier open-loop gain is in excess of 10 000 the error may be disregarded.

For the non-inverting configuration it may be shown that the gain is

$$A = 1 + \frac{R2}{R1}$$

Hence it may be seen that any reasonable gain may be programmed by simply selecting two resistors, and that drastic changes in device open-loop gain will have little effect on the closed-loop gain.

The open-loop gain should be at least 10 dB higher than the closed-loop gain at all working frequencies to maintain frequency response of the amplifier within 3 dB. (See graph of open-loop gain versus frequency.)

The effects of using feedback are as follows:—

Inverting Amplifier

Output impedance is reduced by the loop gain. That is, if the amplifier has a gain of 10 and the output resistance is 150 ohms, the closed loop output impedance falls to 15 ohms.

As the amplifier always tries to keep its input terminal at zero the input impedance is equal to the value of R1.

Distortion is reduced by feedback by I/A.

Common mode rejection is improved and the stability is improved.

Non Inverting Amplifier

Output impedance is reduced by the loop gain.

Input impedance is increased by the loop gain, (but is limited by common mode impedance and resistances connected between non-inverting input and ground.

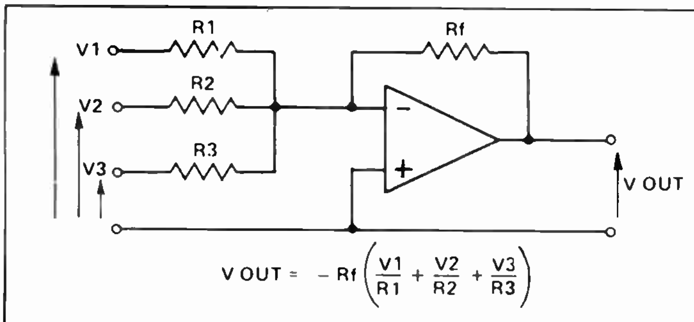


Fig. 8. If extra input resistors are added the amplifier sums the input voltages and amplifies them by the respective resistor ratios.

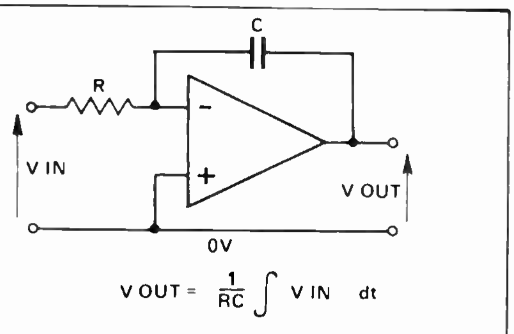


Fig. 9. To make an op-amp integrator we simply replace the feedback resistor with a capacitor. The time-constant of the stage is then determined by R and C.

Distortion is again reduced by I/A.

Common mode rejection is not good with this configuration, the amplifier is less stable than the inverting mode and the gain can never be less than unity.

INVERTING ADDER

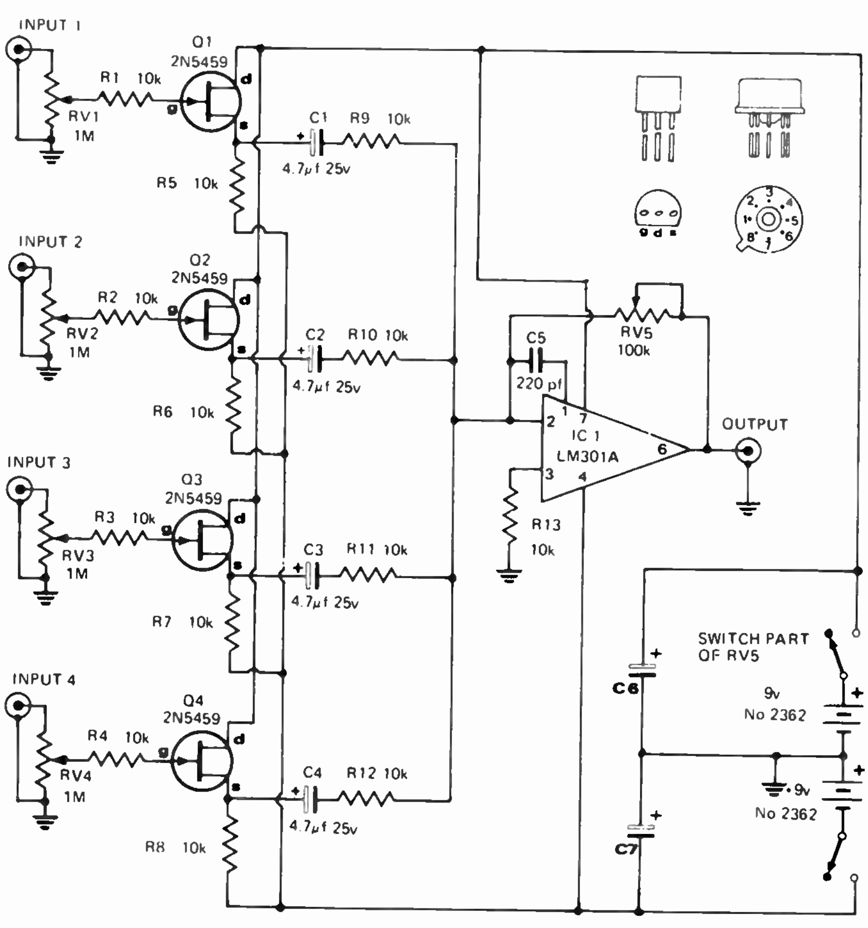
The fact that currents are effectively summed at the inverting input terminal may be exploited as shown in Fig. 6a. The input signal voltages are summed by using a separate input resistor for each signal. The input signals are effectively isolated from each other as the summing point acts like a virtual earth.

Thus the output will be:–

$$e0 = -R_f\left[\frac{e1}{R1} + \frac{e2}{R2} + \dots \frac{en}{Rn}\right]$$

That is the output is the sum of the input voltages, each being amplified by the ratio of the feedback resistor to the individual input resistor. This simple circuit finds great use in audio mixers etc.

Fig. 10. This audio mixer accepts four input signals and combines them. Variable gain is provided by RV5. The circuit is basically a summing op-amp.



USING OP-AMPS TO DO ARITHMETIC

The simple summing circuit described may be used to do arithmetic. For example assume we needed to continuously solve the problem X = A+2B. We could simply apply a voltage proportional to A through an input resistor of 10 k to an amplifier having a feedback resistor of 10 k. The second input voltage proportional to B is applied through a resistor of 5 k. Thus the amplifier has unity gain to A and multiplies B by 2. These voltages are summed to provide our required function of 'X'. This is the basis of analogue computers. Analogue computers consist mainly of a group of operational amplifiers configured to solve a particular mathematical expression.

Other mathematical functions are also easily achieved. Subtraction is done by putting the input to be subtracted through a unity gain inverter before summing.

INTEGRATION WITH AN OP-AMP

In many electronic instrumentation circuits there is need to integrate a signal with respect to time. For example there may be a call for a time delay having a precise timing interval. The integrator circuit uses a feedback impedance that is a capacitor – not a resistor. (The theory still holds when the impedances are of any type – R, L or C or even mixtures; it is the mathematical manipulation that becomes difficult.) It can be shown that the circuit given in Fig 9 has the following output to input relationship.

$$V_{out} = \frac{1}{RC} \int V_{in}\,dt$$

($\int$ integration symbol)

The formula tells us that the output voltage is the true undistorted integral of the input signal with respect to time and that the values of R and C decide the time-constant of the stage.

The integrating op-amp circuit finds use as a ramp generator, as a basis of repetitive signal generation and is invaluable in solving mathematical equations in the analogue computer.

Although only one input resistance is used here it is quite feasible to use a number of input branches to combine summing with integration.

The above circuits, based on op-amps with feedback, provide just a few of the many arithmetical operations that can be obtained.

Combinations of different op-amp circuits can perform multiplication, division, squaring and square root functions, and solve simultaneous and differential equations. In normal electronic practice, such op-amp circuits are used to provide accurate signal processing at low cost. Indeed, in many instances it is preferable to use analogue computational circuits in preference to digital methods — each case must be decided upon its merits in terms of cost, size, speed and versatility.

With these principles in mind, it is relatively easy to untangle what is happening in a seemingly complicated circuit like that given in Fig. 10.

This is a four-input mixer (ETI project 401). In this case the four inputs are amplified by field-effect transistors then ac coupled by capacitors and 10 k resistors to the inverting input of the op-amp which is connected as an adder. The second input is earthed via a resistor of similar value to the input resistors as this aids stability. The 100 k feedback potentiometer, RV5, adjusts the Rf value thereby changing the gain and thus providing overall volume control. A 220 pF capacitor is needed to adjust the performance of the op-amp as dictated by the maker — it is not part of the summing circuit proper.

Figure 11 illustrates one way of using an op-amp integrator. The gain of the stage is 1/RC = 1/10k 100μF = 1. Note again that the positive input is earthed via a 10 k resistor. When an input is applied, the output rises linearly with time and could keep going until the maximum available voltage is reached. Before this occurs the switch is closed, discharging the 100 μF via the 100 ohm resistor. The result is a ramp signal that restarts its climb when the switch is operated and released. The switch may of course be an electronic one and such circuits are used to obtain sawtooth waveforms which are useful for a wide variety of applications.

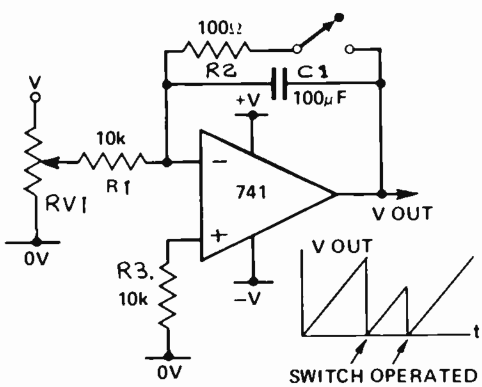
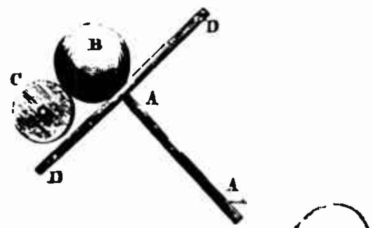


Fig. 11. The integrator may be used to generate sawtooth waveforms.



SIDE ELEVATION.

D, the Disk.
A, the Axle of the Disk.
C, the Cylinder.
E E, the Axle or the Journals of the Cylinder.
B, the Ball.

Ball and disk method of integrating. It was invented by James Thomson in the 1860s. This unit is from 1876.

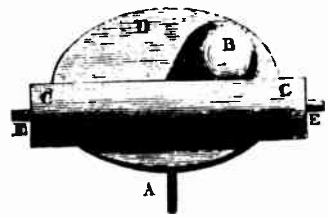
THE DEVELOPMENT OF ANALOGUE COMPUTING

The ancient Greeks had a good understanding of simple mechanical devices such as gears, levers and pulleys and are credited with the development of the first analogue calculators. A calendrical computer, known as the Antikythera mechanism, has been found that dates from around 80 BC. The Chinese had a south-seeking chariot in the first century AD; its manlike figure always pointed in an adjustable preset direction regardless of the directional movements of the chariot. (It was all done with gears forming a differential mechanism).

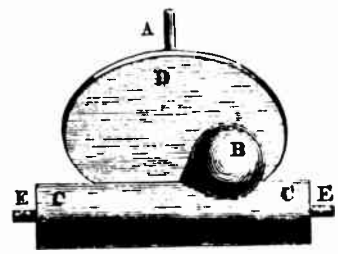
Mechanical design skills were given great emphasis in Europe from the 14th century onward for there was strong movement in fine mechanical skills, particularly in clock making. This influenced the production of many similar devices including several that could perform simple analogue and digital calculations. Pascal, and almost currently Leibnitz, both built mechanical digital adding machines, in the early 17th century.

As mathematics progressed, the need grew to solve the arising equations. Late in the 18th century, scientists began to look for ways to obtain solutions using mechanical devices. They set up mathematical equations and the solutions were produced via the movement of the respective parts of the mechanism. The main areas of need arose in astronomy and in ballistics calculations.

At the same time as such requirements arose, other important (later to be related) developments were taking place for different reasons. One was the concept of closed-loop feed-back control; a sensor monitored the variable to be controlled causing the power source to alter accordingly to



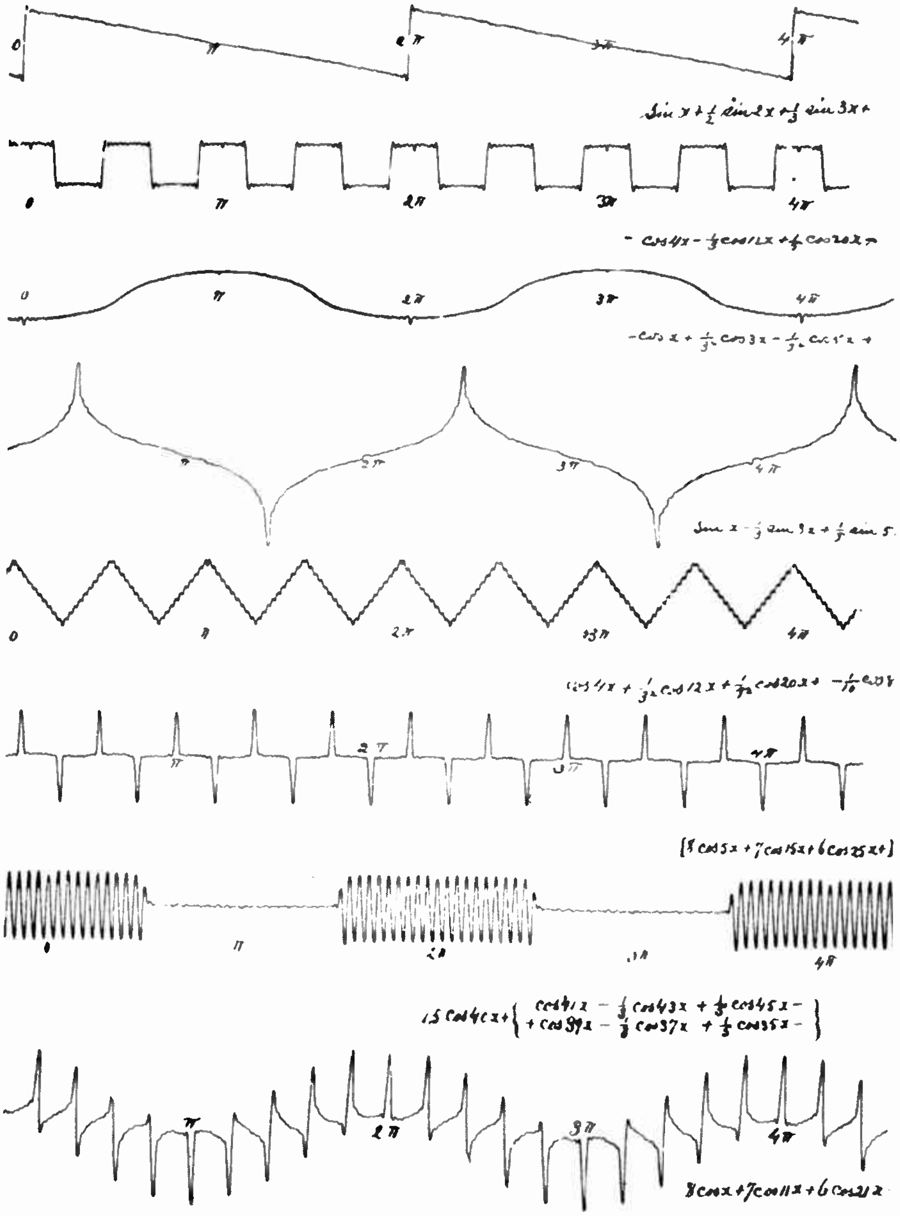
FRONT ELEVATION.



PLAN.

bring the variable back to the preset desired value. The first unit in which the sense and power source functions are clearly separate seems to have been an incubator made in the early 1600's. Watts' ball governor, fitted to a steam-engine in 1787, showed the industrial age what could be done with feedback controls. Feedback control systems were later called servos.

Many mechanical computing machines were developed in the 19th century. James Thomson, brother of the famous Lord Kelvin, devised a mechanical ball-and-disk integrator mechanism in the 1860's. Kelvin later incorporated this into a mechanical tide predictor (in 1873). His machine could draw on a chart the amplitudes of water tides that would be expected from knowledge of the tidal potential of the moon on the seas. In reality, the machine resolved the various harmonics of a signal . . . it was a mechanical Fourier analyser in reverse. Kelvin also



The eighty element harmonic analyser designed by Michelson and Stratton produced these curves by addition of harmonics. That was in 1898.

designed a harmonic analyser along the same lines.

Many similar mechanical calculating devices were made subsequently. In 1898, Michelson and Stratton designed an analyser that could assemble up to eighty harmonics. In so doing, they introduced the technique into the USA. It was there that much of analogue computing was developed over the next half century. A few of the curves created by addition of harmonics as a Fourier series are shown. Michelson, together with Morley, had previously developed his now well known experiment on the velocity of light. Stratton became the key man in the creation of the National Bureau of Standards of the USA, becoming its first director in 1901.

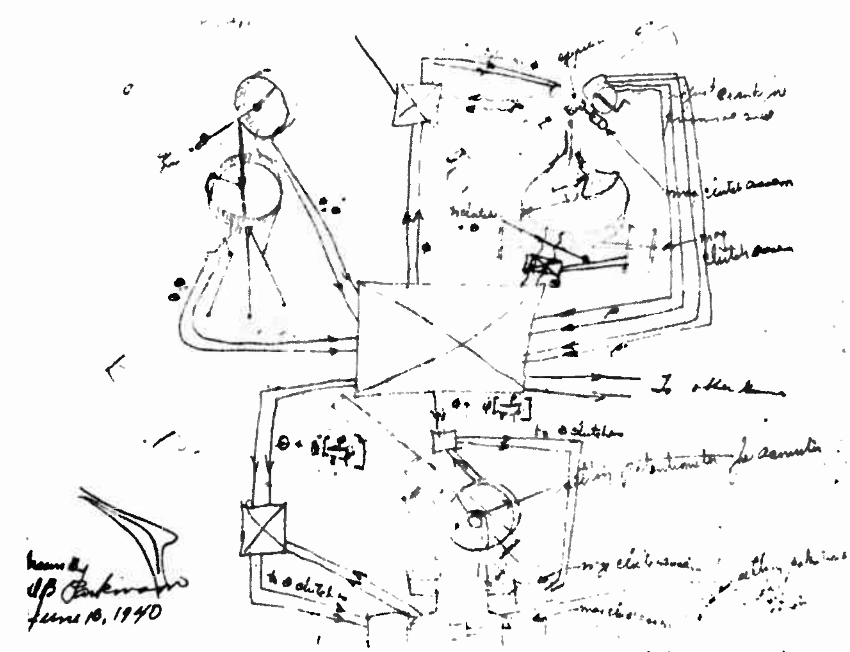
By the start of this century, scientists and engineers had a good grasp of the concept that physical systems that can have their performance described in mathematical terms can be simulated by using another energy form of the same equations. This became known as the method of analogies. But in 1900, electronic devices for amplification and detection were still unknown. Equations had to be modelled with (in the main) mechanical devices. The speed and precision of these depended upon the care exercised in manufacture. They were also slow to develop, for mechanical systems are not as easily modified as are the electronic systems that we use today. They also cost considerably more to make than would be expected today. These factors collectively set the scene for new methods to emerge. And emerge they did as the following four decades revealed. A powerful analogue device, the gyroscope, was developed for guiding ships and torpedos (in 1900-1905). Sperry and Hannibal Ford are now recognised as pioneers of gyro-guidance systems. They even made a system for use in aircraft by 1909. In this way they harnessed the torque produced by a gyroscope, disoriented from its original direction, to move the rudder of a ship or airplane correspondingly. This was a clear start to analogue computation. Sperry and Ford added a windspeed indicator, feeding its signals into the gyro signals thus incorporating analogue computation into a servo system.

It was not long before other such devices were designed. The USA's Army and Navy used a bombsight computer (in 1918) that combined electrical and mechanical components to perform the calculations needed to tell the bomb aimer when to release his charges. It is now clear that electromechanical computing devices were considered to be very important in development programmes of that time. One man's work that was fundamentally important was that of Leonardo Torres. Between 1893 and 1920, he built many equation solvers. Torres, in 1902, showed that any desired mathematical function could be represented mechanically. He built a machine that could find the roots of polynomials up to the ninth degree.

Around 1900 the US Coast and Geodetic Survey began work (under the direction of Fisher and Harris) that would replace their existing tide predictor. It was dubbed the 'Great Brass Brain' when it was put into service in 1914. It could compute up to 37 tidal components. It was a masterpiece of craft, skill and ingenuity. Around that time, the Germans also developed similar machines, because enemy ships that could predict tide heights could make advantageous use of shallow waters. The Germans needed similar skills so that they could follow with their submarines.

In 1922, Minorsky published a paper 'Directional Stability of Automatically Steered Bodies' in which he analysed control problems. He also contributed to the establishment of the control discipline that came into positive existence in the 1950's.

In 1913, Torres had speculated about such a basic control discipline; Harold Black was able to identify its most basic entity; laws and behaviour of negative feedback. Black spent many years researching methods of improving telephone transmission, eventually developing the concept of negative feedback to reduce distortion



D. B. Parkinson's 1940 drawing for the M-9 gunnery control system. It incorporated vacuum tube amplifiers set up as the now familiar operational amplifier.

in electronic amplifiers. By 1936 his work was in practical use on trunk lines. Today the basic concept of negative feedback appears to be almost trivial but it often took enormous effort to realise and simplify it into easily taught principles.

Bush was another pioneer of servo theory. He published a classical paper 'Theory of Servo-mechanisms' in 1934. At that time electronics had become an established medium for doing many tasks previously done mechanically. The trend toward electronic methods was becoming increasingly established by the 1930s.

Many people made significant contributions to the development of analogue computing, too many to name individually. Mechanical machines, although limited in today's terms, were able to provide considerable support to people with calculation problems. One example is striking. In the 1930s, Leontief developed a model of the economy in which 42 parts, called sectors, were identified. He worked out that it would require 30,000,000 multiplications to get a solution. With the best available calculators of the day that would have taken more than ten years of work to reach an answer. He reduced his model to only ten sectors, thereby reducing the multiplications to a work load of only two years. Even that was prohibitive so he instead used a simultaneous equation solver made by Wilbur (at the M.I.T.). It used tapes and tilting plates to solve nine simultaneous equations in a matter of hours instead of years.

The Wilbur machine, (Leontief remarked) could have its coefficients altered a little by simply sitting on the frame of the machine. If they did not change much it was a good indication that the solution was relatively stable! That remark highlights one of the analogue computer's advantages. It provides the user with a far better feel for the process being simulated. It also enables the operator to easily alter the parameters and reveals the general pattern of the process as it proceeds with the solution.

Thermionic valves soon showed that their technology, although clumsy by today's standards, was far easier to use than mechanical alternatives. In the 1930s, designers began in earnest to replace mechanical methods with electronic ones. The stable amplifier arose out of developments by Leeds and Northrup, under Williams. They sought a stable, zero-drift free, dc amplifier for use in an industrial recorder. At that time recorders were still using a simple galvanometer movement in which mechanical feelers followed the needle causing the pen to follow also. With purely mechanical and electrical methods such recorders had only milli-volt sensitivity. Another large area of analogue computer use in those times was in the control and management of power-grid systems. Power authorities used network analysers to tune up their operations. These were the size of large rooms. Hundreds were in existence by the 1940s. In 1936 Beuken's thesis 'Heat Loss in Periodically Activated Electric Ovens', showed that heat flow could be modelled by an electrical network. He, too, added to the concept that a general-purpose analogue computer, using entirely electronic techniques, was possible. In that period, people, such as Bush, were actively looking for a generalised machine for analogue problem solving. Its use would be far reaching and wide ranging. In 1934, for instance, Porter produced his Master's thesis 'The Construction of a Model Mechanical Device for the Solution of Differential Equations with Application to the Determination of Atomic Wave Functions'.

War efforts again stimulated such ideas. Flight simulators were developed for training pilots of multi-engined planes. It was in that work that it was realised that electro-mechanical simulation was not fast enough. Eventually digital techniques overtook analogue in that application. In another area of war effort, that of gunfire control, electronic analogue methods were used in a system known as M-9. Improbably, the system originated as the result of a dream the designer, David Parkinson, had after the Dunkirk saga. He 'saw' guns firing under automatic control based on information about speed and position of aircraft. The dream turned into a reality which worked well against non-evasive targets. In fact M-9 systems shot down 4672 of the 4883 V-1's aimed towards Britain during World War Two. That was significant enough at the time, but more significant still was the concept of the 'operational amplifier' used in the system. It showed how mathematics can be performed in a general way using a single kind of amplifier connected in just a few basic and simple ways to suit a vast set of needs. The modern form of analogue computer had arrived.

Today the analogue machine is finding a rebirth now that it is recognised that digital methods are not always the best. Most engineering organizations now have hybrid analogue-digital machines at their disposal. In this way the best of both methods can be brought to bear on problems in which mathematical methods show the way to proceed yet which cannot be solved by classical, simple methods.

REFERENCES

If you have not already obtained a good set of applications notes on the many uses of op-amps, now is the time to act for they will prove invaluable when building your own systems.

It matters little whose notes are obtained. The basic op-amp circuits have remained reasonably stable for several years now. Perhaps the only point to consider is to ensure that the op-amp specified is both available and an inexpensive choice.

ELECTRONICS – in practice

INTEGRATED circuits (whether linear or digital) are produced in a variety of packages (illustrated in Fig.12). The methods of mounting are different for each form and the beginner would be wise to restrict himself to dual-in-line, TO5 or TO99 versions. Care should also be taken not to overheat the devices when soldering.

Mounting sockets are available for around 40 cents each and these are invaluable for the experimenter – allowing an IC to be used many times without damage. However even at 40 cents the sockets are dear when compared to current IC prices of around one dollar. So a socket may not always be a justifiable expense.

An alternative is to mount the IC on a small piece of matrix board and wire in leads for input, output and power connections. This provides robust connections by which the amplifier may be wired into experimental circuits again and again without damage.

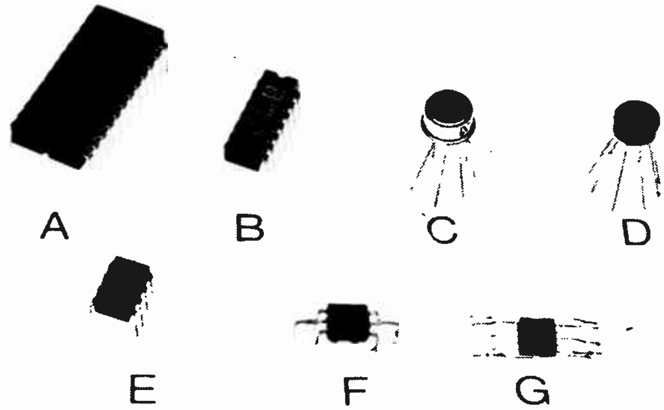


Fig. 12. Typical integrated circuit packages. (a) 24 pin dual-in-line. (b) 14 pin dual-in-line. (c) TO5 metal can. (d) TO 105 plastic. (e) 8-pin dual-in-line. (f) case 643A (used for low price consumer devices). (g) TO86 flat pack.

A MULTI-PURPOSE AMPLIFIER

An operational amplifier is an ideal amplifier for experimental use. It enables a wide range of gains (or attenuation in the inverting mode) to be obtained and is hence very useful in the experimenter's workshop.

It is surprising how often a little extra gain or attenuation is needed when experimenting with electronic circuitry. Often the need is temporary, in order to establish what gain is required in a particular circuit. Once this has been established the unit is replaced with the simplest stage that will do the job – e.g. a single transistor.

This project provides such a general purpose amplifier having a gain adjustable from 0.1 to 100, and with a choice of inverting or non-inverting operation.

The circuit given in Fig 13 is based on a Fairchild application note and uses the currently cheapest op amp, the 709 series. As power drain is low two dry cells may be used thus avoiding the expense of a mains power supply.

To obtain non-inverting or inverting operation, a changeover switch has been included to ground the appropriate input to the op amp. Components C1, R4 and C2 have been chosen to obtain the maximum bandwidth of 0.5 MHz.

The output impedance of the unit is less than an ohm but the peak output current should not exceed 20mA. The power dissipation of the amplifier must not exceed 250 mW so even 20mA may be too much under some circumstances.

The input impedance in the inverting mode is the value of R1. Hence if a higher input impedance is required then R1 should be increased. Note that R2, R3 and RV1 must also be increased proportionately to maintain the gain ratios.

Do not use values in excess of 10 megohm as stray capacity and leakage resistances will then affect stability and accuracy.

CONSTRUCTION

The form of construction is largely a matter of individual preference. We suggest that a small diecast box be used. By mounting the switches and potentiometers through the bottom of the box (rather than onto the lid) the lid may be removed for access.

The gain control should be marked by experimentally verifying gains at various positions of the control. A multimeter may be used to compare input and output, to determine gain, but use a low frequency (e.g. 400 Hz) from a low impedance source, so that the accuracy of the multimeter itself does not affect results.

USING THE AMPLIFIER

Basically the unit is a single-input, variable-gain dc amplifier. We have seen in the theory section however that it can be used for other purposes.

To use the unit for amplifying ac, use a capacitor in series with the input to isolate any dc component of the previous circuit. Make sure that the reactance of the capacitor is less than the value of R1 at the lowest frequency of interest.

To mix signals, simply add additional 10k resistors from each input to the summing point (pin 2).

An integrator may be constructed by replacing R2 and RV1 with a capacitor.

Thus, as well as being a useful tool, the amplifier may be used to increase your understanding of op amp techniques.

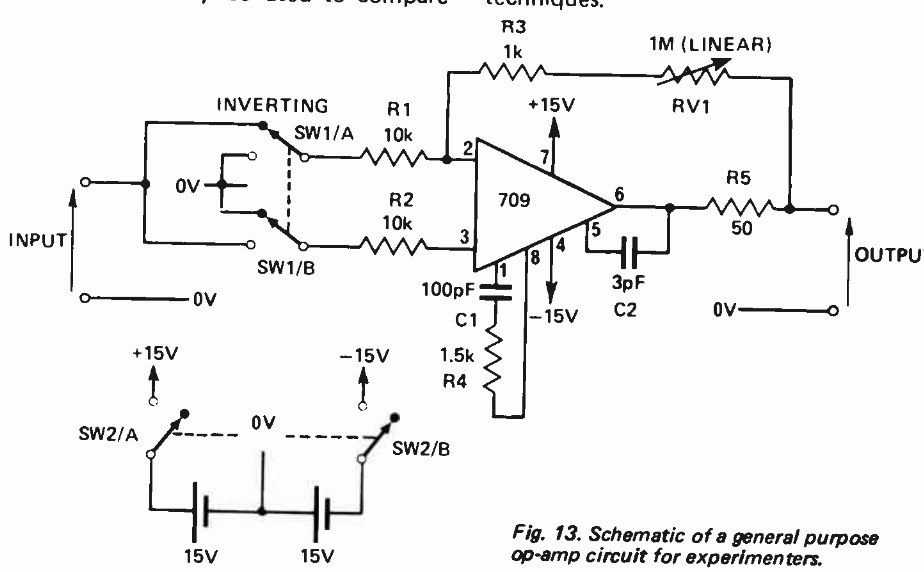
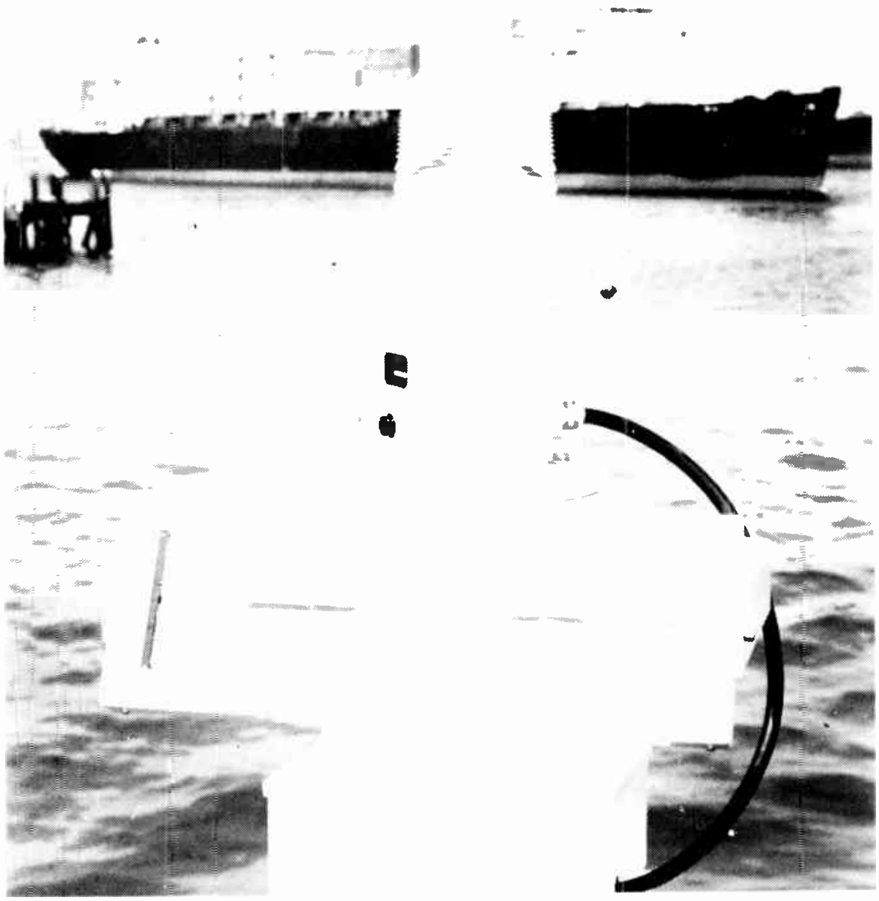


Fig. 13. Schematic of a general purpose op-amp circuit for experimenters.

13 The sources of power



Solar-cell powered buoy.

OUR COURSE, so far, has concentrated on developing basic electronic system blocks from combinations of passive and active components. You will have seen that, with each type of circuit, there is a requirement for some sort of power supply, although, there are some very rare circuits that may be powered by signal energy alone.

The provision of power for electronic circuits, is hence of primary importance. In the circuit illustrations, used so far, power supplies have been of a very simple kind, but, in some circumstances, they may be quite complex and expensive. Hence, before developing our circuitry still further we must gain a better understanding of the types of supply and the methods of implementing them.

The most commonly used source of electrical energy is that provided by the power mains and this, as we know, is alternating current (ac). However electronic circuits, in the majority of cases, need direct current (dc) supplies. Hence a discussion of power supplies for electronic systems must cover firstly the production and secondly the stabilization of dc voltages.

PROVISION OF DC

The source of dc power for electronic circuits, at any particular voltage, must be convenient, economic and easily started and stopped as required.

A wide range of basic power supplies is available to choose from – see Fig. 1 They range from tiny batteries to huge engine-driven generators. Each application has to be considered individually and the appropriate means chosen to suit the requirements of the circuit and the way it is to be used. Can the supply provide enough power? Does it provide the desired conditions of portability? – (in the field the weight of the supply may be critical). Is the method used economic? (batteries may be simple to use but their replacement can be costly). Is a non-portable supply already available for use? (such as the electricity mains). Sometimes a power supply already operating on some existing equipment may have adequate spare capacity.

There are many known methods of producing dc power. Batteries use electro-chemical action; rotating generators move conductors in a magnetic field to generate electricity; the mains supply (derived by rotating generators) is rectified to produce dc, fuel cells combine chemicals (still an exotic way to produce energy); thermo-electric systems generate electricity from thermo-couples or solar cells.

However the two most common sources of dc are firstly from batteries and secondly transformer/rectifier systems driven from the mains ac supply.

BATTERIES

In 1792 Italian anatomist Luigi Galvani, whilst working on dead frogs, discovered that the frog's legs twitched when touched with two dissimilar metals. The same phenomena occurred when the frog's legs were attached to an electrostatic generator. He (wrongly) attributed this to an effect which he called "animal electricity".

However, another Italian professor, Alessandro Volta, investigated the effect in 1800 and, showed that it did not depend on the animal tissue, but upon electrical generation due to two dissimilar metals being separated by a conductive solution. He thus showed two important things – that animal muscle was activated electrically, and that electricity could be generated chemically. (Previously only static electricity was known.)

Volta produced the first practical battery, called at that time a voltaic pile, by placing moistened paper sheets

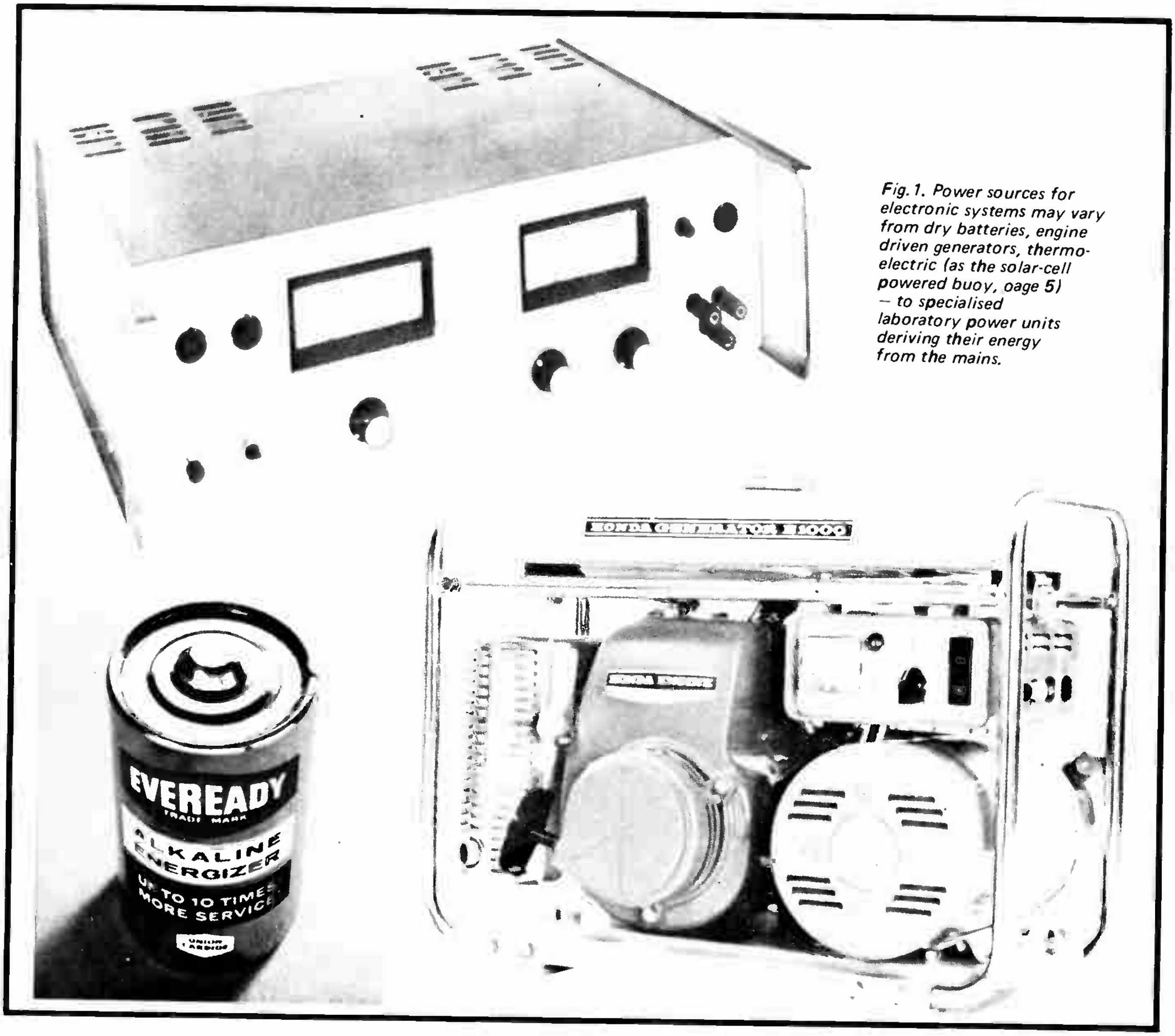


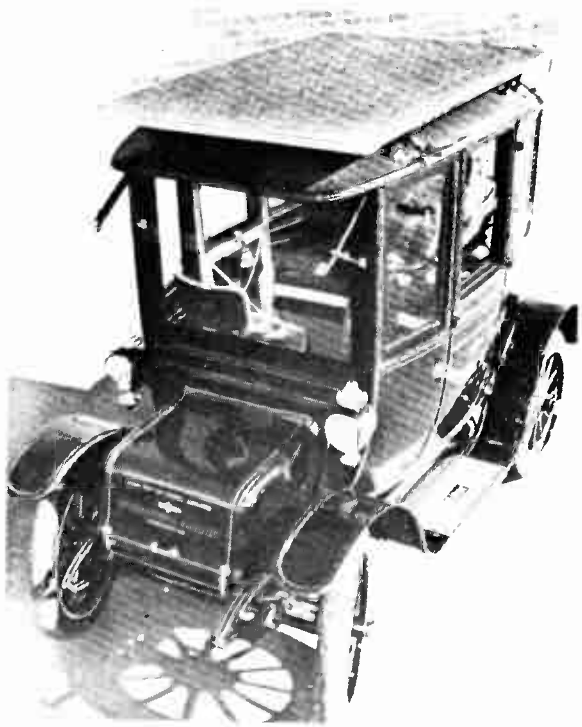
Fig. 1. Power sources for electronic systems may vary from dry batteries, engine driven generators, thermo-electric (as the solar-cell powered buoy, oage 5) – to specialised laboratory power units deriving their energy from the mains.

between alternate sheets of copper and zinc as shown in Fig. 2a. He also made cells in which the separating fluid (now called the electrolyte) was a liquid. His wet-cells used rods of zinc and copper, placed apart, in a diluted solution of sulphuric acid (Fig. 2b). Volta thought that the solution merely separated the electrodes without playing any vital role. We now know differently.

The fluid (it can also be a paste or solid) acts as an electrolyte. That is, the dissolved compound dissociates into positive and negative ions, however, the electrolyte has overall electrical balance.

When the copper and zinc electrodes are inserted an electric field is set up in the boundary layer between each electrode and the electrolyte. With the copper/zinc cell the copper is at a lower potential than the acid and the zinc is at an even lower potential.

The cell thus has an electromotive force between the electrodes which



This 1912 Baker Electric is now driven by solar energy! An array of 10 640 silicon solar cells mounted on the vehicle's roof charge intermediate storage batteries. Final drive is via the Baker's original dc electric motoı

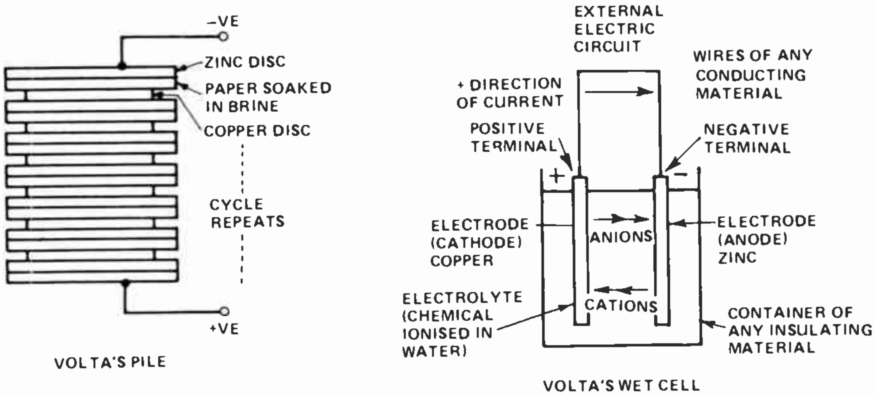


Fig.2. Cross sectional diagrams of the first electrochemical cells – VOLTA'S pile and wet cell.

depends on the difference in potential between the copper and the zinc.

When the electrodes are connected to allow electrons to flow, the dissociated ions move towards the electrode of opposite polarity. For example, in Volta's wet cell, the zinc electrode combines with the negative sulphate ions leaving the zinc electrode with an excess of electrons. These electrons flow through the external circuit to the copper electrode where they combine with the hydrogen ions to produce free hydrogen.

Many combinations of electrodes and electrolytes may be used to form cells in a similar manner. Some arrangements are more useful than others by virtue of higher energy capability, and hence many of the original systems developed have now been discarded as inefficient.

DEPOLARIZATION OF CELLS

The formation of gas on an electrode (hydrogen in the voltaic cell) becomes an effective insulator and may cause the cell to cease working efficiently or even completely. If the gas (or other product, eg solid in some cells) can be chemically removed, as it is formed, the cell will continue to produce power until the negative plate material has been used up – it redeposits on the other plate. Such an additive, which maintains full cell efficiency, is known as a depolarizer.

PRACTICAL BATTERIES

The electrochemical process just described can be optimized to either *produce* electricity or to *store* it for reuse. Cells providing power from an initial chemical charge are called primary cells. Those that are made intentionally to store power are called secondary cells (also called accumulators in earlier literature). Some combinations and designs will act as both, but usually a primary cell is a throwaway item. A secondary cell usually requires charging (the process of storing electrical energy) after manufacture, and may be recharged as often as is necessary.

PRIMARY CELLS

The most commonly used primary cell is the well-known dry-cell (or more correctly, the Leclanche cell, after the original developer who introduced it in 1877). It is made, as shown in Fig. 3, from a zinc can containing a central carbon rod surrounded by, firstly, a depolariser (manganese dioxide) and then the electrolyte which is in paste form (ammonium chloride, zinc chloride, water and a filler material). The basic cell is made in many sizes and is also packaged as groups of cells connected in series and/or parallel to provide either greater capacity at the 1.5 V delivered per cell – or increased voltage. For example 90 V batteries (constructed from sixty 1.5 volt cells) were extensively used in the days of valve-circuit portable radios.

There are many alternatives to the basic Leclanche cell. All have characteristics which make them suitable for low power, portable applications. The characteristics of the different primary cells are given in Table 1.

The mercury cell, developed in the 40's, is far more rugged than the Leclanche cell and retains its voltage better over long periods of light use or storage – several years is typical. These use zinc and mercuric oxide (or graphite) electrodes with alkaline hydroxide electrolyte. A typical arrangement is shown in Fig. 4. They can be made extremely small in size and are ideal for powering very small equipment, such as hearing aids, or for equipment used intermittently such as photographic light meters.

Another cell available today is the alkaline-manganese battery. Its interior design consists of pellets of anode and cathode materials; zinc and carbon are used. The manganese dioxide depolariser is arranged to be more efficient than in the common dry-cell and the electrolyte is potassium hydroxide. This battery has an excellent shelf-life and is capable of sustaining a high discharge rate.

Several other primary cells will be encountered in electronic instrumentation – The Daniell cell 1836 (copper, zinc and sulphuric acid). the Clark cell 1872, and the Weston cell 1892 (mercury, cadmium amalgam and cadmium sulphate solution, as

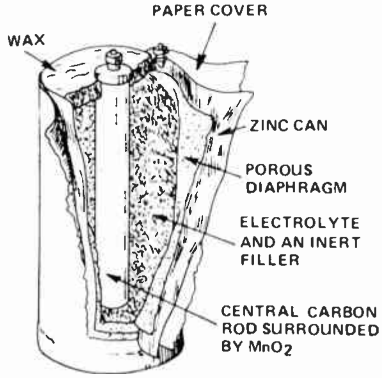


Fig.3. The common dry cell was originally developed by Leclanche in 1877. It produces power for a limited period and is then discarded.

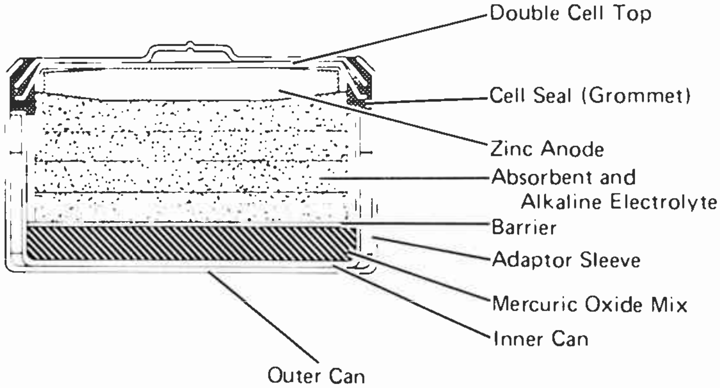


Fig.4. Cross section of a typical mercury cell.

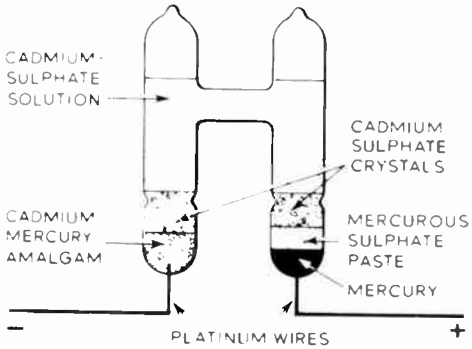


Fig.5. The Weston standard cell delivers 1.0186 volts provided the load is minimal. It has found extensive use as a standard of the voltage unit.

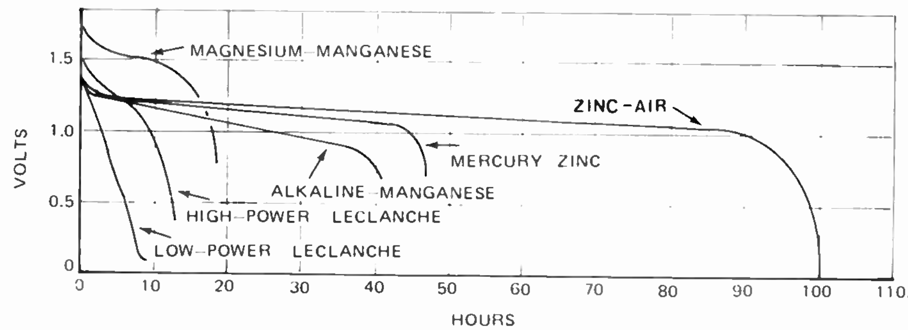


Fig.6. Comparative chart showing voltage characteristics of similar size units of various types of dry cell battery.

shown in Fig. 5) are the three cells which were used internationally at various times to define the standard of voltage. The latest voltage standard has recently been changed to use the so-called Josephson solid-state effect, but the Weston cell is adequate for many voltage calibration tasks (1.0186 volts). Standard cells are used only to provide accurately-known and time-stable voltage, but only at low current. They are not intended for power use.

A more recent development are zinc-air cells. These use a zinc powder anode in contact with potassium hydroxide electrolyte. The cathode is a porous arrangement that breathes to atmosphere making use of oxygen, via an intermediate process and a catalyst, to produce hydroxyl ions which enable current flow to occur.

The silver-zinc primary cell has high energy density and discharge rate but because of high cost, is restricted to exotic applications such as spacecraft electronics.

Each type of cell has its own particular merits. Figure 6 shows the voltage-time curves for an ideal loading condition along with comparative figures for the commonly used cells. Leclanche cells operate best in intermittent service, where high currents are needed, or continuously for low drains. Mercury cells especially suit low current demands for very prolonged periods. Zinc-air batteries work best for high current loads maintaining voltage uniformly over considerable periods. Silver-zinc provides the highest available energy density.

The relative cost of each should be considered in selection along with the requirement. It may well be more economical, in the not too long run, to use the more expensive alternatives.

SECONDARY CELLS

We have seen that the electrical energy provided by a primary cell is derived from a chemical process. From school chemistry we know that, when zinc is dissolved in sulphuric acid, a large amount of energy is released as heat. In the voltaic cell this energy is released as electricity rather than as heat. If the reaction is *not* reversible the cell is a primary cell and is thrown away when exhausted.

There are however others in which the reaction *is* reversible and these are known as secondary cells. For the system to be reversible the electrolyte and electrodes must be capable of being converted back to their original state after discharge. This reversal is *not* spontaneous. The cell must have the electrical energy pumped back into it. That is – it must be charged.

The commonest arrangement (in use since the last century) is the lead-acid battery, such as is used to start cars and to power the auxiliary circuits. The second most commonly used is the nickel-iron cell.

The lead-acid battery consists basically of a plate of lead (negative electrode) and a plate of lead dioxide (positive electrode) immersed in dilute sulphuric acid – as shown in Fig. 7. As the cell discharges, the lead electrode and sulphate ions in the electrolyte combine to produce lead sulphate plus electrons, and the lead-dioxide combines with sulphate ions, hydrogen ions and electrons to produce lead sulphate plus water. The insoluble lead sulphate adheres to the plates, finally shielding them from further electrochemical reaction – the cell is then discharged. The recharging process reverses the reactions, rebuilding the electrode material as the lead sulphate is removed from solutions to produce sulphuric acid and electrode. The nominal voltage produced is 2.0 V. As water is liberated the cell is easiest vented to air, but it can be made as a sealed cell.

The nickel-iron cell, invented by the Edison Company at the turn of this century, uses oxides of iron and nickel as the electrodes together with potassium hydroxide electrolyte. The electrochemical action is similar to the

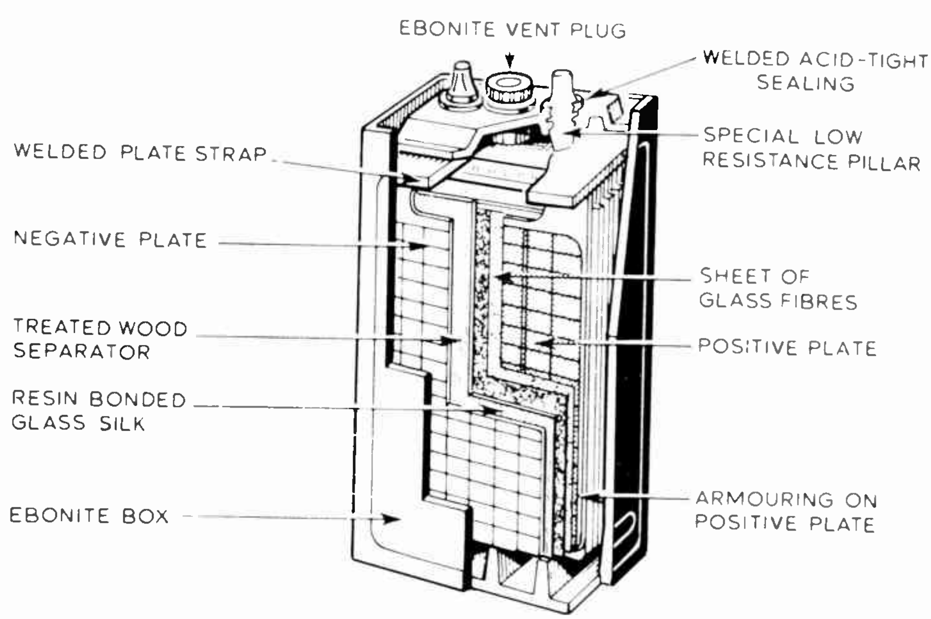


Fig.7. Interior of lead-acid storage cell. Electricity is stored by virtue of chemical reactions induced by charging the cell with electricity.

lead-acid battery — electrodes and electrolyte combine releasing electrons and the process is reversible. These cells can be sealed without difficulty, they are more rugged, give a longer life than lead-acid cells, but cost more.

In the search for more storage capacity for unit volume and weight, research has yielded some exotic battery designs. Silver and zinc are used in a design originated by Andre in the 1930s. Clearly the cost is higher but the considerable gains in weight reduction may make them attractive where weight is a major cost factor — missiles, satellites and man-packed equipment.

As it is now clear that a new kind of storage battery will be in extensive use within this decade we include a brief description of the high-temperature batteries now approaching market production. These cells, also use electrodes and electrolytes, but run at temperatures up to 400°C, and can provide at least four times the storage capacity at the same cost and weight as lead-acid cells. The need for high-temperature operation does, however, exclude them from low power applications. The two main contenders are the sodium/sulphur battery that uses liquid sodium and sulphur electrodes with solid alumina electrolyte (the most developed to date) and the lithium/sulphur battery that uses liquid lithium and sulphur electrodes with molten salt electrolyte (the most theoretically efficient cell). This latter type, will probably be more costly to produce. Both of these types, plus several other high temperature arrangements, have been used in prototype situations — powering electric cars is the dominant requirement, but large scale mains-power, system-float storage will be the main usage in the future.

The range of storage cells available for powering electronic circuits is therefore broad, and the type must be chosen to suit the application. For circuits having only medium demands, electronic flash units, calculator supplies — small rechargeable nickel-cadmium cells are best. These are made in the same shape as mercury or Leclanche cells allowing them to replace primary cells and be recharged when needed.

Table 1—Primary Batteries

	Leclanché (Dry Cells) Cylindrical	Alkaline-Mercuric Oxide: Mercuric Oxide	Alkaline-Mercuric Oxide: Mercuric Oxide Manganese Dioxide	Alkaline Manganese Dioxide	Magnesium-Manganese Dioxide	Zinc-Air	Solid State	Silver-Zinc
Cap. Avail. (Ah)	0.350 - 21	0.075 - 14.0	0.036 - 3.6	0.580 - 10.0	2.0 - 9.0	3 - 25	0.010 - 1.5	1.5 - 220
Open Circuit Voltage (V)	1.55 - 1.70	1.35	1.40	1.50	2.0	1.45	0.66	1.86
Nom. Operating Voltage (V)	1.25	1.25	1.25	1.25	1.55	1.1	0.55	1.45
Recom. Dischg. Temp. (°F)	65 - 85	65 - 130	65 - 130	65 - 115	65 - 130	50 - 100	40 - 120	50 - 90
Recom. Storage Temp. (°F)	−40 - 75	−10 - 80	−10 - 80	−40 - 80	−40 - 120	−80 - 100	−65 - 120	32 - 90
Self-Dischg. Rate/Mo. at R. T. (%)	1.0 - 1.5	0.8 - 0.9	0.8 - 0.9	0.8 - 0.9	0.5 - 2.0	0.2 - 1.0	0.02 - 0.25	N/A
Watt Hr./Lb	15.5 - 34	37 - 48	34 - 52	33 - 38	55 - 60	80 - 150	5 - 10	40 - 80
Watt Hr./Cu. In.	0.9 - 2.7	4.7 - 6.0	4.5 - 7.5	3.8 - 3.9	3.30 - 3.70	10 - 15	0.6 - 1.2	2.37 - 6.51
$/Watt Hr. (approx)	0.04 - 0.17	0.16 - 5.4	0.20 - 4.4	0.04 - 0.41	0.08 - 0.20	0.02 - 0.04	0.30 - 30.0	0.60 - 8.0
Characteristic Features	Inexpensive, available in a large variety of sizes & battery volt.	Excellent shelf life, High energy density	Excellent shelf life, High energy density	Excellent shelf life, High-rate discharge cap.	High Operating voltage, High storage cap.	High energy density	Excellent shelf life, Low temp. operating cap.	High energy density, High-rate cap but very expensive.

Table 2—Practical Secondary Systems

	SEALED NICKEL-CADMUIM			SILVER-ZINC			LEAD-ACID SYSTEMS					
									STATIONARY			
	Cylindrical	Button	Rectangular	Low Rate SZR(L)	High Rate SZR	Fast Activating SZFA	Auto	Motive Power	Antimony	Calcium	Planté	Sealed Gelyte
Cap. Avail. (Ah)	0.100 - 7.0	0.02 - 0.50	11 - 23	1 - 140	1 - 150	1 - 180	33 - 340	180 - 2175	10 - 8000	50 - 2550	8 - 996	6 - 9
Open Circuit Voltage (V)	1.30	1.30	1.30	1.86	1.86	1.86	2.10	2.12	2.06	2.06	2.06	2.10
Nom. Operating Voltage (V)	1.25	1.25	1.25	1.45	1.45	1.45	1.98	1.94	1.94	1.94	1.94	1.97
Nom. End-of-Chg. Voltage (V)	1.48	1.48	1.48	2.05	2.05	2.05	2.53	2.55	2.17 @ Float	2.17 @ Float	2.17 @ Float	2.55
Recom. Dischg. Temp. (°F)	65 - 85	65 - 85	65 - 85	50 - 90	50 - 90	50 - 90	70 - 90	70 - 110	70 - 90	70 - 90	70 - 90	70 - 90
Recom. Storage Temp., Wet Chg'd (°F)	−40 - 80	−40 - 80	−40 - 80	32 - 90	32 - 90	32 - 90	−40 - 115	30 - 77	−40 - 80	−40 - 80	−40 - 80	0 - 50
Recom. Storage Temp., Dry Chg'd (°F)	N/A	N/A	N/A	32 - 90	32 - 90	32 - 90	−40 - 115	32 - 100	−40 - 115	−40 - 115	−40 - 115	N/A
Self-Dischg. Rate/Mo. at R. T., Wet Chg'd (%)	10 - 15	5 - 8	5 - 8	2 - 5	2 - 5	5 - 10	5 - 11	7 - 10	7 - 12.5	1.0	3.0	7 - 12
Watt Hr./Lb.	8.3 - 19.0	10 - 12	7.4 - 9.2	32 - 60	38 - 66	36 - 73	12.7 - 21.8	8.6 - 11.0	4.8 - 9.7	5.7 - 9.7	3.9 - 6.5	14.5
Watt Hr./Cu. In.	0.85 - 2.20	0.64 - 0.90	0.62 - 0.73	1.66 - 4.20	1.95 - 4.61	2.20 - 5.22	0.79 - 1.6	1.08 - 1.37	0.27 - 0.84	0.43 - 0.84	0.22 - 0.58	1.16 - 1.50
Cycle Life (nom. cycles expectation)	250 - 10,000	250 - 10,000	250 - 10,000	25 - 50	10 - 20	2 - 5	150 - 250	1000 - 2000	N/A	N/A	N/A	100 - 1000
Calendar Life (nom. yr expectation)	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	~15	15 - 24	~24	N/A
$/Watt Hr. (approx)	0.90 - 10.00	3.50 - 52.00	1.10 - 1.43	0.75 - 11.00	0.80 - 11.00	0.65 - 11.00	0.012 - 0.024	0.05 - 0.09	0.09 - 0.42	0.08 - 0.28	0.11 - 0.54	0.22 - 0.30
$/Watt Hr./Cycle (approx)	0.004 - 0.04	0.014 - 0.21	0.004 - 0.006	0.030 - 0.44	0.080 - 1.10	0.37 - 5.50	0.00008 0.00016	0.00005 0.00009	N/A	N/A	N/A	0.0022 0.0030
Characteristic Features	Operative in any position, no maint.	Operative in any position, no maint.	Operative in any position, no maint.	High rate capability, High energy density	High rate capability High energy density	High rate capability, High energy density	Inexpensive Excellent high-rate capability.	Excellent cycle life, Rugged const.	Rugged Const. Wide range of available cap.	Lowest float current, Excellent life	Long life, High Reliab.	No Maint. Inexpensive

14 Simple power supplies

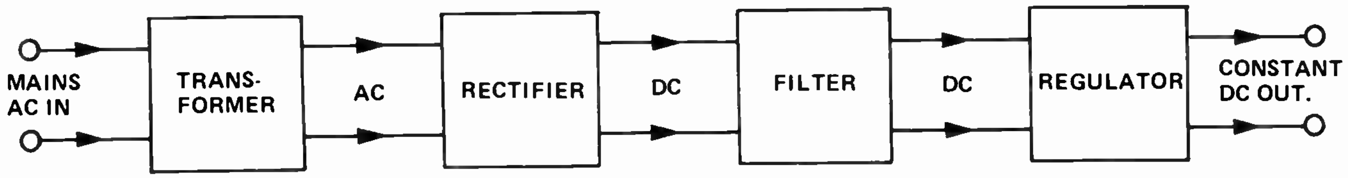


Fig. 1. The various sections required in the process of converting the ac mains supply into a source of dc power.

AN AC SUPPLY provides a sinewave current that changes direction at the supply frequency. Firstly, the ac voltage has to be transformed to the appropriate voltage level. To obtain dc a switch (the rectifier) is needed to reverse polarity of alternative half cycles. This done, all that remains to be added is a method of smoothing out (filtering) the half-sinusoids to obtain a steady current. We will look at each of these steps in turn.

TRANSFORMERS

The principles of inductance were briefly introduced in Part 6 of this course. We suggest that the section be read again.

If two inductors A & B are placed such that the axis of their coils align (as in Fig. 2), and coil A is energised with an ac source a voltage will be generated across coil B.

As we move the coils closer to each other the voltage developed, across coil B, approaches a value which is proportional to that across coil A. The proportion will be equal to the ratio of the number of turns on B, to the number of turns on A.

ie $$\frac{E_B}{E_A} = \frac{N_B}{N_A}$$

Where E_B = voltage across coil B
E_A = voltage across coil A
N_B = turns in coil B
N_A = turns in coil A

The effect is due to the field of one coil cutting the turns of the other and is known as mutual inductance. If the coils are wound on top of each other, and an iron core is used, the coupling is improved to almost unity and we have a device capable of changing ac voltage from one level to another. Such a device is known as a transformer.

There are losses in the transformer due to the resistance of the wire in the coils — these are known as copper losses, and in the iron of the magnetic core — these are known as iron losses.

A transformer can never *create* power — it can only transfer it and change voltage levels. Small transformers have power efficiencies from 60-90%; 85% is typical.

To reduce the iron losses as much as possible the core material (at frequencies below 20 kHz) is usually a special silicon steel called "transformer iron". The core is built up of thin laminations of this iron individually insulated by a thin coating of lacquer. By this means eddy current (circulating currents within the core) losses are reduced to a minimum.

Note particularly that the transformer is an ac device. It will *only* produce voltage in the secondary winding when there is a current change in the primary. A dc current flowing in the primary will not produce a secondary output.

The iron laminations retain the magnetic field ensuring virtually total magnetic linkage between coils. For high frequencies, up to several megahertz, ferrite powder mouldings are often used. In many high-frequency applications, the ferrous magnetic circuit is omitted altogether. Figure 3 shows a range of transformers for use at various frequencies.

In mains-operated power supplies the relatively low frequency of the mains leads to efficient coupling. Hence the ratio of input/output voltage is as the ratio of input turns/output turns. A transformer is, therefore, selected to provide the correct voltage (stepped down or up) and must be designed with wire in each winding heavy enough to carry the currents needed without overheating. Usually selection of a transformer is made from manufacturers' product lists using the nearest listed, with any difference being on the conservative side — higher voltage or higher current capability when the exact requirement is not available. The power capacity of transformers is stated as the volt-amp

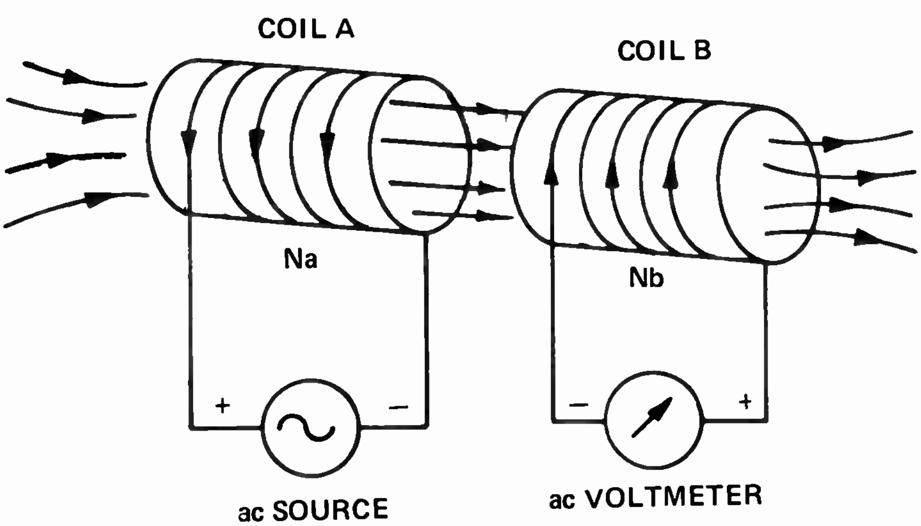
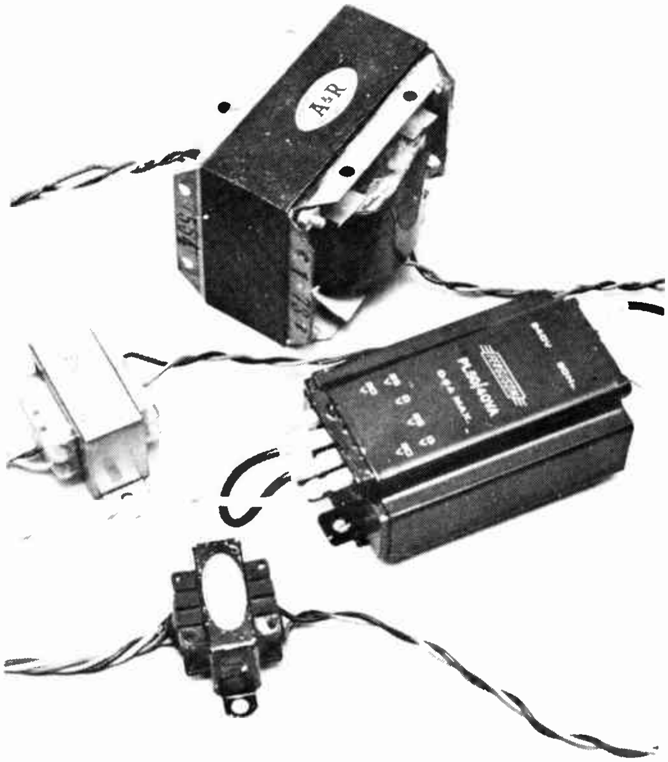
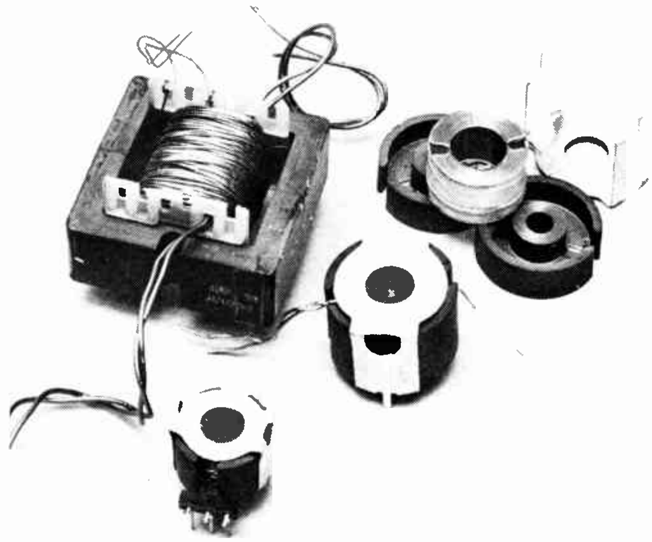


Fig. 2. Transformer relies on the principle that when lines of magnetic force move through a coil, a voltage is induced in the coil which is proportional to the number of turns in the coil.





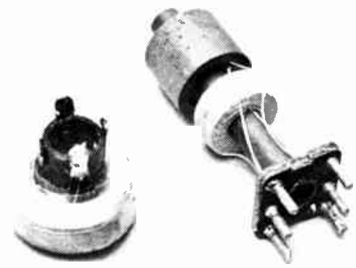


Fig. 3. The design of a transformer depends greatly on the frequency of operation and the amount of power to be handled.
At low frequencies (eg 50 Hz mains) a laminated silicon-steel core is required, (TOP LEFT).
At medium frequencies a ferrite core or slug may well be used to adjust as well as increase inductance (50 kHz to several MHz).
At high frequencies (eg 50 MHz and above) air spaced coils may be all that is necessary, (RIGHT).

Fig. 4. These characteristics of common rectifier arrangements will help you select a transformer to obtain a particular dc output.

COMMON RECTIFIER ARRANGEMENTS

	1 Cycle Output Waveform	Average dc Volts Output	RMS Volts at Output	Peak Volts Output	Peak Reverse Rectifier Voltage	Percent Ripple RMS/dc out
(A) 1 ϕ HALF WAVE		1	1.57	3.14	3.14	121%
(B) 1 ϕ FULL WAVE CCT		1	1.11	1.57	3.14	48%
(C) 1 ϕ FULL WAVE BRIDGE		1	1.11	1.57	1.57	48%
(D) 3 ϕ STAR (WYE)		1	1.02	1.21	2.09	18.3%
(E) 3 ϕ BRIDGE		1	1.00	1.05	1.05	4.2%

product of the total output or input. This can be found as the sum of volts times amps of all of the secondary output circuits plus about 10% for losses.

All transformers have rms rated outputs. In practice this voltage is the *unloaded* output voltage and may vary from transformer to transformer. Additionally, because of the finite winding impedance, the transformer output will drop when loaded. This effect, known as transformer 'regulation', is quoted as the percentage voltage between load and no-load. In prototype designs it is therefore advisable to use a transformer with a number of tappings so that the correct rms output may be selected on test.

RECTIFIER STAGES

Many different rectifier systems may be used, Fig.4 shows those most commonly encountered together with their schematic diagrams and relevant conversion factors. Note that the dc output *is not* the same as the ac input. A mistake commonly made by beginners is to assume that the dc output from the rectifier will be the same or less than the rms output from the transformer.

A single rectifier , as in A, gives half wave operation only and clearly, whilst saving a rectifier element, only allows half the sine-wave through with a resultant drop in average dc output. The gain in saving rectifier elements is offset by the need to provide a higher output voltage from the transformer and a more powerful filter to smooth out the pulsating dc current (121% ripple!).

Clearly, fullwave rectification (that is, use of both half cycles) is better but it requires more rectifiers. There are two main methods. One uses four rectifiers to create a 'reversing' switch — the so-called bridge circuit. Output current from the transformer of one polarity passes through to the filter stage using two of the rectifiers; the next direction of current is then allowed through by the second pair which are connected to accept reverse current polarities. An alternative full-wave method uses only two diodes instead of four. It works as two half-wave systems that alternately connect to the common filter terminals with the same polarity. It uses less rectifier elements than in a bridge circuit but requires a centre-tapped transformer. Rectifier diodes for bridge circuits are available ready-packaged as a full 4-element bridge in a common encapsulated unit.

Where a three-phase (the normal industrial high-power mains) supply is available, other rectifier arrangements are possible — as shown. As the number of phase half-cycles used is increased the dc produced becomes smoother, relaxing the degree of filtering needed. Other more sophisticated six-phase systems (using special transformers) are used industrially.

Originally, rectifier elements were either vacuum-tube diodes (two-element tubes) or specially made contacting surfaces of copper oxide or selenium. Although both of these are still in service, they have been more or less superseded by modern solid-state, two-layer semiconductor diodes (in the simplest form) and by the family of multiple-layer semiconductor devices in which the current can be controlled as well as being rectified. (These devices, SCRs and TRIACs, will be covered later.)

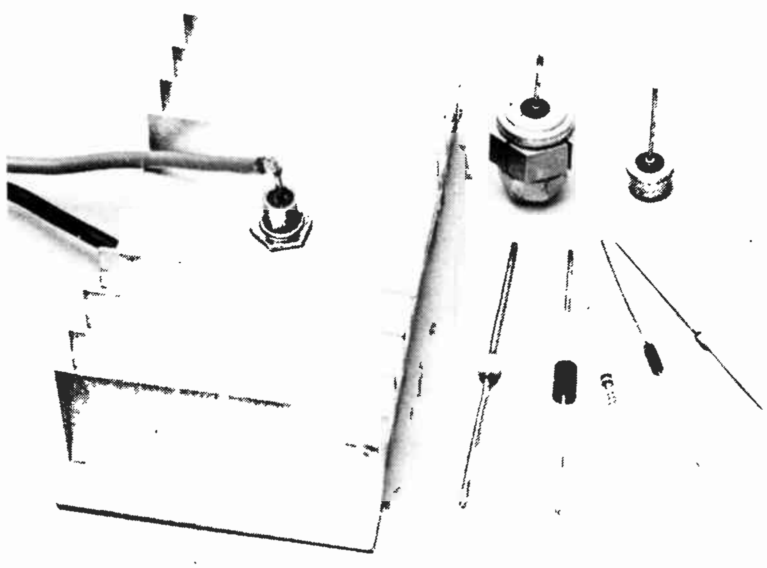
Virtually all diodes designed for power rectification are now silicon devices — although germanium still finds some use for low-power, signal-detection diodes. The power handling capability of a diode depends upon the voltage drop across it and the current flowing through it. These determine the heat to be dissipated at the diode junction. Provided the *junction* itself is maintained below its maximum safe value, all is well. Heat sinks are usually used to help liberate this waste heat, thereby raising the current capacity of the rectifier units. When selecting diodes for power use it is necessary to ensure that they can safely withstand the peak reverse voltage of the waveform — this can be as much as three times the quoted ac value (which is usually the rms value). In the manufacturer's data this is shown as the peak inverse voltage (PIV). In a half-wave circuit supplying, say, a 100 Vdc output, the peak inverse voltage rises to 314 V!

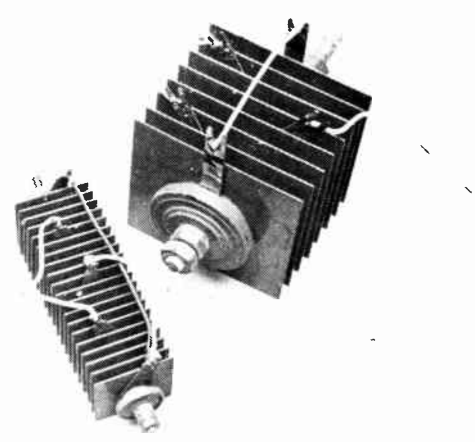
Diodes come in all shapes and sizes as Fig. 5 shows. Large power diodes are intended to be mounted on heat sinks and the manufacturers have built them accordingly to ensure good thermal contact. Special heat-sink extrusion is made for this purpose.

Individual diodes in a bridge circuit must be insulated from one another — nevertheless it is often convenient to mount them on a common heat-sink. Mica washers are often used for this purpose as they provide good electrical

Fig.5a, A selection of commonly used solid state rectifiers. High-power diodes are sometimes mounted on a heatsink to help radiate the heat generated due to internal losses.

Fig.5b, Selenium rectifiers were quite bulky but were extensively used in valve radio days.





insulation whilst allowing heat to pass through.

The current rating needed of the diodes depends upon the rectifier circuit. If half-wave it must be able to handle the full current expected. For full-wave bridge or centre-tapped single-phase arrangements, the diodes only switch on alternate half cycles and, therefore, can be rated for half the output load current. Special care must be used when silicon diodes are used. The initial onrush of current to the uncharged filter capacitors can exceed the safe maximum of the diodes unless adequate limiting resistance exists in the transformer winding or input leads. Typically, the peak current may be as much as 10 times the average dc current.

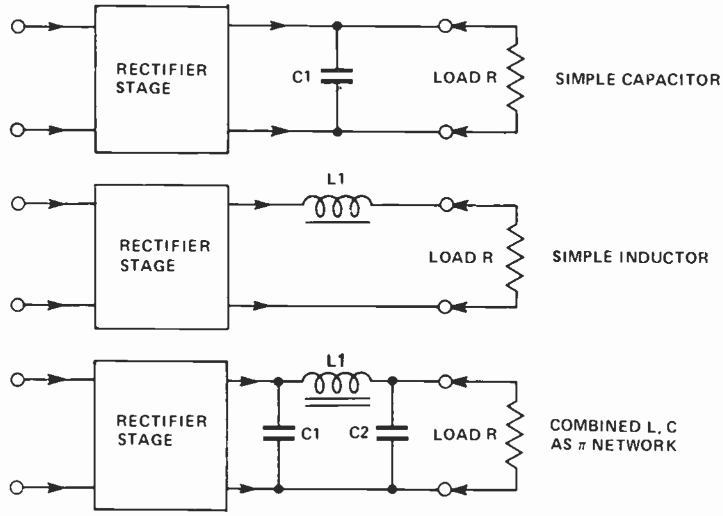


Fig.6. Various types of filter may be used to smooth the pulsating dc from the rectifier. (a) a simple capacitor (b) a simple inductor. (c) a combination of capacitance and inductance (pye filter).

FILTERING

The output of any rectifier system consists of a train of half-sinusoid waveshapes. We know that all waveshapes can be constructed by adding a number of pure sinusoidal signals. Thus the rectifier output is a complex waveshape containing a basic dc level plus many other frequencies. To smooth the signal, therefore, a low pass filter is needed that rejects all frequencies above dc (frequency of zero).

Several alternative methods of filtering are available. The commonest, shown in Fig. 6, is to use a large value shunt capacitor across the output terminals. At each new half-cycle the diodes pass a burst of current into the capacitor to recharge it, making up for charge drawn by the circuit load on the supply. By appropriate choice of capacitor size for a given load and adequately low bridge resistance (this decides how quickly the charge will enter the capacitor), the supply can be made to hold a voltage up near the peak value of the waveform. However care must be taken to ensure that the peak current rating of the diodes is not exceeded.

In applications where a relatively large power level is involved it may be more economical to use another method. The shunt capacitor method, above, provides a short-circuit path to high frequency signals (capacitive reactance falls with increasing frequency) thereby shunting them away. Only dc is unattenuated. The same effect may be achieved if an inductor is used, as shown in Fig. 6 – but this time in series with load. The inductor provides lowest impedance to lowest frequency so dc passes virtually without loss (provided the dc resistance of the winding is low – hence the high cost of effective filter inductors) but provides increasing impedance as the signal frequency rises.

These two basic methods can be taken further again using both together to increase the frequency rejection. We will not pursue the design of sophisticated power supply filters for they tend to be rather specialised. Note, however, that the filtering effect depends largely upon the magnitude of the load current drawn. This can be seen by regarding the filter component reactance and the load impedance as a series or parallel network (see Fig. 7) in which the supply voltage is the output produced across the load impedance.

An increasing load current occurs due to a reduction in load impedance (usually regarded as a resistive load). The series inductive method provides less ripple (the name given to the ac component-present) as the load increases. On the other hand, with the capacitive shunt method the ripple increases as load increases. Hence the two methods complement each other and (as neither is ideal) the two are combined in more advanced filtering methods.

It should now be clear that the rectifier stage design will largely determine the specification of the transformer and that the filter method

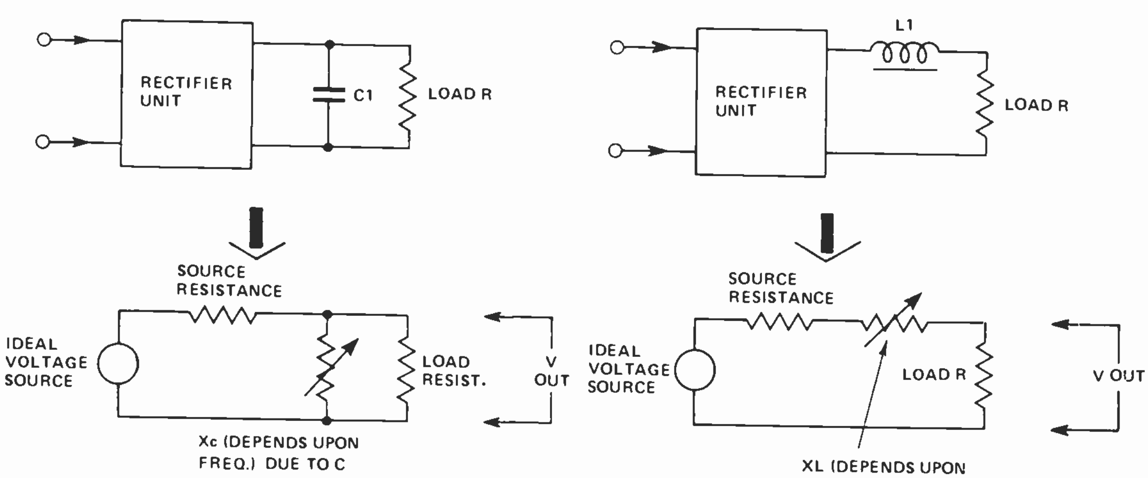


Fig. 7. The performance of a simple filter may be evaluated by replacing the capacitor (etc) by its equivalent resistance at the ripple frequency. Thus we have a voltage divider due to this and the source impedance. We may also from such equivalent circuits calculate the degree of regulation for any given load.

ELECTRONICS -it's easy!

must also be considered in the overall design.

Power supply design is not as straightforward as might at first be thought. Each stage determines the requirements of the other stages so a certain degree of skill and experience is needed to reach a satisfactory design. Furthermore, as we will see later, the design must also make allowances for the way the supply is to be used and for the method of stabilisation employed.

POWER SUPPLY TERMS

The two forms of power supply — voltage or current — as we have seen earlier, can be represented as black boxes which consist simply of a source (voltage or current) and an equivalent resistance value. A voltage source ideally maintains the required voltage regardless of load current. A current source, the reverse situation, provides the required current regardless of output voltage. Practical supplies have a finite resistance value (the ideal of zero output impedance is unobtainable) but it is possible to produce a circuit that is close enough to the ideal for practical purposes.

Let us now see what happens to a voltage supply as the load current increases. We see from Fig. 7 that the voltage appearing across the load is that produced by a perfect generator driving a divider chain. Hence, provided the source resistance is much smaller (at least ten times smaller) than the minimum load resistance, the change in voltage across the load as the load current varies will be negligible. The aim, therefore, in good voltage supply design is to produce a unit with low internal resistance. Factors of one thousandth are typically obtained.

Constant voltage supply is by far the most common requirement, but there are also many applications for constant current supplies. In addition there are other supplies available with special characteristics.

Because of finite internal power-supply resistance the voltage output of basic supplies (caused in reality by the resistance of the diodes, transformer losses and filter resistances) drops as the load current increases. All these effects produce voltage drops that subtract from the original voltage source. The ratio of no load voltage (less full load voltage) to the no load voltage is called the regulation of the supply. This is expressed as a percentage.

IMPROVING REGULATION

In some instances, battery supplies for example, the internal resistance is adequately low and the output remains reasonably constant with time and changing load. A lead-acid storage battery for example will provide voltage constant to about 0.1% for quite a long time as long as the load is fairly low.

Mains derived supplies, however, exhibit poor regulation, unless (costly) stabilising circuits are added. Apart from this their output is also proportional to changes in mains voltage — which can fluctuate by as much as ±10%.

In many electronic systems the voltage must remain constant regardless of changes of mains input and load and changes in component values with time. Consequently, basic sources of dc power are often followed by a unit known as a regulator. Its role is to maintain the output constant to a chosen degree (0.1% changes in output due to load or input changes is typical). The degree of stability obtained relates to cost. Techniques cover a wide range — from a single special diode and a resistor, to multiple transistor circuits and special purpose IC's.

Power systems such as these will be covered in the next part of this series.

REFERENCES

"D.C. Power Supply Handbook" — Hewlett Packard, 1973.

Power Supplies" — B. Doherty, Pt. 1. E.T.I. September, 1971. Pt. 2 E.T.I. October, 1971.

"Batteries", R. Harrison, E.T.I. November 1977.

"Power Supply Handbook" — Kepco Inc. (1966) Flushing, N.Y. 11352, U.S.

ELECTRONICS – in practice

THIS project will provide more experience in the use of operational amplifiers. The circuit uses two amplifiers, illustrates a number of new points and provides a very useful piece of equipment.

A MIXER-PREAMPLIFIER CIRCUIT USING OP-AMPS

The signal provided by a sensor operating at audio frequencies, eg a microphone, a guitar-string vibration sensor, a record-player cartridge — needs boosting before the signal is used to drive a main amplifier or recording unit. The preamplifier shown here accepts signals of around 2 mV level, has an input impedance of

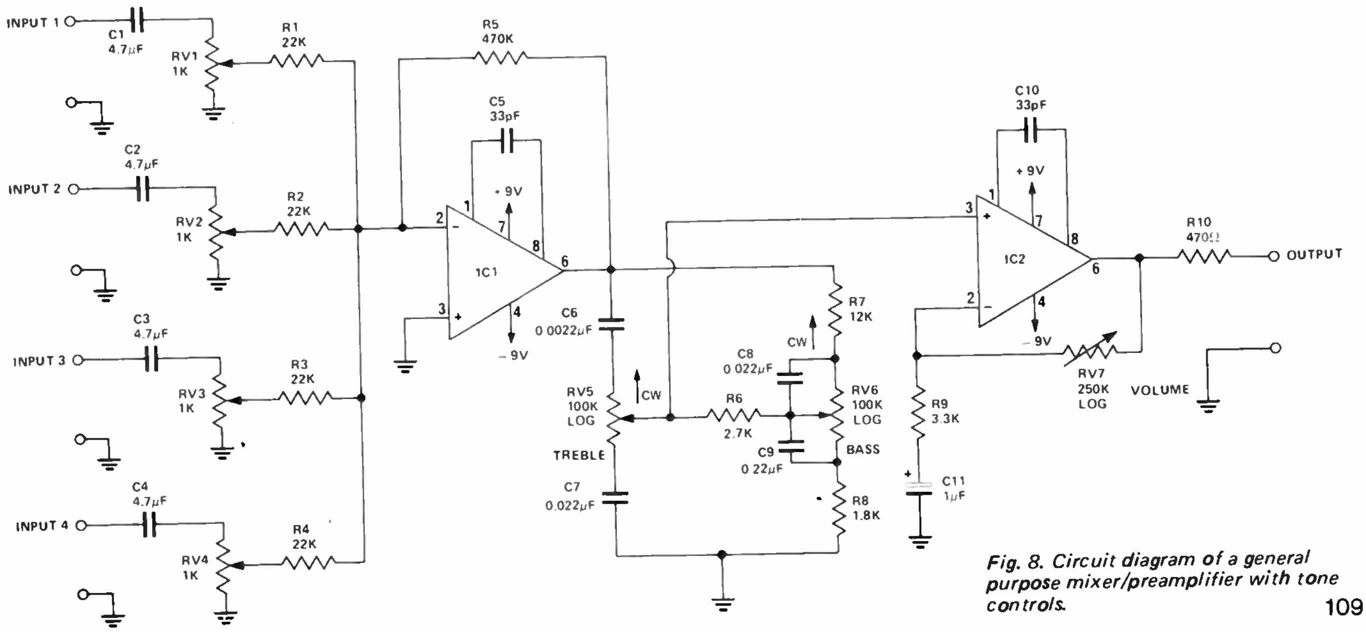


Fig. 8. Circuit diagram of a general purpose mixer/preamplifier with tone controls.

1 k, provides a gain of approx. 1600, and has an output swing of up to 3.2 V for 2 mV input. It introduces comparatively little distortion and is designed to accept four inputs, each having a level control. A special tone control network is incorporated that enables bass and treble signal frequencies to be varied over ±10 dB (at 100 Hz and 10 kHz respectively). Although primarily intended for mixing audio signals in entertainment applications, the circuit can also be used as a single-input, variable-gain unit in any application where gain and frequency adjustment are needed.

OPERATION OF THE MIXER-PREAMPLIFIER

Each input of the circuit given in Fig. 8 is ac coupled and has an attenuating potentiometer that allows the gain of each input channel to be independently adjusted as required. Four such inputs are summed by an inexpensive IC op-amp, connected as a summation circuit, having a maximum stage voltage gain of around 20 (25 dB).

The output of this stage feeds the next stage via a conventional tone control network which either attenuates or boosts bass and treble frequencies according to the settings of each potentiometer. Note that the second stage op-amp is connected as a non-inverting (the output signal has the same polarity as the input) single-input amplifier stage having a maximum gain of about 80 (37 dB). The feedback resistor, in this case, is a potentiometer allowing the overall gain of the unit to be varied. Thus this potentiometer acts as a master gain control.

In the circuit diagram (Fig. 8) the power supply connections are not shown. This is usual in op-amp circuitry to avoid over complicating the diagram. The connections are — positive to pin 7 and negative to pin 4. These connections are, of course, made on the printed circuit board.

A simple power supply (Fig. 9) may be used if batteries are unsuitable. This provides the positive and negative

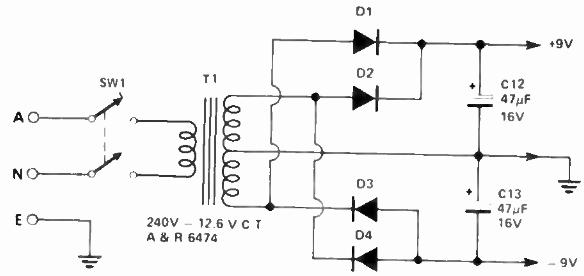


Fig. 9. Circuit diagram of an unregulated power supply suitable for use with the preamplifier of Fig. 8.

supplies necessary for the op-amp. At first glance the circuit appears to be that of a full-wave bridge. In reality it is two separate supplies, driven from different sides of a centre tapped transformer each being connected in the opposite way to provide opposite polarities.

Note that the transformer supplies a total of 12.6 volts rms, that is 6.3 volts on either side of the centre tap. This, when rectified and filtered, provides 9 volts dc (capacitor charges to peak of waveforms that is $\sqrt{2} \times 6.3 = 8.9$ volts). Hence the capacitors must be rated for at least 9 volts — a little more is usual, say 12 volts, but not too much higher as the rated capacity of some capacitors falls if not worked at near full design voltage.

The diodes must have a peak-inverse rating of *twice* the peak voltage, 18 volts in this case, because at the time when the diode is non-conducting it has the charged capacitor voltage on one side and the full peak reverse voltage from the transformer on the other. In practice modern silicon power diodes have voltage ratings starting from about 50 volts and the EM401 specified is rated at 100 volts — much more than is required.

BUILDING THE UNIT

A printed circuit-board layout for the pre-amplifier is given in Fig. 10 along with the component overlay that shows where each component is placed. Take particular note of the polarities of the diodes, the ICs and the electrolytic capacitors when fitting them to the board.

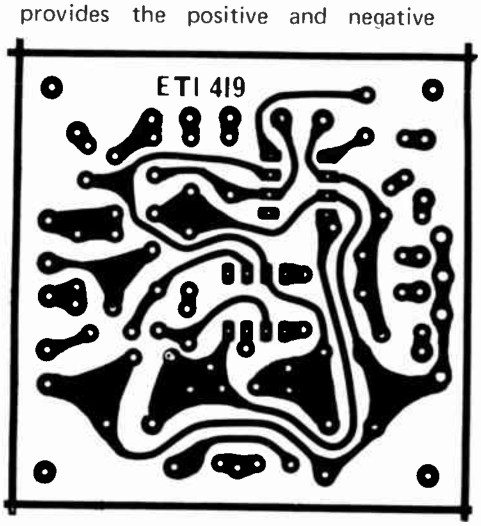
The power supply components (watch the mains connections — they must be made safe) and the board may be conveniently housed in a diecast box or one of the plastic (Clipsal) boxes made for electrical use. Mark each control clearly for ease of operation.

Parts List for mixer/preamplifier

R1	22 k	½ watt	5%
R2	22 k	½ watt	5%
R3	22 k	½ watt	5%
R4	22 k	½ watt	5%
R5	470 k	½ watt	5%
R6	2.7 k	½ watt	5%
R7	12 k	½ watt	5%
R8	1.8 k	½ watt	5%
R9	3.3 k	½ watt	5%
R10	470 k	½ watt	5%
RV1	potentiometer		1 k log.
RV2	"		1 k "
RV3	"		1 k "
RV4	"		1 k "
RV5	"		100 k "
RV6	"		100 k "
RV7	"		250 k "
C1	4.7μF	10V	
C2	4.7μF	10V	

C3	4.7μF 10V
C4	4.7μF 10V
C5	33 pF ceramic
C6	0.0022μF polyester
C7	0.022μF "
C8	0.022μF "
C9	0.22μF "
C10	33 pF ceramic
C11	1μF 25 V tag tantalum
C12	47μF 16 V electro. P.C. mount
C13	47μF 16 V electro. P.C. mount
IC1	LM301A
IC2	LM301A
PC Board	ETI 419
SI	DPDT toggle switch, 400 V, 1 AMP
T1	transformer A + R 6474 or similar
D1 4	EM 401 or similar

Constructional kits for project ETI 419 are suitable.



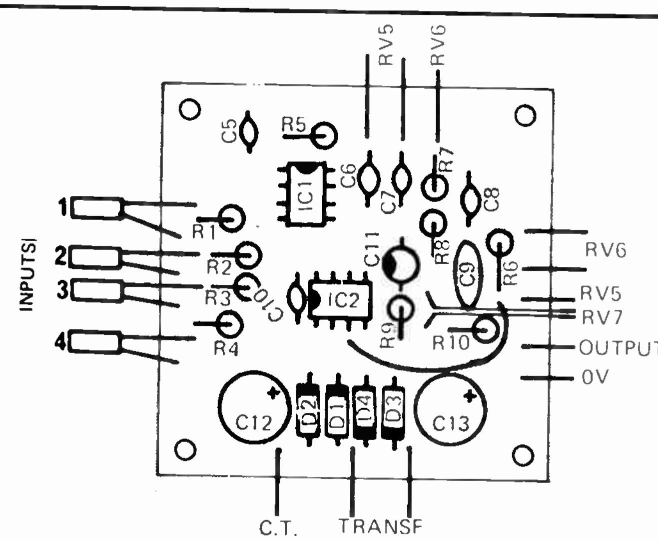


Fig. 10a. Printed circuit board for the mixer preamplifier. (b) Component overlay for the preamplifier incorporating the components for the ac power supply (except the transformer).

15 How regulated power supplies work

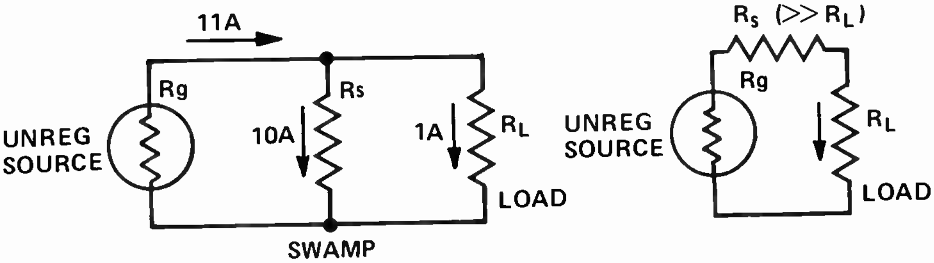


Fig. 1. Regulation can be slightly improved by:
(a). Voltage regulation – an additional shunt load.
(b). Current regulation – an additional series load.

WE HAVE SEEN how the regulation of a power supply is related to its effective internal resistance. A voltage source requires low internal resistance, whilst a current supply requires maximum internal resistance for best operation.

PASSIVE METHODS

A simple and very elementary method of improving the regulation is to impose a dummy load on the supply that is much greater than the normal load. Figure 1a shows a swamping load of 10 amps in parallel with 1 amp. If the real load, (the smaller) varies by a value comparable with 1 amp, far less change occurs in the total load drawn from the supply – the output voltage, therefore, changes less.

This method improves regulation for load changes, but does nothing to guard against input supply changes. Furthermore, it is clearly inefficient. Note that R_L is now connected to a lower source resistance – that of R_g in parallel with R_s. The reduction in resistance is, however, not great.

If a constant current through the load resistance is required a similarly crude method is to place the load in series with a resistor that is much greater than the load value – as shown in Fig. 1b. It can be seen (from Ohms law) that I will now remain constant over a wider range of R_L variation.

Again there is a disadvantage, the input voltage must be raised to drive the same current through the increased resistance circuit. Furthermore, if this is done, R_s wastes considerably more power than is used in the load. As required, for better current regulation, R_L now sees a higher source resistance.

Both circuits are used occasionally but their real relevance is that the same basic principle (modifying the impedance of the supply) is used in more sophisticated supplies. These supplies use special non-linear components and active devices to provide much better regulation with considerably less loss of power.

NON-LINEAR DEVICES

Before low cost semiconductors became widely available in the form of regulating diodes (Zeners) and regulator integrated circuits, designers used the barretter current regulator and the gaseous-tube voltage regulator. These are still found in older equipment but would not normally be used in new designs.

The barretter contains an iron filament in a hydrogen filled envelope and is used in series with the load. Over a reasonably wide range of voltage (100-200 V) the load current remains constant to within 20% (typical value would be 300 mA). By today's design standards they waste power and run extremely hot. Regulation occurs because an increase in current through the filament (see Figure 2a) causes its resistance to increase thus tending to reduce the current to its previous level – R_s, in Fig. 1b, increases with increasing current. Note that the current itself provides a feedback effect (via heating of the filament) that controls the current. The use of feedback (but in a more effective way based upon active elements) is the secret of obtaining really good regulation, as we shall see later.

The gaseous-tube voltage regulator is a gas-filled two-electrode valve which, once the gas is ionised into conduction, provides a reasonably constant voltage drop between its electrodes with varying current values flowing through the ionised gas. Again these are seldom used today, being more suited to voltages much larger than those needed in semi-conductor work (they strike and operate at around 100 V).

In use, the regulator is wired in parallel with the load, the two being fed from the supply via a series resistor, as shown in Fig. 2b. If the input voltage increases (assuming a constant load) the total current must rise. But as the VR tube maintains constant voltage across itself, the load current remains steady and all the excess current flows through the VR tube. Thus the voltage drop across the series resistor increases so that the voltage applied to the load remains the same.

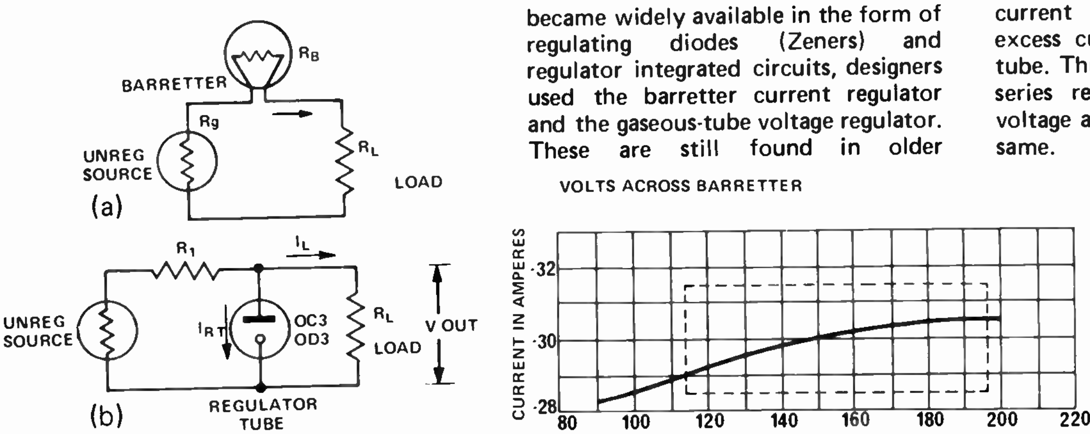
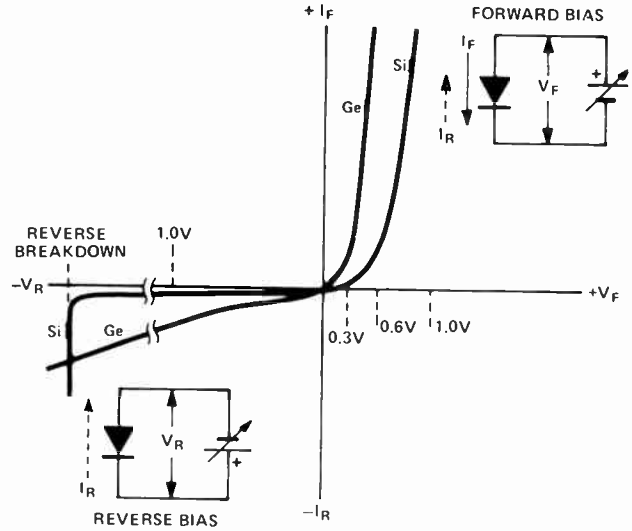


Fig. 2 (a). The barretter tube regulates by the increase in resistance with temperature (and hence current) of an iron-wire filament.
(b). The gaseous-tube voltage regulator operates by virtue of the constant voltage which appears across a gas discharge over a wide range of current.



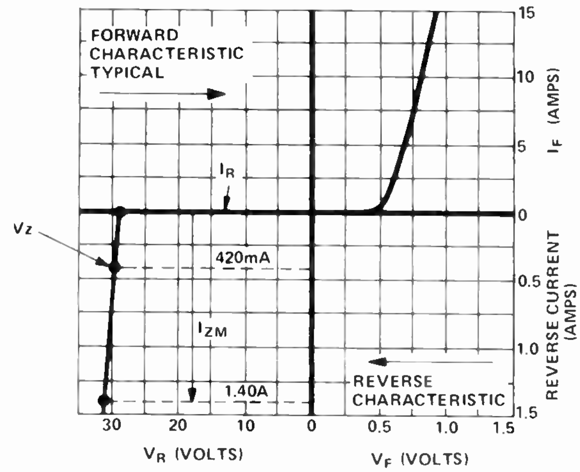


Fig. 3. Forward and reverse bias characteristics of germanium and silicon diodes.

Fig. 4. In a zener diode the reverse-breakdown region is used to provide a constant voltage as current demand varies.

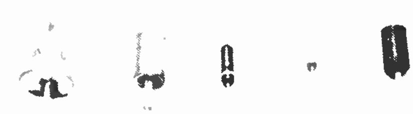


Fig. 5. Zener diodes are manufactured in a wide variety of packages. The larger capacity (not shown here) units are usually mounted on heat sinks if run at their maximum rating.

The effectiveness of the compensation depends upon the rate of change of the voltage-current characteristic of the device. The barretter represented in Fig. 2a changes some 200μ A/V. A flatter curve would imply a current that changes less per volt and this is to be preferred.

Neither of these two devices has a particularly low V/I ratio and neither, therefore, is able to provide close control over wide ranges of input change.

Negative temperature coefficient NTC resistors – more commonly called thermistors or varistors – have a similar voltage-current characteristic but the slope is in the reverse direction, that is, increase in current increases their temperature which, in turn, decreases their resistance. They are not suited to regulator design where constancy is desired but are useful in providing the reverse effect, for example, when wired in series with a load, that could be damaged by switch-on current surges. An NTC resistor suitable for such use might have a resistance of 3000 ohms cold reducing to 200 ohms when heated by 100 mA passing through it.

ZENER DIODES

The current-versus-voltage characteristics of both germanium and silicon diodes are illustrated in Fig. 3. In the forward direction (positive voltage at anode with respect to cathode) the devices operate, as shown, in the right-hand region of the graph.

It can be seen that once the forward voltage across the diodes reaches 650 mV for silicon (or 350 mV for germanium) it remains substantially constant over quite wide excursions of the forward current. Thus forward-biased diodes could be used as constant-voltage regulator elements but only at the low voltages mentioned above or at multiples of these voltages (using diodes in series).

If a conventional diode is reverse biased its operating characteristic will be as shown on the left hand side of Fig. 3. Very little current (microamps) will flow until the reverse voltage reaches a comparatively large value when the current starts to increase much more rapidly. In a germanium device the voltage across the diode will still increase relatively slowly with increasing current, but in a silicon device the voltage across the diode now remains substantially constant regardless of further increase in current. This point is known as the **Reverse Breakdown Point.**

In a normal diode the rapid increase in reverse current causes the semiconductor junction to overheat and the device may fail. This breakdown effect occurs at voltages between two and a half volts and several thousand volts depending upon the material and construction of the semiconductor junction. However, this seeming disadvantage may be put to work in specially constructed devices known as Zener diodes.

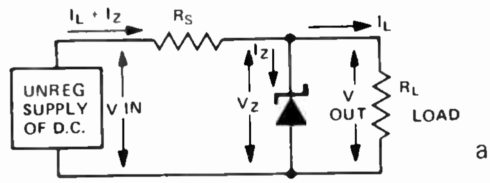
Zener diodes are invariably silicon devices which have been specially designed to operate within the reverse-breakdown region, without damage, provided that the maximum-specified power dissipation (VxI) is not exceeded.

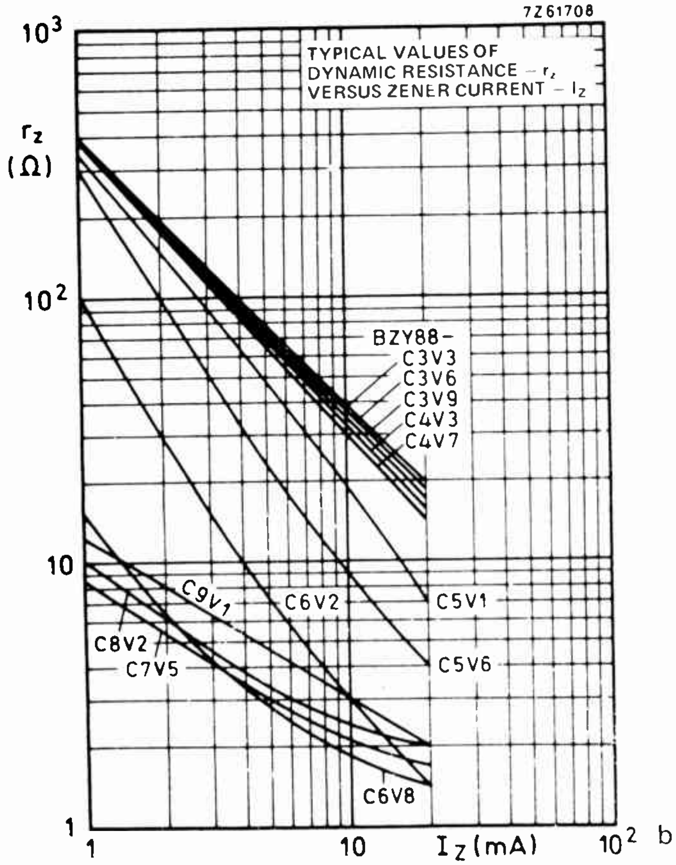
DYNAMIC RESISTANCE OF ZENER DIODES

Ideally, a Zener diode should maintain a constant voltage across itself with varying current through it. However, practical devices don't behave quite like that. In Fig. 4 we see, from the typical characteristics of a 30 volt, 50 watt device, that if the current through the device changes by 1.0 amp, the voltage across it will change by 2 volts. This may be expressed as a resistance as follows:–

By Ohm's Law $\frac{\Delta V}{\Delta I} = R$ i.e. $\frac{2}{1} = 2$ ohms

As this resistance is the ratio of changes of voltage with respect to current it is a dynamic quantity and is therefore known as the **Dynamic**





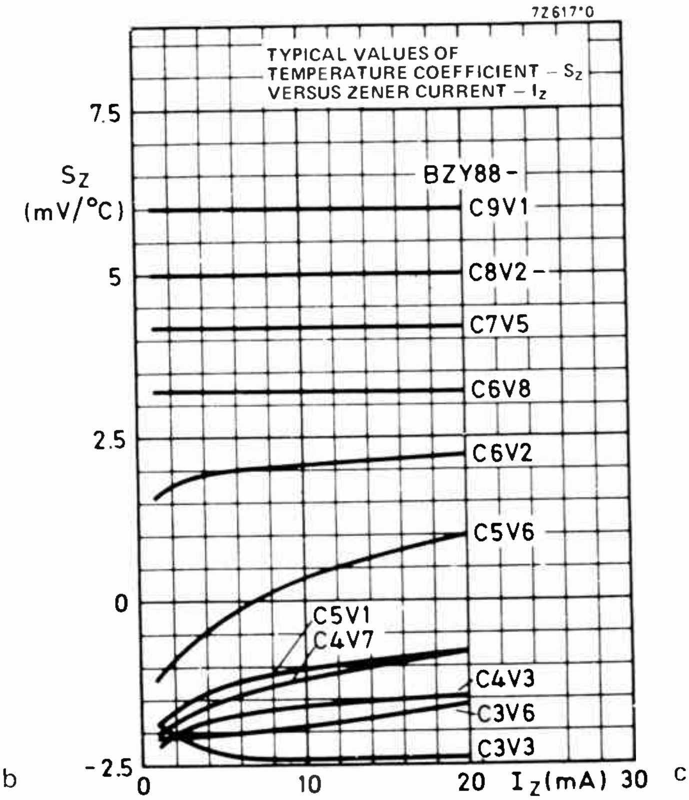


Fig. 6. (a) Typical Zener diode regulator. (b) dynamic resistance of BZY88 series Zeners. (c) temperature coefficients of BZY88 series diodes.

Resistance of the Zener. It tells us how well the Zener will regulate the voltage with changes in load current.

Thus a Zener having the desired reverse-breakdown voltage may be used to replace the gas-regulator valve shown in Fig. 2b. Any load connected across the Zener will see a source impedance which is the parallel combination of the dynamic resistance of the Zener, and the internal impedance of the power supply.

TEMPERATURE COEFFICIENT

The reverse-voltage characteristic of the Zener is temperature dependent, the extent of this dependancy being determined by the designed Zener voltage and power dissipation. For example in typical 400 mW devices the temperature coefficient ranges from -2.5 mV/C° for a 2.7 volt Zener to +26 mV/C° for a 30 volt Zener. Zero temperature coefficient is obtained with a device having a nominal voltage of 5.6.

Where Zener regulators are required to have minimum temperature coefficient and higher than 5.6 volt rating, several diodes with temperature coefficients which cancel may be used in series. For example, if 9.8 volts is required with zero temperature coefficient, a 3.6 volt -2.0 mV/C° diode may be used together with a 6.2 volt +2.0 mV/C° diode.

REGULATOR DESIGN

The circuit of Fig. 6 is that of a typical Zener-diode regulator stage. The series resistor, R_s, must be large enough such that when the load current is at its minimum (Zener current at maximum) the power dissipation rating of the diode is not exceeded, and small enough to ensure that when the load current is maximum (Zener current at minimum) the voltage across the load does not fall below E_z (Zener voltage). Additionally the Zener current should always be at least one tenth of the maximum load current. The optimum value of R_s may be calculated from

$$R_s = \left[\frac{E_{s1} - E_z}{1.1\, I_{L1}}\right]$$

Power in $R_s = (1.1\, I_{L1})^2\, R_s$ and maximum Zener dissipation may be calculated from

$$P_z = \left[\frac{E_{s2} - E_z}{R_s}\right] I_{L2}\; E_z$$

where
- E_{s1} = minimum supply voltage
- E_{s2} = maximum supply voltage
- E_z = zener voltage.
- I_{L1} = maximum load current.
- I_{L2} = minimum load current.

For example assume that we have a car battery supply that varies from 11 to 14 volts and from this we wish to obtain a stabilized 6 volt supply at currents from 40 to 60 mA.

The nearest available Zener voltage is 6.2.

Thus $E_{s1} = 11$ $\quad I_{L1} = 0.06$ A
$E_{s2} = 14$ $\quad I_{L2} = 0.04$ A
$E_z = 6.2$

$$\therefore R_s = \frac{11 - 6.2}{1.1 \times 0.06} = 72.7$$

use nearest preferred value 68 ohms.

The power rating of this resistor must be

$$(1.1 \times 0.06)^2 \times 68$$

= 296 mW — a ½ watt resistor will do.

$$P_z = \left[\frac{14 - 6.2}{68}\right] - 0.04 \quad 6.2$$

$$= 463 \text{ mW}$$

Hence a suitable device from the Philips range would be a BZY96C6V2 which has a power rating of 1.25 watts and a nominal Zener voltage of 6.2.

Where Zeners with power ratings greater than one watt are used a heat sink will usually be necessary. Note also that the supply voltage must *always* be greater than the Zener voltage if regulation is to be maintained — at least 10% higher is a safe minimum value.

The Zener voltage regulator is widely used throughout electronics. It may be used as a basic regulator as in Fig. 6, or, it may be used to provide the reference voltage for more accurate and powerful regulators which make use of active devices as well.

The aim of good voltage supply design is to achieve lowest practical effective internal resistance. The Zener does this reasonably well, for the load sees only the dynamic resistance of the Zener which is much lower than the source impedance. As a rough guide the dynamic impedance R_z of the Zener, varies (according to device) from 30 ohms per volt of the Zener, downward to fractional ohms per volt. If one Zener stage cannot provide enough stability it is quite practicable to join stages in cascade. Each stage thereby lowers the effective source resistance (because stages are connected in parallel) but, more significant is the fact that input voltage variations are more adequately attenuated, each stage running from a progressively better stabilised source. Fig. 7a shows a typical dual stage supply. Figure 7b is the preferred method, of providing the same illustrated 5.6 V, for in this alternative all of the diodes have optimum temperature stability.

Zeners also have other uses — to clip and hold voltages at fixed levels and to convert sine-wave signals to square-wave signals.

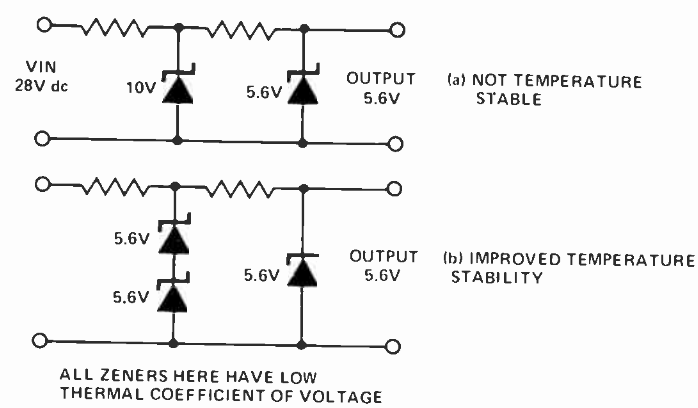


Fig. 7. (a) Cascading Zener regulator stages improves voltage stability. (b) Selected Zeners are often joined in series to achieve the lowest temperature error.

ACTIVE COMPENSATION

Although the Zener can provide a relatively low dynamic resistance value, it is possible to provide still lower resistance by incorporating active amplifiers into the regulator circuitry.

If the actual load voltage is compared with a constant reference-voltage source, it is possible to determine if the output is greater or smaller than required and by how much. Having made such a comparison, the difference, called the error signal, can be used to modify the incoming signal accordingly. This is the principle of feedback. Figure 8 shows how feedback is used in the electro-mechanical type of supply regulator. These regulators are used where loads are high and the unwanted changes occur only relatively slowly.

In operation the output of a basic rectifier is smoothed by a capacitor (C) to provide the required output voltage. The output voltage is compared with a reference voltage and the difference between them (that is the error) is amplified. The amplifier output drives a motor such that a tapping on the ac transformer is changed — thus reducing the error.

Thus, by using feedback, changes in output voltage are rapidly sensed and the input quickly compensated. The feedback amplifier and control actuator (the motor in Fig. 8) need not be precision devices — they can be quite crude in fact — but the reference voltage must have better stability and accuracy than is required from the output.

The reference voltage is quite often, and effectively, supplied by a Zener. As the Zener now only has to supply a reference voltage, and not operate over a wide range of current, its operation will be much more stable. That is, its dynamic impedance will not be a source of error. Additionally, the Zener current may be set at a level which gives optimum temperature stability.

Although electro-mechanical

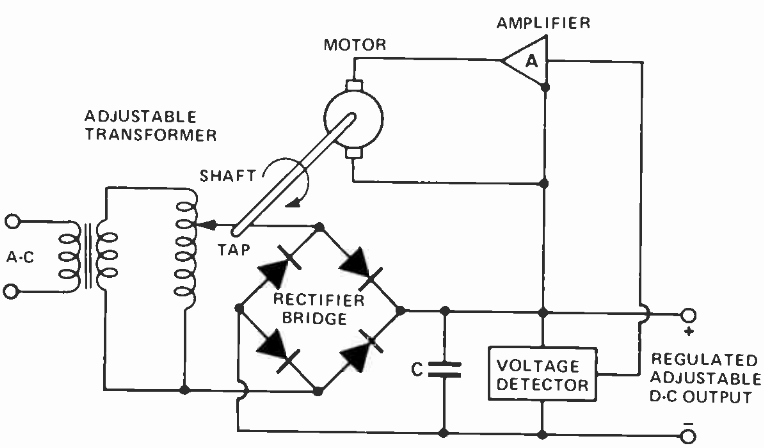


Fig. 8. Superior regulation is obtained by using feedback — as demonstrated by this electromechanical form of regulator.

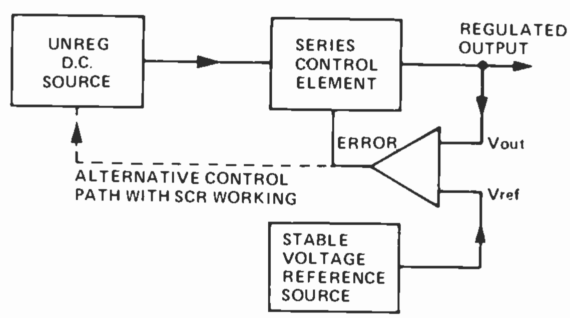


Fig. 9. Generalized diagram of active feedback type regulator.

regulators have their uses, the majority of regulators for low-power electronic systems now use totally solid-state components to build systems such as that shown in Fig. 9.

A wide range of control methods are used using such devices as transistors and integrated circuits, silicon controlled rectifiers (SCRs) and saturable reactors to achieve fast and accurate regulation.

The voltage reference is again generally derived from a Zener network. However, precision units may use a Weston standard cell or a special high-stability Zener arrangement. Yet another kind of regulator may use an external varying voltage as the reference in the feedback system. With such supplies the output is made to track the varying input voltage – these are called programmable supplies.

SHUNT REGULATORS

We have seen how the basic Zener arrangement may be used to provide a shunt path around the load thus stabilizing load voltage. By reducing the current range required of the Zener diode it is possible to improve the regulation and to reduce the power handling required of the Zener. Figure 10 illustrates how a transistor is added to the basic Zener circuit to produce a more precise shunt regulator.

Now the Zener regulator only has to supply the base current of the transistor which in turn controls the much larger collector current.

To see how the regulator works let us assume that the current demanded by the load falls. The voltage across the load would tend to rise (due to less voltage drop across the series resistor R) and this would cause the base-emitter voltage on the transistor to rise (as V_R is held constant by the Zener). Hence the transistor draws more current to compensate for the current shed by the load. That is the current drawn through the series resistor R is held constant, current not needed by the load being shunted by the transistor.

As the transistor provides current gain ranging from tens to hundreds, the current variations demanded from the Zener are reduced by the same factor, with consequent improvement in regulation. Although not immediately obvious, feedback is used in this circuit. Voltage changes across the load appear at the base of the transistor which acts to reduce the original change to zero.

One vitally needed characteristic of a general-purpose power supply is that the output should be capable of being short-circuited without damaging any components. The shunt regulator does just this – a shorted output merely connects the emitter of the transistor to collector, thereby reducing the voltage applied to the device. The transistor therefore cannot be damaged by a shorted output. Such a supply is, however, inefficient, especially at light loads, for shunt regulators act always to dissipate the same maximum amount of power – either in the load, in the shunt element or in both. Hence, at zero load the unit wastes as much power as the maximum safe load would consume, and if ever the load is disconnected the transistor must be capable of passing the full load current.

SERIES REGULATORS

The Zener reference may be used to control a transistor in series with the load such that a constant voltage appears across the load. Figure 11 is the circuit of such a basic series regulator.

The operation of a series regulator may most easily be understood by considering the transistor and load to be an emitter follower circuit. We know from our previous theory that an emitter follower maintains its emitter at V_{be} (0.6 volts for silicon) less than the voltage at its base, regardless of the value of the collector supply. Thus the transistor, because of the Zener reference voltage at its base, varies its impedance and hence the voltage dropped across itself, in order to maintain a constant voltage across the load, regardless of load current and supply-voltage variations.

As the transistor has a large current-gain the Zener diode again only has to supply a small current range and regulation is therefore improved. However, the transistor must be capable of withstanding the full load current and of dissipating fairly high power. The series regulator is more efficient on light loads than the shunt regulator but if the output is short-circuited the transistor in a series supply will be destroyed (unless protected in some way), as the full supply voltage and base drive is applied to it.

IMPROVING SERIES REGULATORS

The simple series regulator, just described, is a great improvement on the simple Zener regulator but may still be improved further by additional circuit refinements.

Figure 12 show the schematic diagram of a typical series regulator supply. The ac transformer has two secondaries the first of which provides dc to the regulator and hence the load via a bridge rectifier and smoothing capacitor. The second winding provides a separate supply to a Zener regulator. As this winding does not have to supply the varying load current – only the steady Zener supply, regulation of the reference

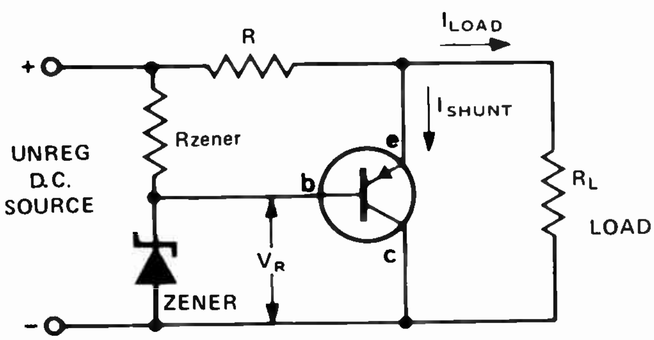


Fig. 10. Basic shunt regulator uses transistor to maintain constant load current on the unregulated dc source.

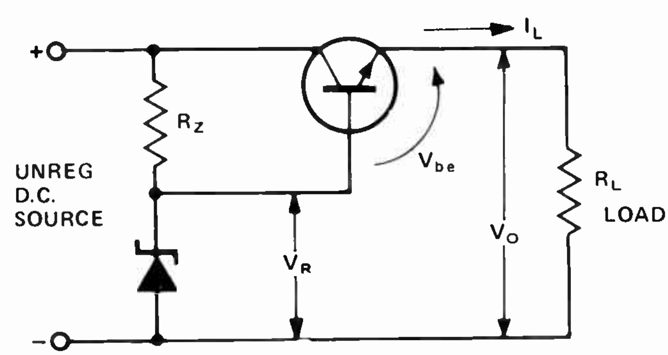


Fig. 11. The series regulator also relies on feedback to control voltage drop across a series-pass transistor.

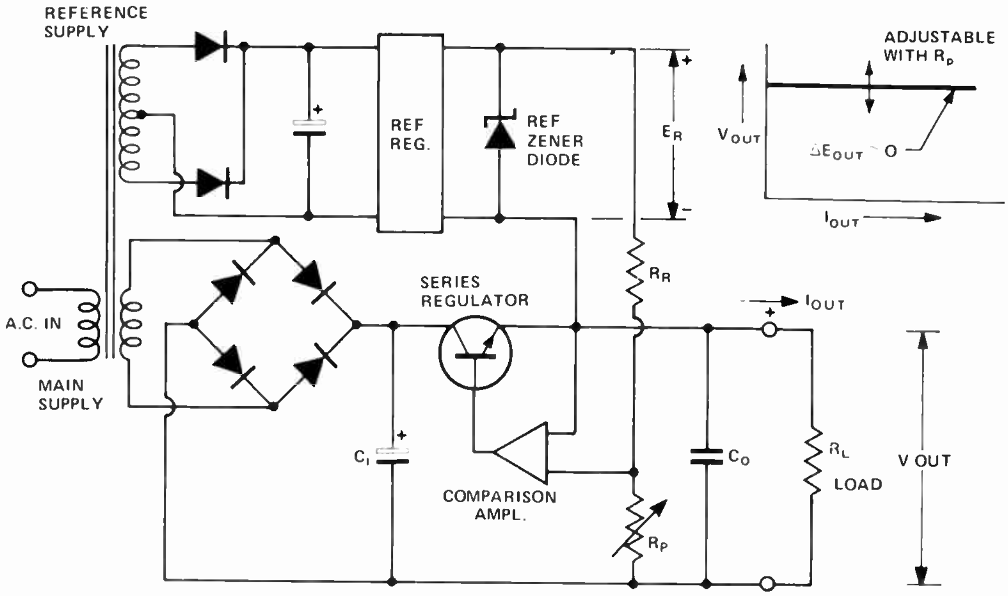


Fig. 12. This schematic diagram of a Hewlett-Packard series – regulated constant voltage power-supply illustrates the design philosophy of precision supplies.

Zener is considerably improved. In addition a temperature compensated Zener may be used which could have a temperature coefficient as good as 0.01%/C°.

This very-stable Zener reference is compared to the output voltage of the supply by a differential-operational amplifier. Thus the Zener does not have to supply any appreciable current. This results in still further improvement in regulation. The operational amplifier provides a change in base current to the series regulator in such a direction as to correct any error between the output voltage and the Zener reference.

CONSTANT CURRENT

In some applications – magnetic circuits, focussing coils, semiconductor testing – the requirement is for constant output current regardless of load changes. Loads connected to such supplies are connected in series, rather than in parallel as is the case for voltage regulated supplies.

Ohms Law tells us that a certain value of current is related to voltage via a fixed value of resistance. Hence constant current supplies can make use of a small series resistor to monitor the output current by virtue of the voltage developed across the resistor. This voltage is then compared with a reference (the actual value is a matter of design choice, the lower the series resistance the lower the voltages and losses involved) in much the same way as for a stabilised voltage supply. The differences in circuitry needed can be seen by comparing Fig. 13 – that of a well-designed constant current unit – with Fig. 12.

By combining these two concepts into one supply a combined constant voltage and constant current unit is formed (denoted CV/CC). This holds a constant voltage up to a preset maximum load current whereafter it provides constant current.

Power supplies designed to provide variable output for experimental purposes, or equipment (or component) testing are often subject to severe or extended overloads. Units such as these are generally equipped with various protective devices that safeguard not only the power supplies themselves but also the loads that they are driving. These various protective circuits will be described in the next part of the series.

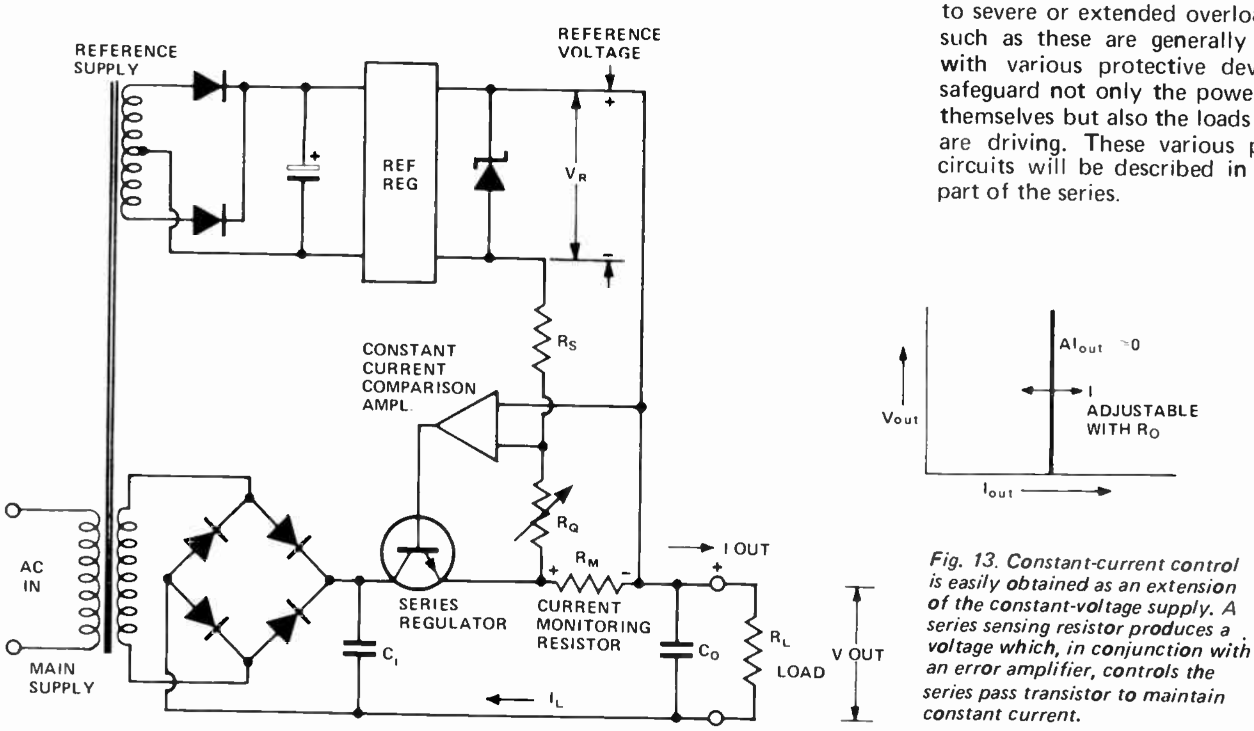


Fig. 13. Constant-current control is easily obtained as an extension of the constant-voltage supply. A series sensing resistor produces a voltage which, in conjunction with an error amplifier, controls the series pass transistor to maintain constant current.

16 General purpose supplies

SUPPLIES built permanently into equipment as fixed parameter units usually operate within reasonably well-defined operating limits. Fault conditions are, therefore, less likely to occur on the supply output, but are still a reality.

By contrast, general-purpose supplies designed to provide variable output for device testing, circuit development and multi-purpose use are prone to a number of fault conditions which could destroy components.

With the reduction in cost of power supply circuits, it is now practicable (and wise) to always employ protective devices that sense external or internal fault conditions and apply protective measures to the supply which prevent damage to both the supply and, possibly, the load being driven.

A wide variety of faults can occur. The supply itself may be damaged by excessive input overvoltage which may occur either as a steady-state overload or as a brief excitation transient. Protective measures include using a simple wire-fuse or magnetically-operated circuit breakers that can break the circuit with greater reliability and speed than fuses. The same transients may destroy the bridge-rectifier diodes, these cannot effectively be protected by fuses or breakers. One technique to overcome this is to use diodes that are much heavier than really needed; another is to supplement these with an RC network across the output of the transformer – to provide a reduced impedance path for voltage transients.

On the output of the supply, protection is needed to prevent too low a load impedance drawing excessive current. We know that a constant current supply uses a series monitoring resistor to produce a control voltage. The same method can be used to limit output current. A simple method often used in series regulators is illustrated in Fig. 1. The diodes D1 and D2 do not conduct until the voltage drop across the small series resistor reaches the forward voltage of the diodes (0.4 for germanium, 0.7 for silicon). If the output of the supply is short circuited the diodes will conduct and limit the output current to a value

$$I = \frac{V_{D1} + V_{D2}}{R_s}$$

Thus if germanium diodes were used and R_s was one ohm the supply would limit at

$$\frac{0.4 + 0.4}{1} = 800 \text{ mA}$$

Another serious condition that can occur is for the supply to produce (if only momentarily) *overvoltage* at the output. Protection against this is essential, for excessive voltage can destroy semiconductor loads connected to the supply. The technique used is very rapidly to sense when overvoltage occurs and immediately connect a very low resistance shunt across the output terminals. For obvious reasons this device is known as "crowbar" protection!

A well-designed "crowbar" takes into account operating times (microseconds is quite feasible), recovery time after triggering; triggering sensitivity and many other features.

The simplest crowbar arrangement is to place a suitably rated (and voltage value) Zener diode across the output, as depicted in Fig. 2. If the voltage exceeds the Zener voltage the Zener conducts, clamping the output to a point just above the normal maximum voltage available from the supply. Excessive current is controlled by the current limiting protective arrangement presumably built into the supply.

More sophisticated methods compare the output voltage to a reference source and use any sudden difference to trigger (see Fig. 3) a silicon-controlled-rectifier which shorts the output thus either blowing the supply fuse or putting the supply into current limit.

TRACKING SUPPLIES

Many circuits, IC op-amps for instance, require dual voltages – that is, positive and negative values referred to a common zero voltage. Some circuits require that both supplies provide exactly the same value of voltage, regardless of differences in load currents or fault conditions which may affect one output only. Another

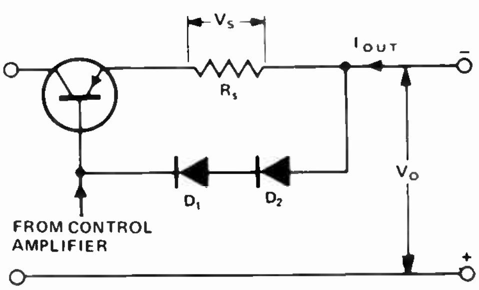


Fig. 1. Automatic current limiting is obtained with a series sensing resistor. This method of control is often used in series-pass voltage regulators.

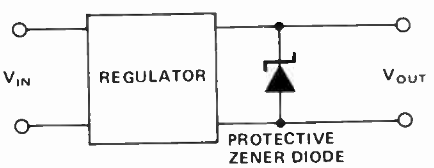


Fig. 2. A Zener diode across the output provides "crowbar" protection against overvoltage.

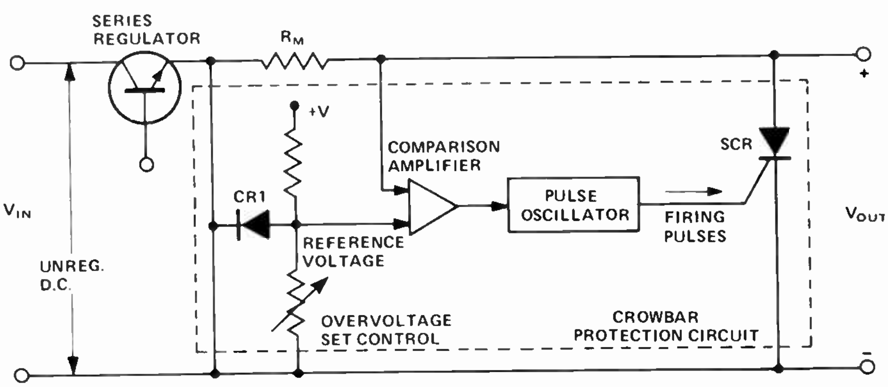


Fig. 3. Active control can provide more effective crowbar protection. This circuit uses a silicon-controlled rectifier to short the output should overvoltage occur. The supply of course must be able to withstand such a short.

need for uniform supply outputs is where a number of slave supplies are required to follow a master unit. Supplies that have this inbuilt facility to follow external voltage are said to possess auto-tracking capability.

Auto-tracking is provided by comparing the two (or more) outputs using any (negative or positive) resultant error signal to control the regulator of one supply. One arrangement is shown in Fig.4.

Auto-series operation is also available in some proprietary units. This enables a number of units to be connected in series in order to provide increased voltage. Sensing circuitry ensures that the voltage is shared evenly across each unit.

HEAVY-DUTY SUPPLIES

The series pass regulator transistor is capable of medium-demand currents. Several transistors may be paralleled to increase the total capacity. However, the method becomes wasteful at high power levels as considerable power must be dissipated in the series pass elements.

A common way to provide greater power is to use the mechanically-driven variable transformer arrangement shown in Figure 8, Chapter 15.

Another method uses special transformer designs that provide a reasonably wide degree of self-regulation, by varying magnetic leakage between primary and secondary windings or, by resonating the transformer windings with a tuning capacitor (transformer core saturates at a constant level).

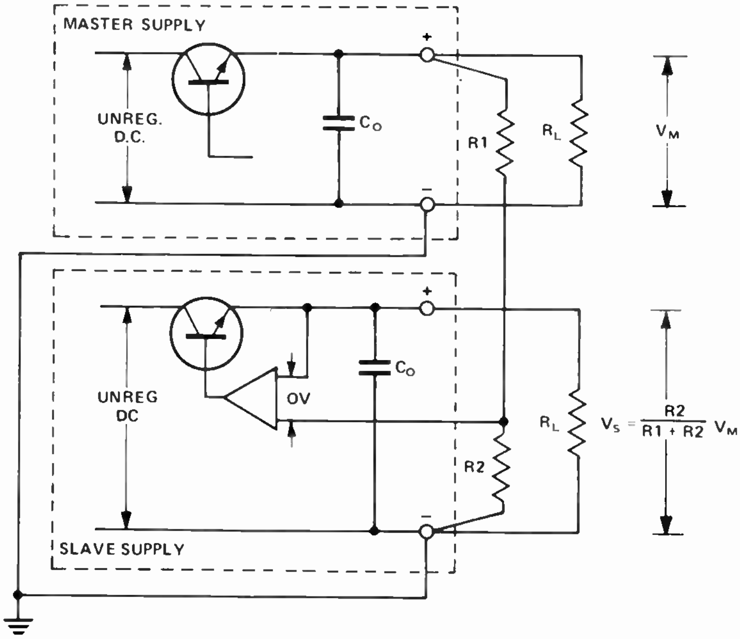
If a switch were to be incorporated, instead of the series-pass transistor, it could be operated with an on/off ratio such that the average power allowed to pass is controlled. This chopped waveform may then be filtered to provide smooth dc. Switching regulation, as this kind of operation is known, is one of the most efficient forms of regulator design because very little power is wasted. The output voltage is compared with a reference value, as before, and the resulting error signal is converted into an equivalent variable-rate, on-off digital signal. This, in turn, is used to control the on-off ratio of the series transistor switch. Where power loss must be minimized, switching regulators are essential.

SILICON CONTROLLED RECTIFIERS

Another kind of switching regulator uses the silicon-controlled rectifier SCR diode (see Fig. 5).

Silicon controlled rectifiers, unlike ordinary silicon diodes, have four semiconductor layers and three terminals (anode, cathode and gate).

Fig. 4. Auto-tracking of two supplies can be achieved by using feedback to sense the error between them and applying correction to one of them.



Like a normal diode the SCR will conduct when the anode is positive with respect to the cathode. But, unless the gate is also positive, the SCR will not conduct at all! The SCR may be switched on at any point in the positive cycle by a positive voltage on its gate. Once the SCR is switched into conduction the gate loses control until the anode-to-cathode potential falls to zero. Thus in operation a single positive pulse will switch the SCR on at any desired time within the positive half cycle and, by varying the time at which this pulse occurs we may control the average power passed by the SCR.

To pass both positive and negative

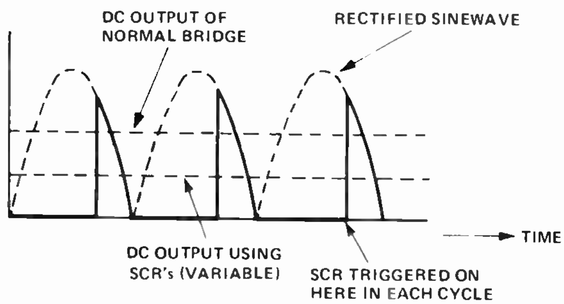
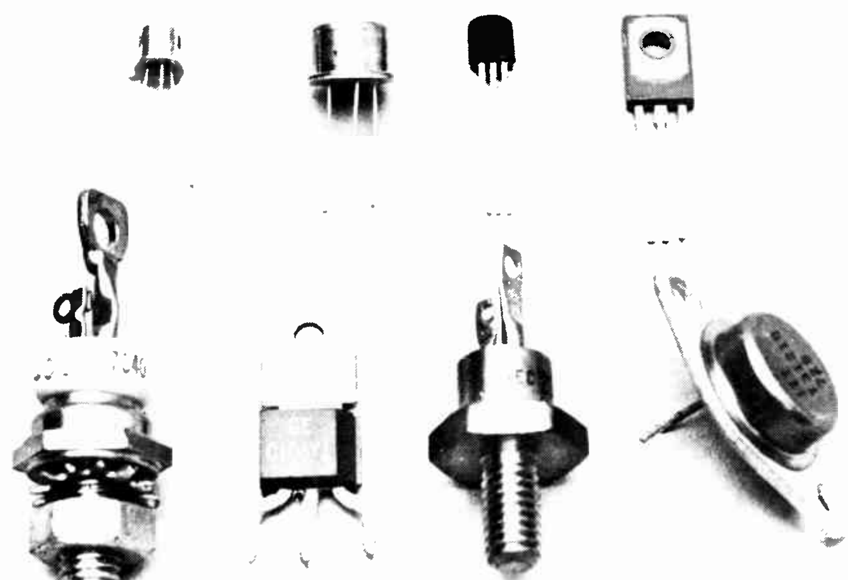


Fig.5. A range of silicon-controlled rectifier SCR's. The waveforms show how only a portion of each half cycle of the rectified sine-wave is switched through thus reducing the effective output voltage.



ELECTRONICS -it's easy!

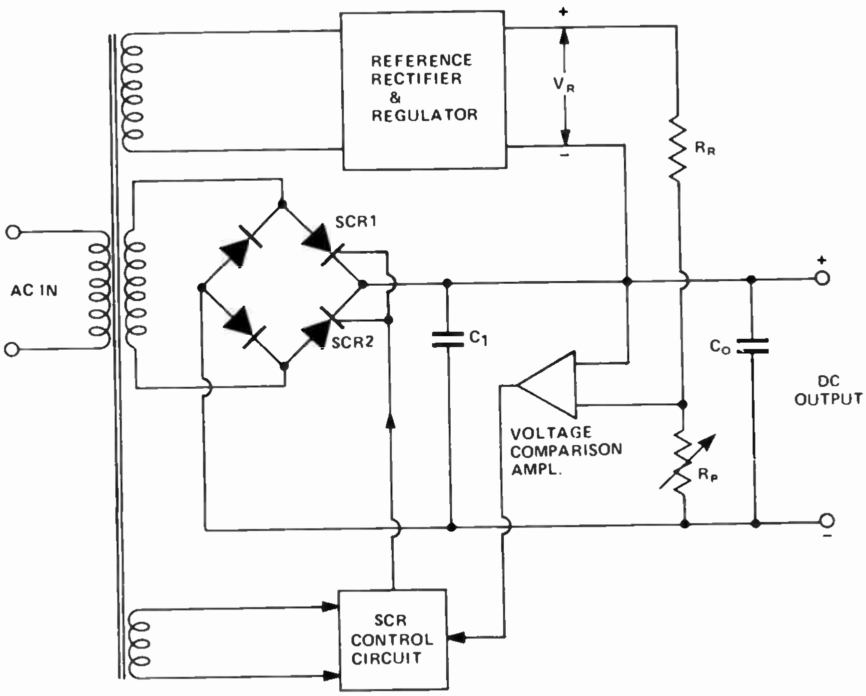


Fig.6. Use of SCR's in the power supply rectifier-bridge stage. The SCR control circuit output decides the average output of the bridge.

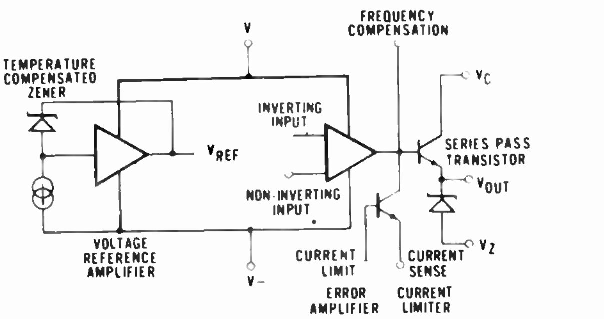
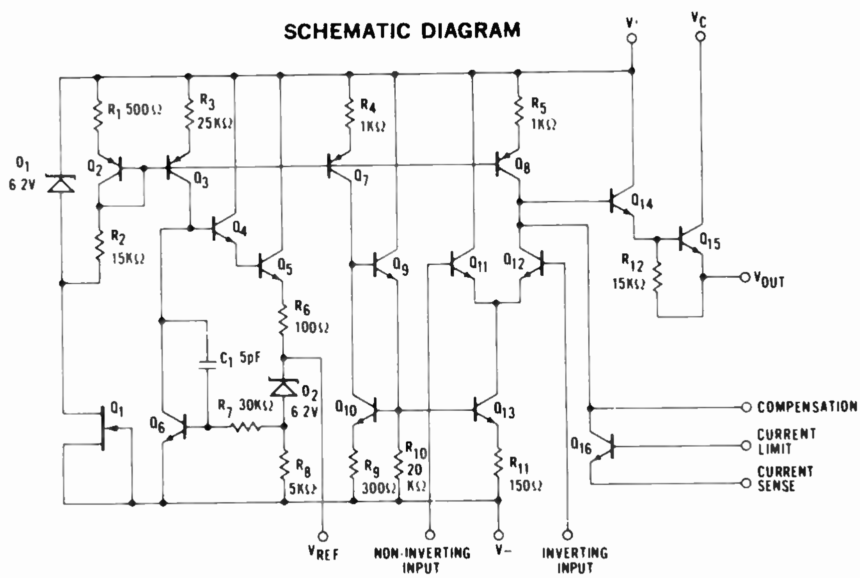


Fig.7. Circuit diagram and system schematic of uA723 I.C. regulator chip. An external series-pass power transistor stage is needed to complete the regulator.

SCHEMATIC DIAGRAM



half cycles we must use two SCRs connected appropriately or use a special device — called a TRIAC — which can be switched on for either polarity.

The schematic of a regulated supply using SCR power control is shown in Fig. 6. The power handling limits of SCR devices range to thousands of amps. As their main feature is control of power by switching, not by dissipation of unused energy, they do not need the same degree of cooling for a given load as would the series-pass transistor method.

INTEGRATED CIRCUIT REGULATORS

In recent years, special purpose IC components have become available that include a reference voltage supply, a comparator and a drive circuit for controlling an external series-pass regulator transistor all in one small device.

One such chip is the μA723 shown schematically in Fig. 7. From the internal circuit diagram it is clear that these units are capable of providing excellent regulation. The output voltage is adjustable on demand by altering the proportion of the output voltage which is compared to the reference voltage by an error amplifier. Ancillary built-in circuitry provides current limiting and crowbar action if needed.

Also available are regulator IC's which have the series-pass power stage formed on single silicon chips. The LM109 and LM309 are such regulators (circuit shown in Fig. 8) and provide 5 V with output load in excess of 1 ampere. The LM309 has internal thermal overload protection, internal current limiting and is virtually blow-out proof. As is shown in Fig. 8, nothing could be simpler to use if a fixed voltage is needed. The cost is a mere three dollars — thereby, powerfully demonstrating that today's electronic discipline is a matter of system rather than component design.

REFERENCES

References given in the article before last are most relevant to the design of regulators. On the more specific aspects of design and use, the following are worth considering for purchase:

"Zener diode handbook" Motorola, 1967.

"Thyristor projects using SCRs and Triacs" R.M. Marston — Butterworths, 1974.

"Silicon rectifier handbook" Motorola, 1966.

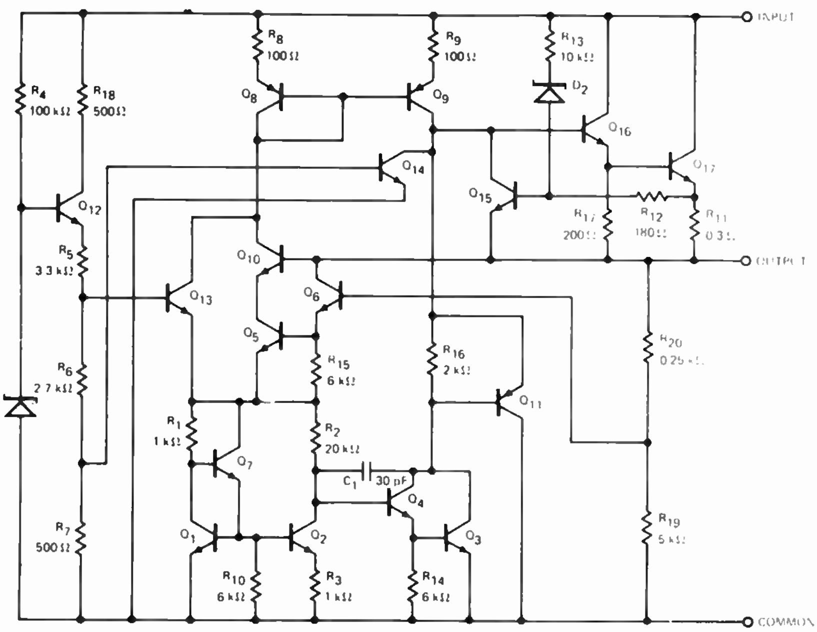
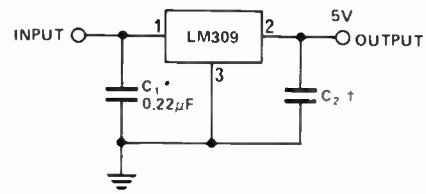


Fig. 8. The LM309 IC regulator has the series-pass transistor (Q17) formed on the same chip. The circuitry is most sophisticated. Different applications are shown schematically.

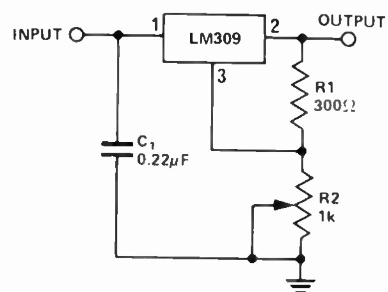
FIXED 5V REGULATOR



* REQUIRED IF REGULATOR IS LOCATED AN APPRECIABLE DISTANCE FROM POWER SUPPLY FILTER.

† ALTHOUGH NO OUTPUT CAPACITOR IS NEEDED FOR STABILITY, IT DOES IMPROVE TRANSIENT RESPONSE.

ADJUSTABLE OUTPUT REGULATOR



ELECTRONICS – in practice

Simple regulated supply provides 1.5 to 15 volts at up to one ampere.

Earlier in this book details were given of an unregulated power supply that provides a dc output varying from 18V at no load, dropping to 10V at maximum load.

This unit can be extended by the addition of a series-pass regulator that employs an IC regulator chip and a power transistor. It incorporates current limiting and the output can be preset to provide any voltage between 1.5 and 15V with a load current as high as one amp.

The circuit diagram for the complete regulated supply is shown in Fig. 9. By studying this, in conjunction with Fig. 8, it can be seen that adjusting RV1 provides the comparator error amplifier with a reference voltage up to the full 7.15V produced by the built-in Zener reference circuit. Resistors R_2 and R_3 provide a divider chain that is tapped to enable the actual output voltage to be attenuated by a factor of 2.2. Thus the output is controlled so that it is 2.2 times the reference voltage provided by RV1.

The capacitor connected between pin 9 and pin 2 provides frequency compensation; improved feedback control performance is obtained by appropriate selection. Resistor R_{SC}, in series with the output, is a current-sensing resistance. Its value, and hence the voltage developed across it, (at pin 1) determines the current limit point. Pin 10 clamps the emitter of Q1 if V_{Rsc} exceeds 0.6V. The maximum output voltage and current obtained from the unit is a function of components used. Using a 15V centre-tapped transformer with 1A capacity, it provides 1A at 10V and 0.5A at 15V. Output voltage can be changed by altering the ratio of R_2 and R_3 with smooth manual control being obtained with RV1. Maximum output must not exceed 25 volts.

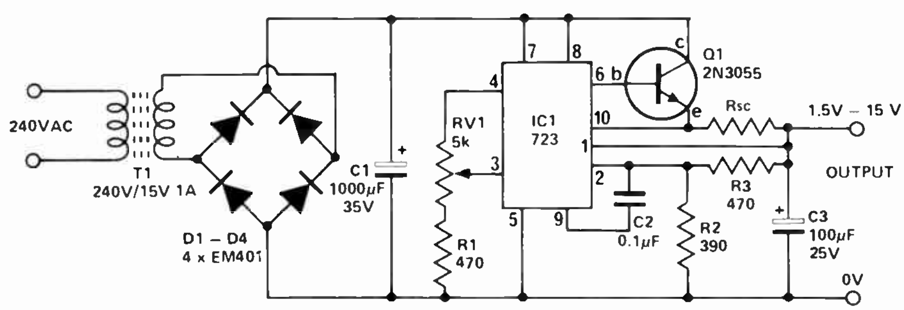


Fig.9. This regulated supply, based on a μA723 IC, provides an output adjustable between 1.5 and 15 volts.

Handy battery eliminator provides 4.5, six or nine volt output from mains.

Many battery operated portable appliances are provided with a socket to enable them to be connected to a suitable external dc power supply.

This external power supply, or 'battery saver', may be used to energize many different types of tape recorders, record players, transistor radios, etc. The unit is mains operated and is intended for use in the home.

It is very simple to construct, provides adequate regulation and has sufficient power handling capacity to operate practically any small domestic, normally battery operated, appliance.

As the majority of battery operated appliances use a nine volt supply, the unit described here has been designed for a nominal nine volt output. However for some purposes a six volt or a four and a half volt output may be required. This may be readily achieved by replacing the components ZD1 and R1 as follows:-

OUTPUT VOLTS	ZD1	R1 OHMS
8	BZY88C10	330
6	BZY88C6V8	680
4.5	BZY88C5V1	680

The completed unit when finished, should be mounted in a suitable box that is properly earthed to the mains.

The unit has been designed so that it will not be damaged if the output is accidentally short circuited. Nevertheless a continual short circuit must not be applied as this will cause excessive heat to be generated within the ten ohm resistor.

If the appliance already has a socket for an external power supply this will almost certainly be of a type in which the plug cannot be accidentally shorted. If no socket is fitted then an external power supply socket should be installed. Standard plug/sockets for this purpose are readily available from most parts suppliers, but note that plugs/sockets intended for nine volt use are not interchangeable with those intended for six volt use — the centre pins are of different diameters.

The socket should be of the type which has a contact for disconnecting the internal battery when the power supply is plugged in.

The 2N 3055 'series pass' transistor is much larger than required. We have specified this device as it provides very good overlead capability and is readily available at low cost.

Q1 is a 'series pass' transistor and drops the supply voltage to the required regulated output voltage. The output of the transistor is controlled by the Zener diode ZD1. Resistor R1 supplies current for the correct operation of ZD1 and also provides base current for Q1.

The ten ohm series resistor prevents damage to the transistor if the output of the unit is accidentally short circuited.

PARTS LIST (ETI 511B)

Mains operated battery eliminator

1 — transformer, 240 volt to 12.6 volts (100 mA minimum)
1 — transistor, 2N 3055
1 — zener diode, BZY88C10
4 — diodes, EM 401
1 — resistor, 10 ohm, 5 Watt, 10%
1 — resistor, 330 ohm ½ Watt, 10%
1 — capacitor, 10μF, 25 volt electrolytic
1 — capacitor, 220μF, 25 volt, electrolytic
Sundries: plug, tag strips, cable, solder lug etc.
Components quoted above are for nine volt output. See text for alternative output voltages.

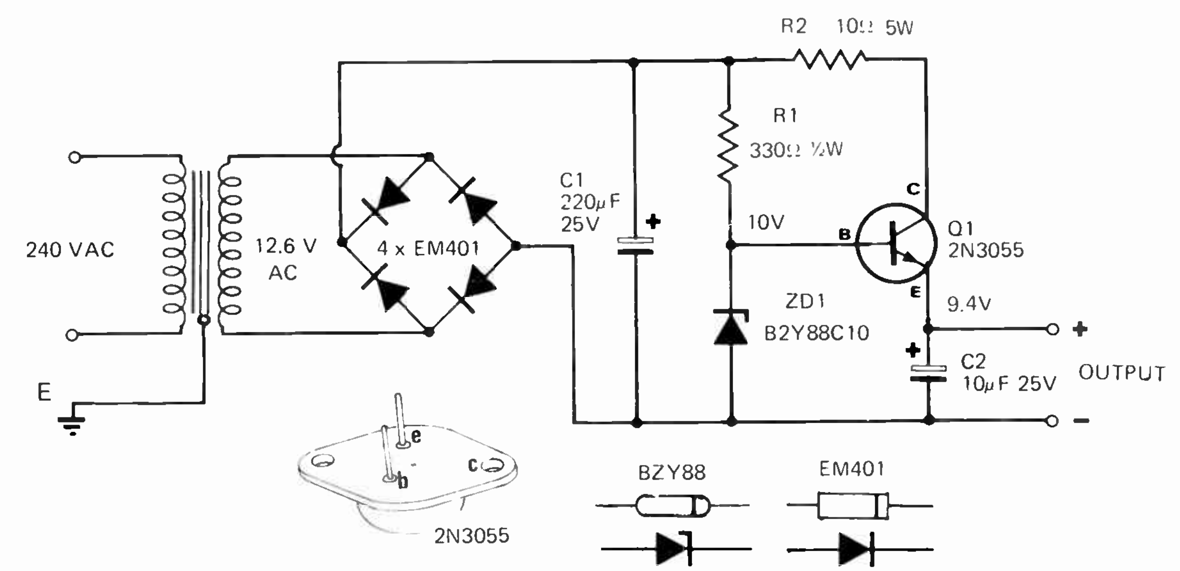


Fig. 10. Circuit diagram of mains operated unit.

17 Generating signal waveforms

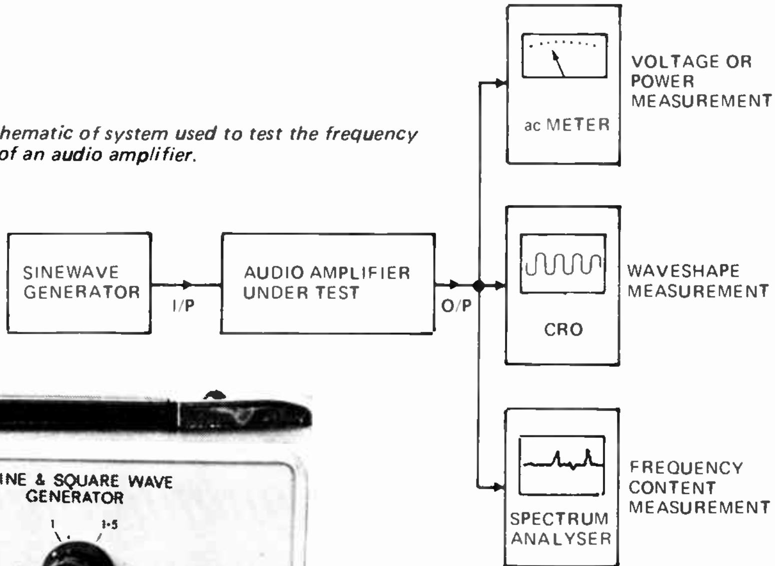


Fig. 1. Schematic of system used to test the frequency response of an audio amplifier.

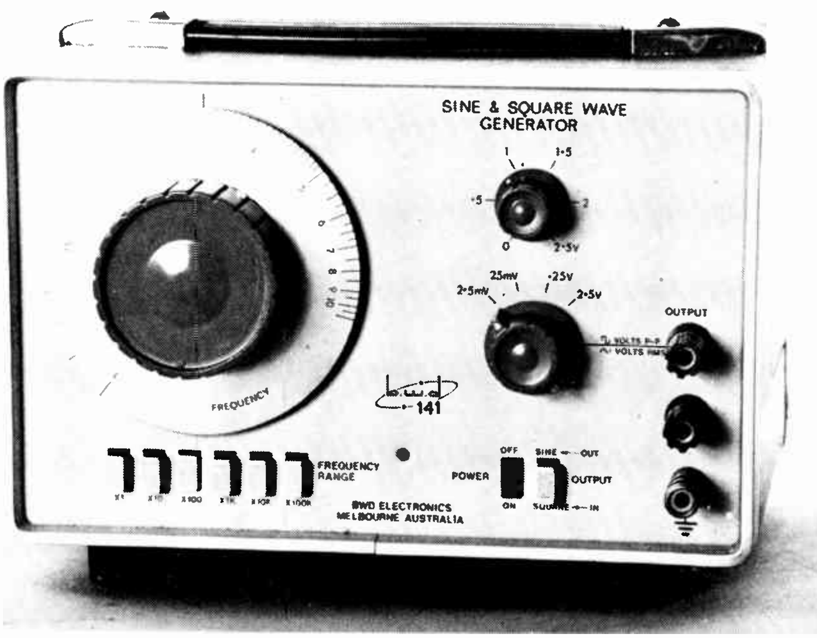


Fig. 2. This audio frequency generator, from BWD, is a medium cost unit that provides adjustable amplitude sine and square waves at frequencies from 1 Hz to 1 MHz.

THROUGHOUT electronic systems, repetitive waveforms of specific shape are commonly processed in order to achieve the desired function. Typically used are – sine waves, square waves, triangular and sawtooth waveforms as well as many types of pulse train.

For instance one may need a sinewave signal generator to determine the response at various frequencies of an audio amplifier. This is done by feeding a sinewave into the input (as shown in Fig. 1) whilst monitoring the amplitude and wave shape of the output. By such means it is possible to establish an amplifier's gain and distortion over its working frequency range. A suitable signal source for this task, such as that shown in Fig. 2, would need adjustable frequency range over the audio band (10 Hz – 20 kHz), adjustable amplitude, and above all an adequately pure sinusoidal wave shape. Any distortion in the test sine-wave would of course affect measurements.

Similarly radio and television receivers are tested by injecting known sine-wave frequencies, only here, the frequencies used are much higher.

In the construction of a radio transmitter, or a multiplexed signal system, a basic carrier is required upon which the modulation is impressed. This technique was described in Part 5 (Vol.1.) of this series. Here however, the requirement is for fixed frequencies; there may not be the same need for waveform purity as in the testing of a superior audio amplifier.

An excellent example of the need for signal generation is in electronic musical instruments such as the electronic organ or music synthesizer.

Both types of instruments require numerous waveforms to be generated over a wide range of frequencies. These waveforms are then mixed and modified in amplitude and time to create an enormously wide variety of sounds. Virtually any sound can be created – from surf breaking on the rocks to conventional musical instruments.

In instrumentation a measuring bridge similar to that shown in Fig. 4 is often used. This particular configuration, known as a Wheatstone bridge works as follows.

If, when an excitation voltage is applied to the bridge, R1 equals R2

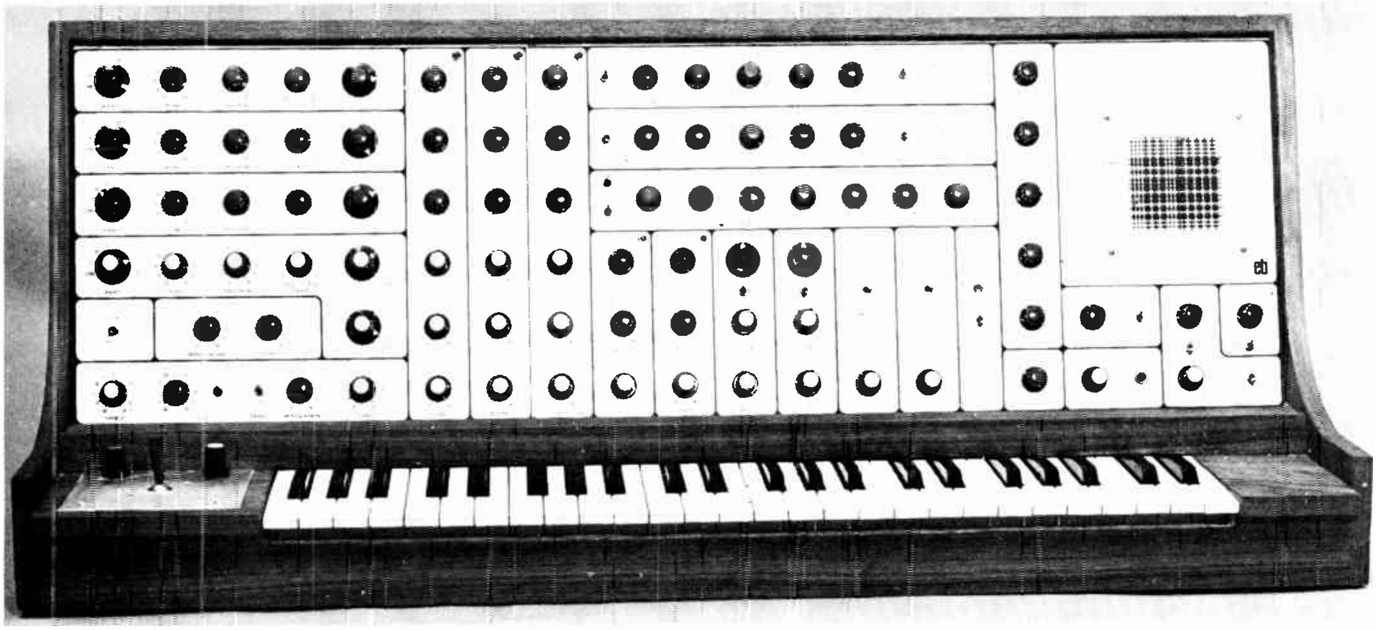


Fig. 3. Electronic music synthesizers, such as the ETI 4600 shown here, create musical sounds from signal generators and waveform shaping modules — not from musical instruments.

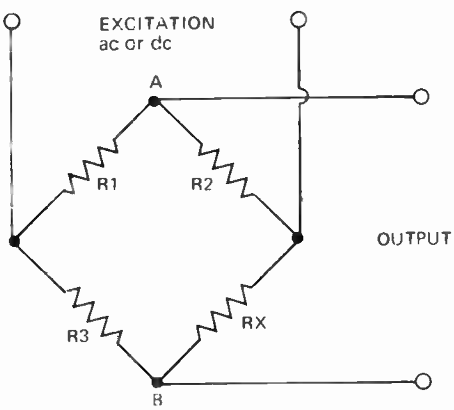


Fig. 4. The Wheatstone bridge is commonly used in instrumentation for measurements with various types of sensor. The bridge is usually excited by an ac source so that the output may be more readily amplified.

Fig. 5. Three common, fixed-rate implantable pacemakers.

and R3 equals RX there will be equal voltages at points A and B, and hence, there will be no output.

Resistor RX is usually a sensor, such as an LDR, thermistor etc which changes its resistance proportionately to the quantity (light, temperature pressure etc) being measured. As RX changes in value the bridge will become unbalanced and an output will be obtained, which is proportional to the change in light level etc.

The excitation used is most often ac because then the output will also be an ac voltage. An ac voltage is more easily amplified and more easily extracted from noise than a dc voltage.

Some instrument applications require a steady sine-wave to excite resonance in a mechanical component, this, in turn, is used to maintain the signal at a constant value. Many accurate clocks work on this principle.

The pacemaker, used to strobe the failing human heart into regular pumping action, puts out a steady train of stimulating pulses. In this application the design demands are for long-life, utmost reliability and, in the case of implanted units, small size, as is depicted by the units shown in Fig.5.

The time-base generator as used in the cathode-ray-oscilloscope, or in the television camera and monitor, provides a signal that steadily increases in amplitude in order progressively to deflect the electron-beam of a cathode-ray tube across the imaging surface. Once the end of the trace is reached it may retrace back again at the same rate or fly back at a much higher speed to start again. The generation of sawtooth and triangular signals that fulfil timing operations — the sweep action of a flat-bed plotting table is another example — is an often met need.

Digital calculators and computers (really one and the same thing in many cases, see Fig.6), require a steady source of timing pulses (a square or rectangular wave train) that pulses the digital-computational circuits along pulse by pulse. This is called the system clock. In many systems, clock rates may be as high as several million pulses per second.

In some applications a short burst of waveform only is required — not a continuous train. In this case the unit must generate the signal needed and then stop, waiting for the next demand. Ultrasonic and radar distance measuring devices are examples of this need: a single pulse is fed to the transducer or antenna, the time taken to go to the target, bounce off and return, is then measured against a time base. The cycle is then repeated for the next measurement.

Although noise is usually regarded as having only nuisance value in electronic systems, the occasional need arises where noise must be generated. The obvious way to test the performance of a system under noisy conditions is to feed it with signal and noise mixed together. By adjusting the signal-to-noise ratio the performance of the system on noisy signals may be measured.

One last example is the provision of tremulant (amplitude modulation) in electric guitars. This is simply the amplitude modulation of the artist's created musical sounds with a constant low-frequency signal. Once the string vibration has been transduced into an electrical equivalent signal form this is modulated by the low frequency signal.

These are but a few of the vast number of applications. The range of performance specifications vary so widely that we must employ many different generating methods to cover all needs. What is good for audio frequency is of no value at UHF frequencies. It is, therefore, necessary to study many alternatives if a really useful knowledge of electronic systems is to be gained.

SIGNAL TYPES

In Chapter Four we saw how all wave forms are composed of numerous sine-waves of different frequency and amplitude. In theory we need only generate sine-waves and mix them together as needed to obtain any desired waveform. In practice, however, this is rarely the way that wave shapes are created. There are many much simpler ways. For example, consider the generation of a low frequency square wave. All we need is a mechanical switch that repetitively opens and closes a dc circuit. To provide a ramp we need only use the voltage building up across a capacitor as it is fed from a dc source.

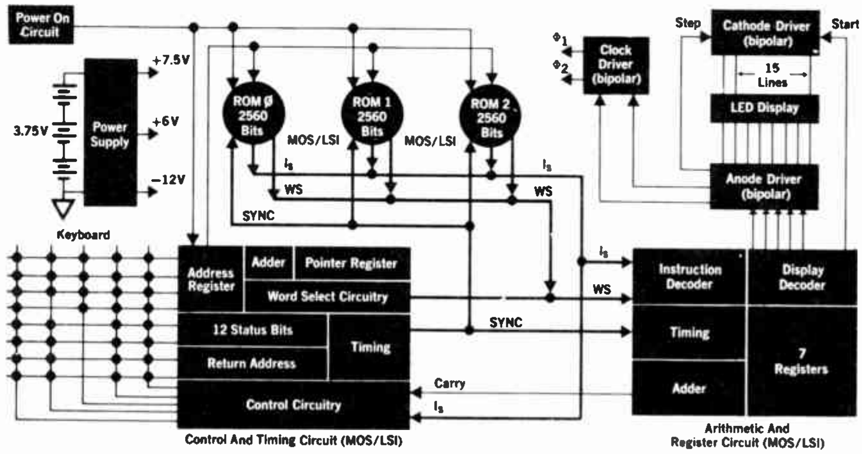
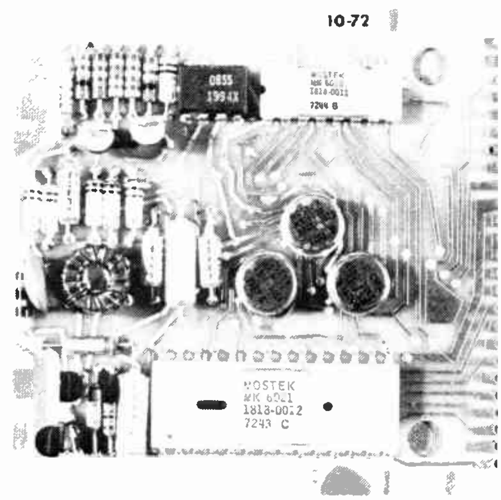


Fig. 6. Calculators work by processing pulses under the control of a train of clock pulses and instructions stored within a digital memory. Shown here is the schematic and logic card of the Hewlett Packard HP35 scientific calculator.



This is not to say that the Fourier-analysis method is of no value. Indeed, by recognising that waveforms *can* be built up from sine-waves we can improve wave shape by following design procedures that observe this rule. A perfectly-adequate, pure sine-wave can be generated by filtering the higher harmonics out of a square wave signal; a much used technique.

A variety of wave forms is shown in Fig. 7. They are photographs taken of signal generator outputs applied singly and in various combinations to the

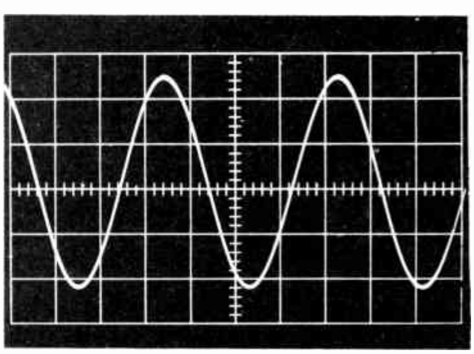
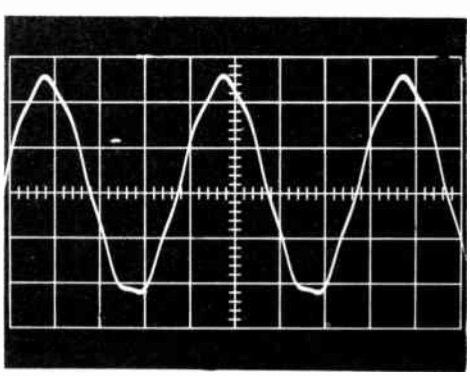
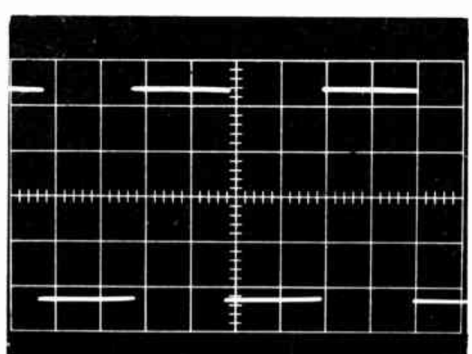
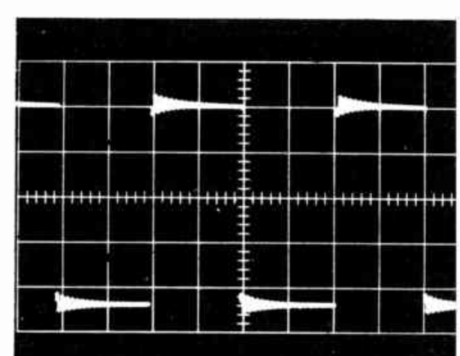


Fig. 7. Various waveforms as photographed on an oscilloscope screen.
(a) Sinewave
(b) Sinewave with harmonic distortion
(c) Square wave
(d) Square wave with ringing.
(e) Sawtooth or ramp.
(f) Triangular (g) White noise







vertical "Y" amplifier input of a cathode ray oscilloscope. The CRO trace is swept across the screen by its own in-built time-base generator. With practice it is possible visually to assess the quality of a wave shape to within a few per cent by studying its geometric shape. A mask, cut out in the shape of the required function, if placed over the CRO graticule, is a useful aid in wave-form distortion studies. Serious work, however, requires the use of expensive frequency analysers which record the amplitude (and possibly phase) of the frequencies present. A pure sine-wave will have no energy at any other than the frequency desired. Thus, in practice , distortion shows up as energy at other higher frequencies.

The square wave should be perfectly sharp and square-cornered — deviations from perfection show up as an obviously rounded rise and fall or as decaying ringing oscillations at the step transitions. Signal analysers are again ideal for studying the imperfections but are not always available due to their high cost.

Most general-purpose signal generators will provide both sine and square waves — this is because the square wave is easily derived from a sine-wave by amplification and clipping of both half cycles. More expensive generators will provide ramps, and triangular waveforms as well, and perhaps single shot pulses of any waveform type, the pulses being manually initiated or triggered by an external signal.

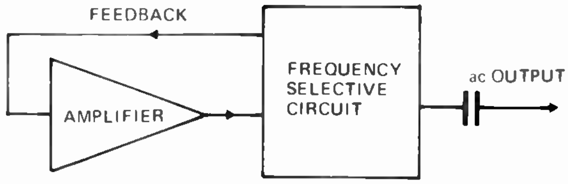


Fig. 8. Basic block diagram of a feedback oscillator.

More expensive, instrumentation signal-generators generally fall into two groups; precision sine-wave (and other shapes) generators with two outputs that can be varied in phase with respect to each other, at the other extreme is the pulse generator which provides digital forms of signal.

For practical reasons, generators cover a specific range of frequencies. There are low frequency generators that provide signals of various waveforms with repetition rates of cycles per hours to several kilohertz. (What is low is highly subjective — the earth scientist regards frequencies of hundreds of years as definite ac signals; the electronic engineer would treat these as pure dc, ac signals being to him those from one cycle per second upward. The optical engineer works with frequencies of terahertz! It is all a matter of relativity.

The audio range is covered with generators providing from around 10 Hz to 20-100 kHz, there being little need for higher or lower frequencies in audio studies.

Radio frequency, (RF) generators provide frequencies needed in radio work, for example, 500 kHz to 1.5 MHz for the broadcast band. Yet higher, are systems for testing and driving radar networks (in the GHz range).

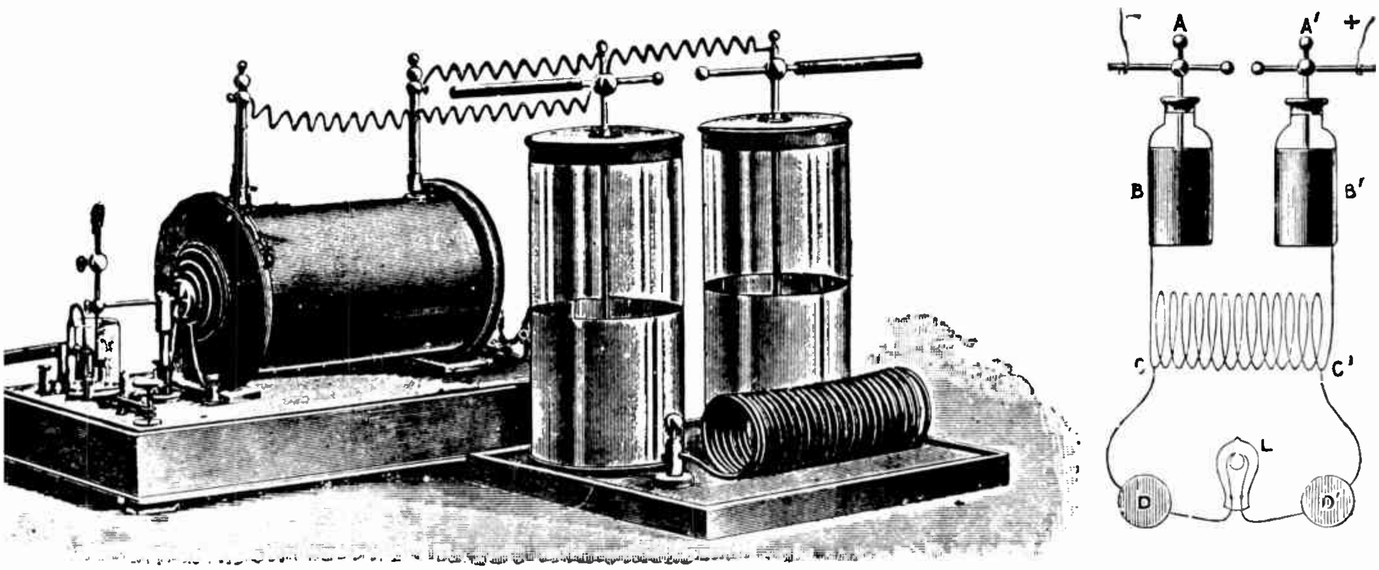
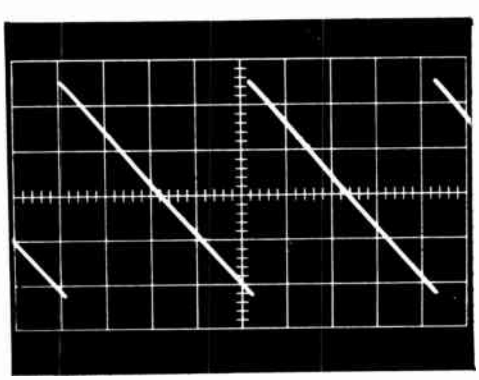
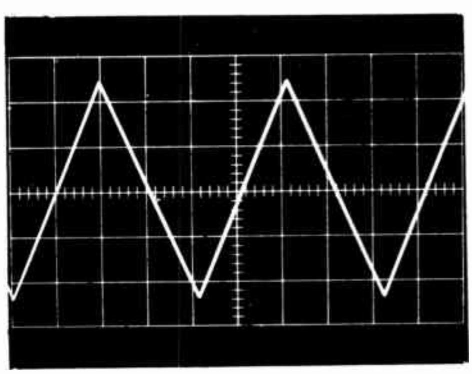
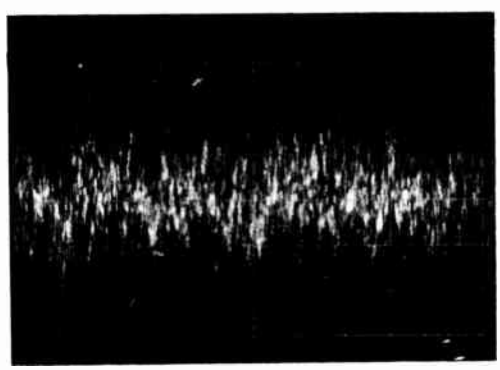


Fig. 9. High frequency generation using a spark discharge from a tuned LC network. The D'Arsonval high frequency apparatus employed two Leyden jars where the outer coatings were joined by a helix, and the inner coatings fed by an induction coil.







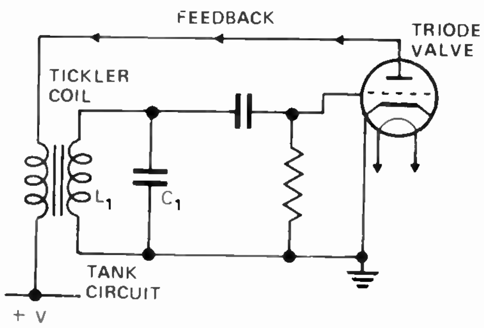


Fig. 10. Schematic of Armstrong's oscillator – the first practical unit to apply positive feedback to the thermonic valve. Developed in 1914.

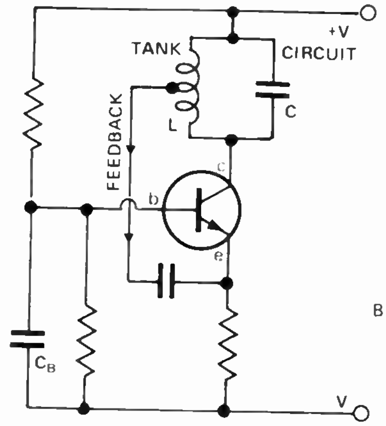


Fig. 11. The circuit of one of the basic forms of feedback oscillator. This type (feedback from tap on tank coil) is known as a Hartley oscillator.

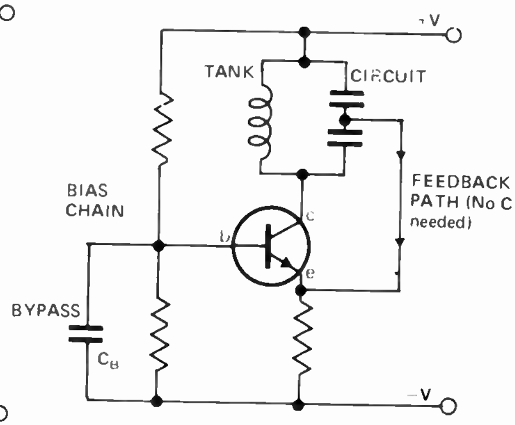


Fig. 12. An alternative arrangement of feedback taps the capacitor rather than the coil. This circuit is called a Colpitts oscillator.

Noise generators provide basically white noise (white noise is defined as having constant power at all frequencies) within the range of interest. There is no sense in building the generator to provide megahertz frequencies when the unit is intended for 1 Hz system testing. By appropriate filtering noise generators may also provide 'pink noise' (noise which falls off uniformly as the frequency rises) or one of many forms of 'grey noise' (non-ideal white noise where the power at various frequencies follows no set law).

Some generators can provide bursts of signal waveform at preset intervals between them. These are useful in radar and sonic distance gauging, in telegraphy and in electronic-music synthesizers.

You might by now feel that the variety of needs implies an overwhelming number of techniques to comprehend. Fortunately the techniques used are far fewer than the possible applications, and many techniques are common to a number of frequency ranges.

BASICS OF GENERATION

Open loop – The desired waveform may be created by using an appropriately shaped mechanical device. A time-varying waveform can be produced by rotating the mechanical part at the speed required.

For example, the experiment described in Chapter Five used a rotating blade to alter the light level on a light-dependent resistor. The photo-optical method of generation is economic where very low-frequency waveforms of great complexity are needed. One commercial unit uses a circular transparent disk that can be masked as needed. One unusual application of such a device was to simulate the respiration signals of a snake. This allowed the data processing equipment, used in research on snakes, to be tested without a live snake (snakes vary their respiration rate randomly from minutes to hours between breaths).

Rotary mechanical generators generally produce signals in 'open-loop' – no use is made of the signal produced to modify itself. Tone wheels for electronic organs and strobing lines on turntables also operate this way. In general, however, most signal generators used in electronics make use of feedback, or closed-loop systems.

Closed-loop – On several previous occasions we have seen how negative feedback (that is, the sign of the feedback voltage is opposite to the signal input) helps to stabilise circuits. Examples are, the use of an emitter resistor to improve thermal stability in the basic ac transistor – amplifier stage, and in the operational amplifier networks where a desired gain is obtained from a much higher gain amplifier; the advantage of negative feedback being greater precision and stability.

The converse also applies: positive feedback leads to enhanced instability. By arranging for the output of the active device to feed back into its input with like sign (similar phase) any small change in input leads to rapid build up of output – up to an amplitude governed by the circuit. If the circuit can then be so arranged

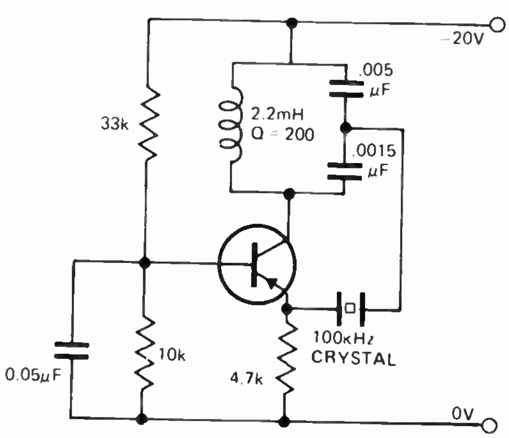


Fig. 13. Oscillator using feedback via a crystal in series-mode.

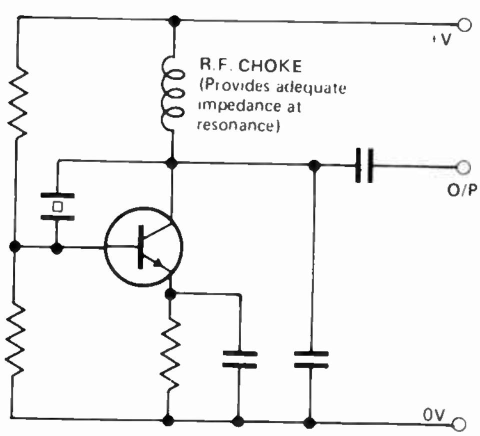


Fig. 14. The Pierce oscillator uses the quartz crystal in its parallel mode of vibration.

that once this limit is reached the source of energy is disconnected the output will fall to zero again. In the meantime the source becomes energised again and the device output again rises. The process repeats providing a steady cyclic signal of wave form decided by the mechanism used. Some oscillators, as such generators are called, produce square-waves, some sinusoids, others ramps.

In general, feedback oscillators can be shown as a system, with a block diagram as in Fig. 8. An amplifier feeds a wave-shaping circuit that controls the time-behaviour of the all-important positive feedback. This diagram should also help to explain why high-gain circuits often oscillate, for any in-phase feedback will turn what is intended to be a stable circuit into an oscillator – the higher the open-loop gain the less in-phase signal needed to cause oscillation.

The natural frequency of oscillation of a feedback system depends upon component values and can be estimated with reasonable accuracy in most designs. There are three basic criteria that must be satisfied if a feedback circuit is to oscillate. Firstly, the phase shift through the amplifier and forming circuit must be zero, or close to zero at the frequency of oscillation required. Any shift toward 180° provides increasingly greater stabilising action. Secondly, the voltage gain of the amplifier and forming circuit must be greater than unity at the frequency needed. Thirdly, the voltage gain must drop to unity once oscillation has begun. If it did not do this, the amplitude would keep on building up with time.

These rules will become more evident as we now look at several design alternatives.

BEFORE ELECTRONICS

Before thermionic valves were invented the generation of high-frequency signals was particularly difficult. One way was to use the oscillatory discharging action of a magnetic-induction coil coupled to Leyden jars. The two formed a tuned circuit which, when the jars were fully charged from the dc source, discharged across a spark gap. A typical layout is shown in Fig. 9 along with contemporary equipment used in medical-electricity treatment around 1900. These produced a train of bursts of decaying oscillation.

Also commonly known at the beginning of this century were the Faradic coil generators. These were based on the electric-bell principle. A magnetic coil was fed with current. As the magnetism built up it pulled in a small armature that opened a contact

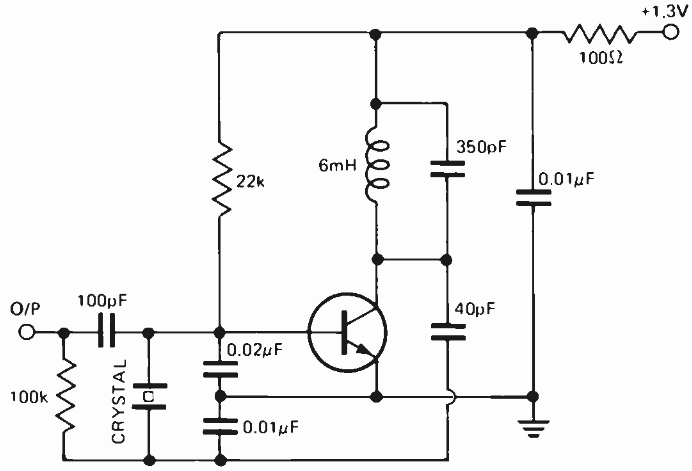


Fig. 15. This very stable oscillator, with a temperature coefficient of one part in 10⁸, is used for calibration purposes.

thus disconnecting the coil. This then allowed the armature to return closing the contact again. The system, therefore, oscillates at a characteristic frequency. High voltage pulses were produced from this chopped primary current by induction in a secondary winding having a large number of turns.

These two devices were commonly known when the first triode active element was invented in 1907. They probably led to the creation of the first useful thermionic valve oscillator built by Armstrong in 1914.

ARMSTRONG OSCILLATOR

Armstrong's arrangement is shown in diagramatic form in Fig. 10. The resonant circuit L_1 C_1, if initially charged, will oscillate at its natural frequency, decaying to zero in a few cycles as the energy is dissipated in resistive losses in the coil. When

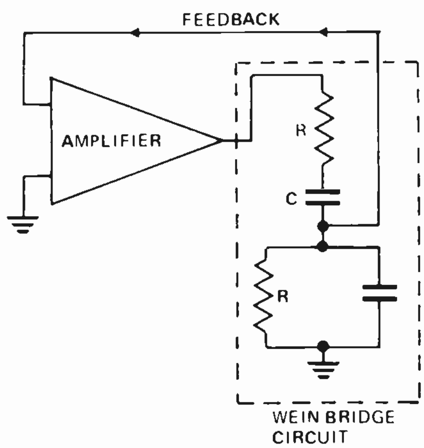


Fig. 16. The Wien bridge network (in the dotted box) provides feedback which is a maximum at one particular frequency. If an amplifier, having sufficient gain, has positive feedback via such a network it will oscillate.

combined with an amplifier, the valve senses the periodic changes in the tuned circuit and feeds a signal back (via the tickler coil), into the tuned circuit, which reinforces the resonance, making up for the losses. Provided the phase is correct the oscillation builds up to a level limited only by non-linearity of the amplifier stage. It is like pushing a child on a swing. In this case the resonant circuit provides the frequency control for it will not resonate at any but its natural frequency. Away from resonance the system responds too weakly and oscillation dies away.

HARTLEY OSCILLATOR

If the inductor of the resonant circuit (called the tank) is tapped, the signal produced at that point can be used to provide the correct amount of positive feedback as shown in Fig. 11. In this case the base of the transistor is effectively earthed through the low impedance of C_b. Thus variations in the tank circuit provide positive feedback to the active element. (Feeding the signal to the base is negative feedback, to the emitter is positive feedback.)

COLPITTS OSCILLATOR

The same effect can be obtained by splitting the capacitor instead of the inductor of the tank circuit. Figure 12 shows the so called Colpitt's arrangement. The output may be taken from the collector, the emitter or from the tank inductor by transformer action.

In each of the above resonant-circuit forms of oscillator the amplifier acts to inject a pulse of current, rather than a smooth change, into the tank circuit. When the amplifier operates this way it is called Class C mode of amplification. The tank circuit provides the quality of waveform. Think of the swing again. This oscillates sinusoidally yet is pulsed to keep it going. In Class -C operation the

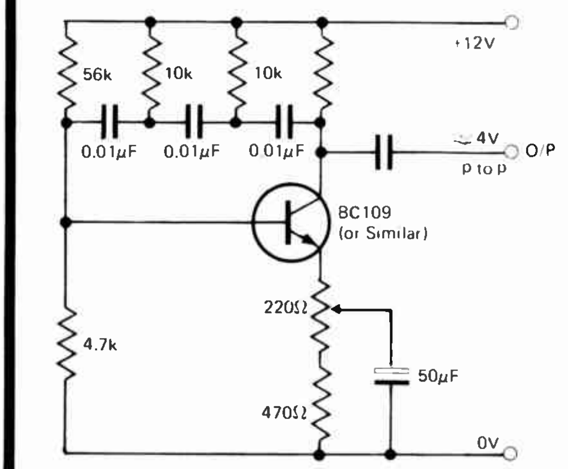


Fig. 17. Inexpensive sine wave generator uses a ladder network to phase shift the output signal 180° at the required frequency of operation.

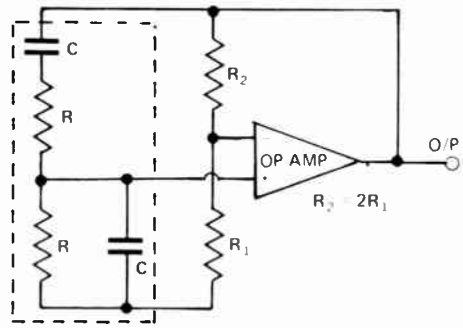


Fig. 18. Wien bridge oscillator using standard, linear op-amp IC (LM301, μA741 etc). The frequency of oscillation $f_o = 1/2\pi CR$. Thus if $f_o = 1000$ Hz suitable values for R and C would be 16 k and 0.01 μF respectively.

active element does not need to respond in the linear region of its characteristic curves. The practical advantage is that it can handle considerably greater power levels in this mode. The actual signal in the active element is, however, a highly distorted version of the output.

QUARTZ CRYSTAL OSCILLATOR

In some designs the frequency generated must be extremely stable. The degree of stability is related to the quality of the resonant circuit in the resonant-type of oscillator. Normal LC tank circuits will achieve quality factors (Q, that is the ratio of reactance of the coil to its resistance) of about 100 but not better because in passive electrical tuned circuits the Q is limited by resistive losses in the inductance.

Mechanical tuned circuits, especially those based on the resonance of quartz crystals, have Q's of 10^5 – they will resonate with about 1000 times greater stability than LC tank circuits.

Quartz is particularly useful for it also possesses sizeable piezo-electric effect. Voltages applied to deposited contact areas will cause the thin slice of quartz to change dimensions. Upon removal of the voltage the crystal resumes its shape generating a voltage whilst doing so. Thus we have an electro-mechanical transducer which is compatible with the electric circuit and which can be used as the equivalent of a superior electrical tank circuit.

A series mode quartz-crystal oscillator is shown in Fig. 13. The crystal "tank" couples the Colpitt's connection back to the active element. The collector tank circuit, which can vary a little due to its low Q, will become synchronised to the natural frequency of the quartz crystal.

Quartz crystal oscillators are useful for frequencies well above audio – in the region of 100 kHz to 100 MHz – but find little application at low frequencies due to the natural limitation of quartz resonance within practical crystals. Where precision low frequencies are needed a quartz oscillator output is divided down to obtain reduced cycle periods. The quartz crystal can also be used in its parallel mode of resonance (decided by the way it is cut and connected). The Pierce crystal oscillator, shown in Fig. 14, uses the parallel mode.

Quartz crystal clocks and watches operate on this principle. A highly stable signal frequency is counted giving time from the known period of the waveform. A highly stable crystal is given in Fig. 15. The tank circuit is tuned to nominally 100 kHz and maintains its value to within one part in 100 000 000 per degree celsius change or per 10% change in supply voltage.

PHASE SHIFT OSCILLATORS – WIEN BRIDGE

Another method of obtaining sinusoidal oscillation is to provide the feedback by a network that only provides zero phase shift at one very specific frequency. The stability of frequency then depends upon the network configuration and the stability of the component values.

The Wien-Bridge oscillator is the most commonly encountered phase-shift arrangement. Referring to Fig. 16 it can be shown mathematically that the right-hand network (taken from a Wien bridge arrangement) feeds back a signal which is sharply maximized for one particular frequency and which has zero phase shift. Provided the amplifier has adequate gain the system will oscillate.

The main advantage of phase-shift circuits is that, for low frequencies in

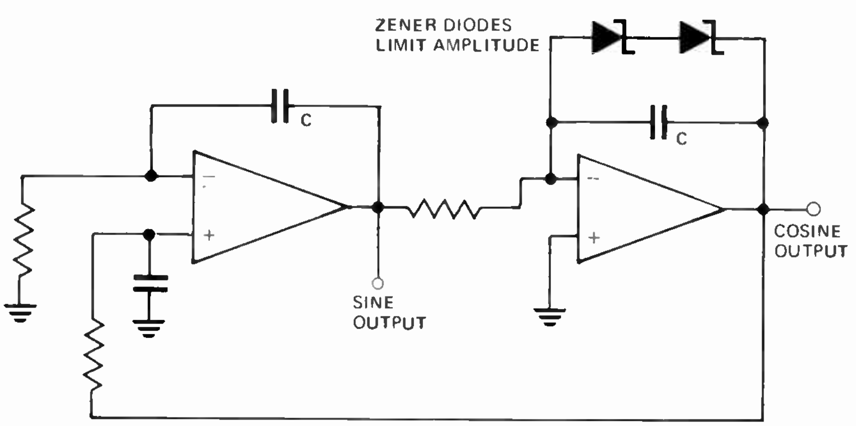


Fig. 19. Generation of sine and cosine waves by implementation of a second-order, differential equation by means of two IC stages. Both stages are essential even if only one output is required.

the audio region, large and expensive inductors are avoided.

LADDER NETWORK PHASE-SHIFT OSCILLATOR

The appropriate phase shift can be obtained if the output of the amplifier is fed back to the input by a chain of RC filter stages. A typical circuit, set to generate 800 Hz, is given in Fig. 17. Note that it is a conventional ac amplifier stage (refer to Part 10 Vol.1) with the ladder network added between the collector and the base to provide the selective positive feedback.

USING IC's

Although many designs of oscillator are based on a single transistor amplifier it should be clear that any amplifier can be wired up with feedback to provide oscillation. Thus, the now inexpensive integrated-circuit operational amplifier is capable of providing better performance with lower output impedance than designs built with discrete components.

Application notes include numerous designs — we show one, that of a Wien bridge oscillator in Fig. 18. For this the bridge circuit, added to the op-amp, is balanced at $f_0 = 1/2\pi CR$. The greater the gain the closer the system holds to this value — hence the improvement obtained by using an op-amp opposed to a single transistor (gains 100 000 and 100 respectively).

In our earlier discussion of operational amplifiers (Part 12 Vol.1) it was stated that capacitance feedback around a single stage provides an integration action. The solution of a second-order differential equation (those that describe the behaviour of the spring-mass system, L-C resonant circuit and pendulum movement) is a sinusoidal signal. If there is no effective damping the system resonates continuously. We can therefore, using two op-amp integrators, set up such an equation and set it going to produce sinewave and cosinewave signals. Figure 19 shows how this is done. The Zener diodes clamp the output to a set maximum value; without them the output would increase to uncertain limits.

In the next part we will continue with other classes of oscillators — those that produce non-sinusoidal waveforms, those that provide a signal frequency that is electronically adjustable, those that provide digital signals and those that must be used to generate signals at very high frequencies where techniques of amplification are quite different to those based on transistors.

REFERENCES

Most books on transistor circuit design explain the principle and theory of oscillators. Few devote much space to the subject however the following will be found helpful.

"Electronics for the Physicist" — C. F. G. Delaney, Penguin, 1969.

"Transistor Manual" — General Electric, 1969.

"Electronic Instrumentation Fundamentals" — A. P. Malvino, McGraw-Hill, 1967.

Books and articles on electronic music hardware cover many circuit ideas.

"Electronics in Music" — F. C. Judd, Neville Spearman, 1972.

For operational amplifier designs consult manufacturers' application notes and books on the subject such as —

"Operational Amplifiers — G. B. Clayton, Butterworths, 1971.

"Modern Operational Circuit Design" — J. I. Smith, Wiley, 1971.

"Operational Amplifiers — Design and Application" — J. G. Graeme and G. E. Tobey, McGraw-Hill, 1971.

ELECTRONICS — IN PRACTICE

THE BEST WAY to become familiar with oscillators is to build a few basic circuits such as those given here. Figure 20 shows the circuit for another form of phase-shift oscillator that uses a twin-T (a form of frequency selective bridge) feedback arrangement. Its frequency output is stable to within a few parts in 10 000 for 10% supply variations. To obtain best temperature stability (0.2% change for a temperature change of — 20°C to 80°C) use stable capacitors such as polycarbonate types.

The advantage of this circuit is that it will operate at low frequencies. If $C_1 = C_2 = C_4/2$ and $R_3 = R_4 = 2R_5$ then the resonant frequency will be at $f_0 = 0.159/R_3C_1$. (f_0 in hertz, R_3 in ohms, and C_1 in farads). With the values shown the circuit operates at 70 Hz.

BIRD SOUND GENERATOR

As an example of synthetic sound generation the circuit given in Fig. 21 produces a chirping sound similar to that of a bird. The adjustable resistor alters the tone of the chirp.

If a light dependent resistor (ORP12 for instance) is inserted in series with the adjustable base-bias resistor, the chirp will only occur when the ambient light level is high — the bird goes to sleep at night. An amusing trick is to mount the entire circuit inside a small box — with the lid is opened the bird chirps.

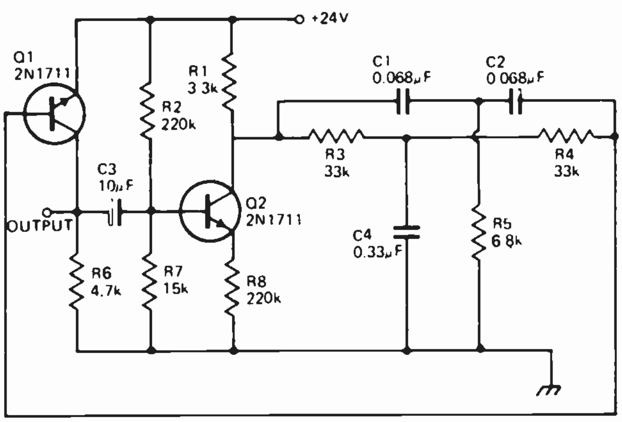


Fig. 20. Phase shift oscillator based on a twin-T, frequency selective network.

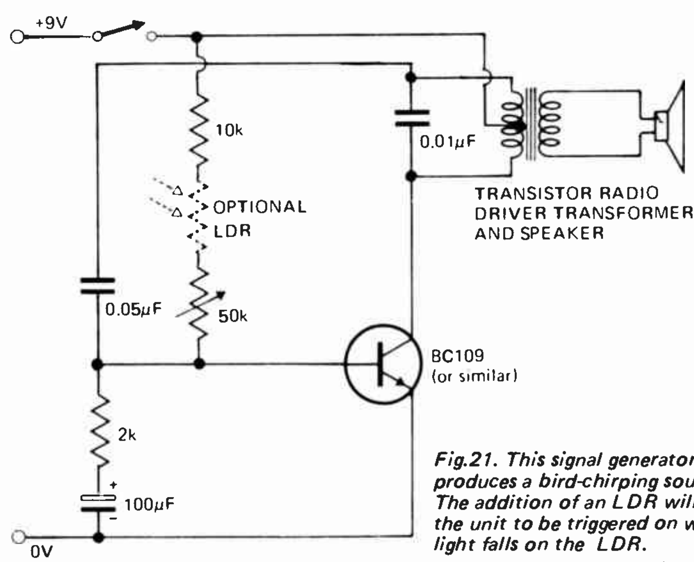


Fig.21. This signal generator produces a bird-chirping sound. The addition of an LDR will allow the unit to be triggered on when light falls on the LDR.

18 Generating non-sinusoidal waveforms

THE UNIJUNCTION TRANSISTOR

The unijunction transistor, UJT for short is a three terminal semi-conductor device somewhat akin to the normal transistor. The exception is that by appropriate choice of the manufacturing placement and thickness of material, it has entirely different characteristics between the currents and voltages of its three terminals. It becomes a device in which a current flows once it is triggered on. It features a stable triggering voltage Vp, a very low value of firing current Ip, a negative resistance region (where a rise in voltage between terminals is related to a fall not a rise in current) and a high current carrying capacity once it has been pulsed on. As there are 50 different kinds we can only give a general impression of their operation here. The symbol is shown in Fig. 1.

When the emitter, E, is reverse-biased no current can flow between B_1 and B_2 bases. When V_E rises sufficiently it alters the state between the bases and, quite sharply, I_{B2} commences to flow. This is seen by studying the representative static emitter characteristic curve given in Fig. 2.

Until V_E rises to close to Vp (the actual value depends upon the standing value of V_{BB} and ranges typically from 3 V to 20 V for V_{bb} of 4 V – 30 V) I_E remains virtually zero. After the I_{EO} value, emitter current can be absorbed and V_E drops back exhibiting a negative resistance region. In the unijunction, therefore, the equivalent of the collector-emitter current flow of normal transistors takes place between the emitter and base 1; base 2 acting as the input that decides at which point the circuit goes into the conduction state.

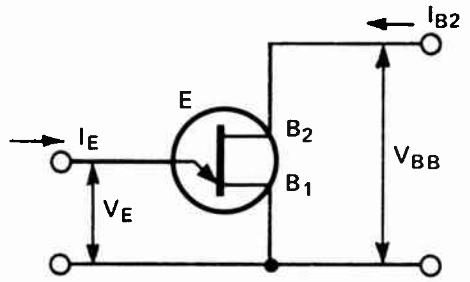


Fig. 1. The basic unijunction symbol.

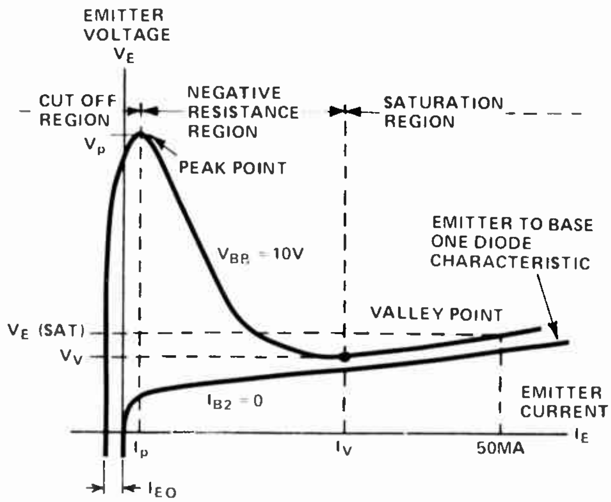


Fig. 2. The unijunction characteristics.

A BRIEF description of how a cathode ray oscilloscope operates was given in Chapter Four. There we saw how a sweep generator is used to provide a signal that causes the spot to sweep across the screen thus tracing the waveform. A television receiver uses the same principle as indeed does a television camera. This effect is achieved by steadily increasing the voltage applied to a deflection plate (or deflection coil in a TV system).

When the spot reaches the limit of travel needed it is swept back again, by reversing the voltage change, so that it eventually arrives back at the original level. If the rise and return of such a waveform are both at the same rate the resultant waveform is triangular; if the waveform returns to the starting point virtually instantaneously – it is a saw-tooth. Sweep times required vary from hours down to a few nanoseconds depending upon the particular application.

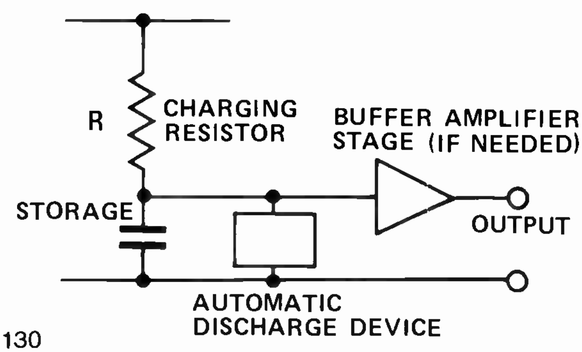


Fig. 1. Functional block diagram of basic relaxation oscillator.

A common use of saw-tooth-like signals is in the modulation of tones – eg, the rise and rapid fall of modern police sirens. The tones used in some telephone systems and in organs and electronic music synthesizers are other well known examples of the application of saw-tooth waveforms.

Most simple sawtooth generators are based upon what is known as the relaxation principle. In this method a capacitor is charged (Fig. 1) though resistor, R, which limits the current and hence the rate of rise of voltage across the capacitor. When the voltage across the capacitor reaches a preset limit, some form of device is actuated that discharges the capacitor back to its initial point. Inductors could also be used in a similar manner but the use of capacitors is more usual. Once the capacitor is discharged the device becomes inoperative and the voltage rises again to repeat the cycle. The simplest form of relaxation oscillator

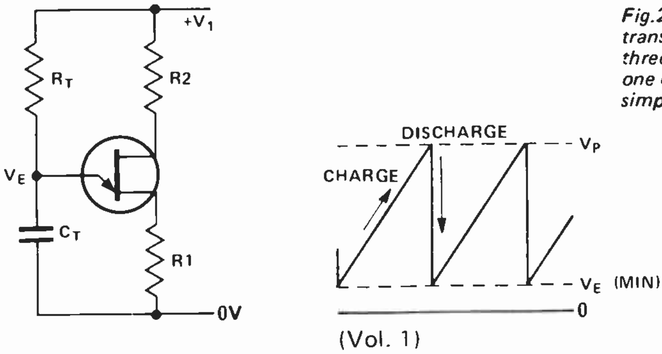


Fig.2. The unijunction transistor together with three resistors and one capacitor makes a simple sawtooth generator.

uses no more than a capacitor, a resistor and a neon lamp as the device to discharge the capacitor. The details of this method were given in Part 7

As with most of electronic technique (and life in general) the simplest method is not necessarily the best. The neon method requires a high voltage supply, by today's standards, and the charge and discharge slopes are exponential rather than straight linear rises and falls.

Improvement can be made by making the charge process more linear, that is, by using a method that gives a purer integration. The first steps toward improvement are to use a much smaller part of the exponential charge curve of a capacitor, for this will be more linear, alternatively we can provide a more linear charge technique. These methods however, usually call for a more sensitive trigger to discharge the capacitor, that is, an active trigger element. Whereas it is quite feasible to use transistors as discharge elements the more practical method, usually employed in relaxation designs, is based upon a device known as the unijunction transistor.

The unijunction transistor has three terminals labelled base 1, base 2 and emitter. The emitter to base 1 resistance of a unijunction is normally very high, however, when the emitter is raised to a voltage known as the peak point, emitter to base 1 resistance drops to a very low value. This property may be used quite effectively to discharge a capacitor once it has reached the peak point voltage of the unijunction.

In the circuit of Fig. 2, capacitor C1 charges through R_T from source V_1 therefore, no emitter to base circuit is made. When V_E reaches the peak emitter voltage V_p of the unijunction (decided by the circuit values R_1, R_2) the unijunction changes state and the emitter to base-one resistance falls to a low value. This discharges C_T through R_1. When the voltage has fallen to V_E(min) the emitter to base 1 discharge circuit becomes a high resistance again and the cycle repeats. Provided R_1, R_2 have small values compared to R_T the oscillation frequency is given by

$$f = \frac{1}{R_T C_T \ln\left(\frac{1}{1-\eta}\right)}$$

where η is the intrinsic stand-off ratio quoted for the unijunction device. It has values typically around 0.6.

If V is kept large compared with V_P the capacitor is charged from a more constant-current source improving linearity. With this technique the linearity, however, still has an error of 10% or so. Using a separate, even-higher charging voltage further improves linearity but at the expense of a more complex supply. Another method is to use a transistor in the charging path to provide constant-current flow to the capacitor.

A much superior circuit is given in Fig. 3. The capacitor C_2 and the output buffer stage improve the linearity by stabilising the voltage across the charging resistor feeding the 100 ηF capacitor. Components R_1 and C_1 are added as an integrating compensating network that further improves the linearity of the charge process. Variation in R_1 is provided to trim the wave shape rise characteristic from concave through linear to convex. As shown, the circuit generates a 1 kHz sawtooth. Note the ability to provide two anti-phase signals and the input that enables the system to be started in synchronism with an external event – as is required to trigger, say, a CRO trace upon demand.

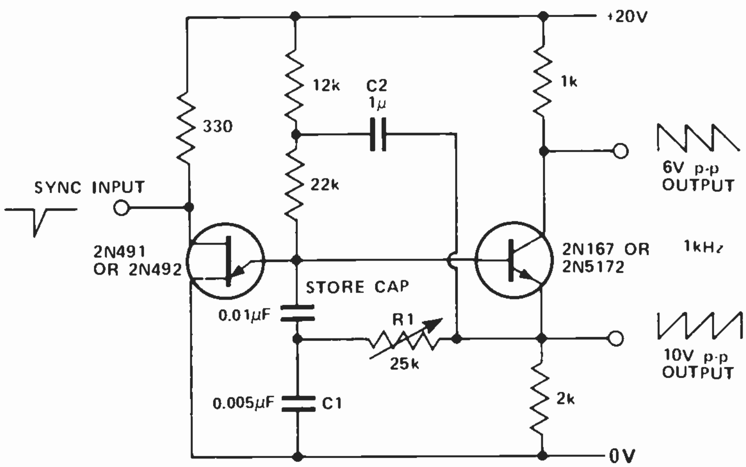


Fig.3. This generator adjusts the sawtooth to a high degree of linearity.

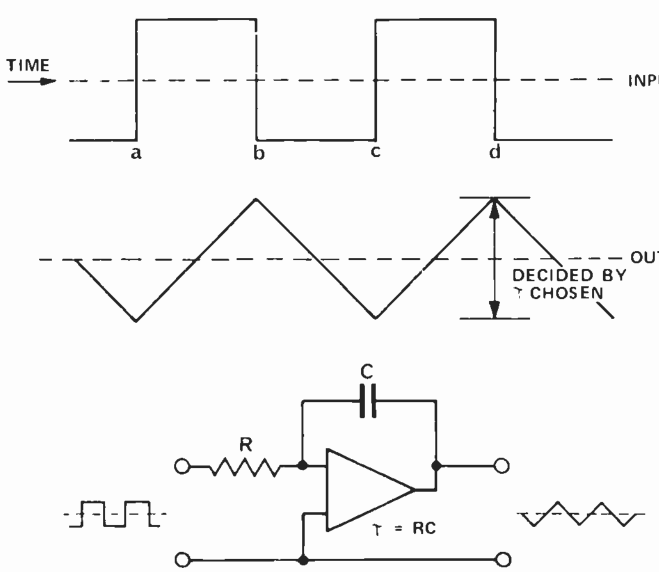


Fig.4. A symmetrical triangular wave may be generated by integrating a square wave.

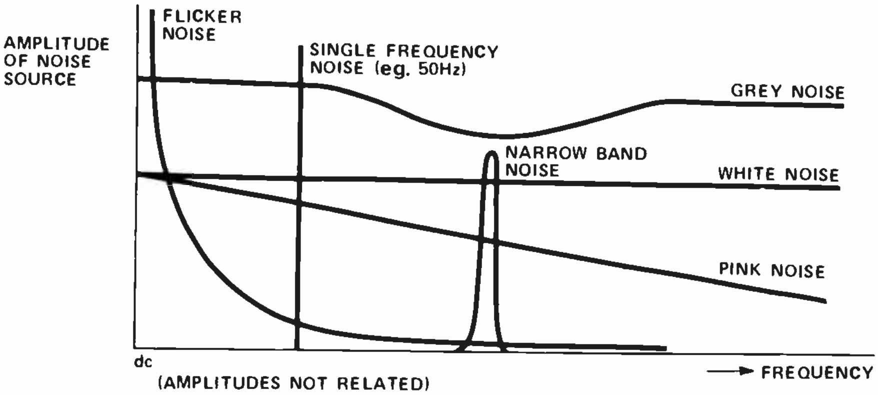


Fig.5. Frequency spectra of commonly encountered noise sources.

By appropriate design, relaxation oscillators can be made to provide sawtooth, triangular and pulse wave shapes. They are also the basis of timing circuits — for the time-constant of the capacitor and the trigger relaxation level effectively defines a time interval.

Yet another way to produce a sawtooth is to generate a number of sinewave signals of chosen frequencies and amplitudes covering the spectrum of the sawtooth. These can then be combined to produce the sawtooth. This method is suitable for synthesizers or other precision generators but would normally be prohibitively expensive — over ten generators would be needed to provide a reasonably accurate waveshape.

A symmetrical triangular waveform may be generated by starting with a square-wave source and integrating it with an op-amp type of integrator. This is shown diagrammatically in Fig. 4. At point 'a' the integrator output starts to rise in the positive direction. When the square wave reaches 'b' the integrator input reverses and the output starts to fall until 'c' is reached. It is, however, not fundamentally possible to have different rise and fall times if the amplitude is to be held constant: different rates require different integration time constants for both directions of signal change.

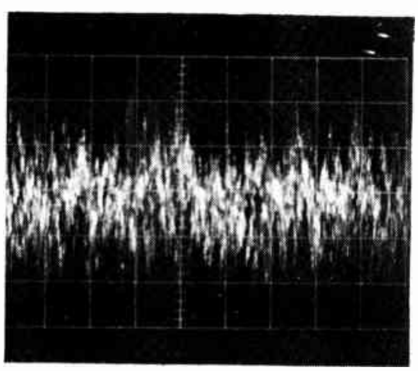


Fig.6. How random noise looks on a CRO screen.

NOISE GENERATION

So far we have said little about noise, that generally unwanted signal that must (usually) be kept to a satisfactory minimum in circuit design. Indeed, it might seem strange that we should sometimes want to *generate* it, when the usual aim in design is to eliminate it.

Although noise may be any unwanted signal and, therefore, can consist of any combination of an enormous variety of waveforms, the noise usually referred to will be what is known as random noise. Random noise is a signal that has the interesting, but frustrating property that one cannot predict the exact level of signal at any particular instance. We can only characterise it by the use of random statistics that will tell us, if we know the type of noise, the *chances* of certain levels occurring at a given time.

Various kinds of noise are termed white, pink and grey. Each is typified by the nature of its frequency spectrum. White noise, the usual one considered (but in reality not always the one that really occurs) has equal energy at all frequencies. The energy level of the signal will be the same at 1 kHz as it is at 100 kHz. There is equal noise energy at all frequencies with white noise. In practice noise energy may fall off uniformly with rising frequency (pink noise), or it might not be quite white in that there may be variation in energy at various points of the spectrum (grey noise).

Unwanted white noise mainly arises due to thermal agitation in resistors. This effect is called Johnson noise. It is a basic effect that can only be reduced by reducing resistance values, or by operation at lower temperatures.

As noise exists at all frequencies, reducing the band-width of a system reduces the total noise power occurring at the output.

Another noise phenonemon is known as excess or flicker noise. It is also sometimes called 1/f or hyperbolic noise. This is noise that rises in level as the frequency is reduced. It occurs in all semi-conductors. It is usually less than the resistor-generated white noise (above 1 kHz) so is not a problem at high operating frequencies. The various types of noise are depicted by a representative plot of their frequency spectra as in Fig. 5. In contrast to noise of the random kind the spectrum of induced 50 Hz hum is a single line. Random noise, usually presumed to approximate white noise, appears as shown in Fig. 6 on an oscilloscope screen. Audibly it sounds like hiss because the ear is most sensitive to frequencies in the 1.5 – 6 kHz region, thus the ear subjectively attenuates frequencies above and below these rough limits.

The amount of noise generated internally in an electronic system is a limiting factor. The noise performance is usually specified as the *noise figure* of the system. Noise figure is the ratio of signal-to-noise at the output to signal-to-noise at the input expressed in decibels. Thus a noise figure of 2 dB is much better than one of 6 dB.

One way of ascertaining the noise contributed by a system is to measure the total output power of the system under test with a suitable driving (wanted) signal and then without the test signal. The residue is noise power. The usual method of stating noise power is as the RMS level of the random process.

In another class of tests, noise of a known level and character is added until the noise output of the system is doubled. The amount added then equals the amount internally generated.

White noise generators can be built using a wide-band amplifier to raise the signal level of Johnson resistor noise. This method, however, is seldom used in practice due to the comparatively lower noise output from resistors compared to other alternatives.

For example, a Zener diode generates much more intense internal noise than does a resistor. Two simple noise generating circuits are given in Fig. 7. One, Fig. 7a, will provide white noise suitable for audio work. Capacitor C_2 (if added) filters the output reducing the noise level as the frequency rises — thus providing pink noise. The other, Fig. 7b, is suited to VHF work as the bandwidth extends beyond 150 MHz. Resistor R is adjusted to pass about 6 mA through the circuit. Capacitor C_2 should be a ceramic capacitor. Output

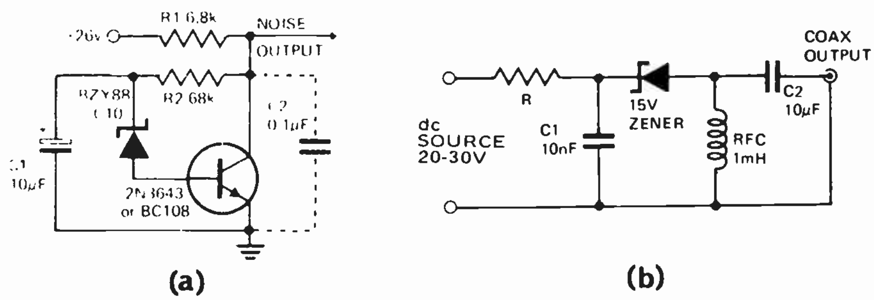


Fig.7. Two noise generators. (a) noise generator suitable for audio work. (b) for VHF use.

via a coaxial cable, in this case, is essential to preserve the bandwidth of the signal.

Another noise source sometimes encountered relies on the variation of contact resistance in an electro-chemical cell — this produces a good signal at relatively low frequencies. It is also possible to approximate noise in certain cases as a binary (that is, two-state only) signal that switches between states in an apparently random fashion — such generators are called pseudo-random binary sequence generators, PRBS for short. These generate their signal by virtue of specially connected ring-counters, a technique we will study later in the series. The output of these can be averaged with a CR filter to provide analogue noise.

NON LINEAR AND NON-REPETITIVE ANALOGUE WAVE SHAPES

The waveform producing circuits considered so far generate sinewaves or linear ramps. In some instances the need may be for a special shape other than those producable by standard circuits. If you are lucky the distortion of some oscillators may be the waveform needed — the exponential rise of single relaxation oscillators can sometimes suit the non-linear characteristics of CRO tube deflection systems.

Provided the cyclic frequency needed is not too high, that is, up to about 10 kHz, it is possible to make use of an optical-disk generator. In this method a transparent disk, on which is placed a mask of the required signal shape rotates at a controlled speed. Light passing through a portion of the disk is integrated by a photo-detector and collection system providing an output proportional to the degree of masking at each point. Figure 8 shows disks suitably masked to provide random noise, sawtooths and heart beats. This method is admirably suited to the generation of very low frequency complex waveshapes — down to 0.001 Hz but suffers from the possible disadvantage that the waveshape period is rigidly related to frequency. It is not possible to retain a fixed cycle time with changing repetition rate.

Given a mini-computer facility it is possible to generate any waveshape as a repetitive event, or as a "one-shot" event, by controlling the signal flow with time. In hybrid computer operations (those combining analogue with digital methods) the mini-computer operates switches that gate voltages to the output. Each change in the circuit alters the rate of rise of the output, that is, the instantaneous slope is controlled. It is, therefore, possible to create a waveshape by successive linearisation of the originally smooth curve into one made up of a string of different slopes joined end to end. The number of stages used decides the degree of accuracy of generation.

It is also possible to generate unusual voltage-versus-time functions using diodes in conjunction with op-amps. Figure 9 shows the schematic diagram of an analogue, op-amp set-up that generates an output voltage as would come from, say, a potentiometer driven back and fourth by a badly worn mechanical linkage. It also simulates gear backlash and a crude approximation to magnetic hysteresis. All resistors are equal and the integrator time-constant is very small. As it rises from zero there is no change in I_o initially, as there is no current flow into the integrator because of the reversed-biased state of the diodes. When the input reaches, say E_2 a diode conducts, starting the integrator which operates until the output causes the same diode to cut off. If e_i now decreases, the output is held high since the integrator cannot discharge until e_i falls to a value which causes the other diode to conduct. The functional relationship between input and output that results is that depicted in Fig. 9. This example shows how seemingly difficult-to-produce functions can in fact often be quite easily produced using op-amp techniques. The batteries are included to set the voltage at which the diode conducts. The need for batteries may be eliminated by using Zener diodes in place of the simple diodes used in Fig. 9.

SQUARE WAVES

Square and rectangular waves are most important in digital circuitry because they have signal levels that can be only one of two definite states (the transition times being considered negligible). They also are used as the starting point for generating pulse trains in which the signal consists of narrow pulses. Three main methods are used to generate square waves. Two start with sinewaves, converting them to square waves, the other generates the square wave directly from a dc voltage.

If a sinewave of the same frequency as the squarewave needed is greatly amplified, the slopes of the sinewave at the zero crossing point are raised more toward the vertical. Also if the amplifier is overdriven the upper and lower limits of the original sinewave will be clipped. A crude square wave results. A more positive clipping

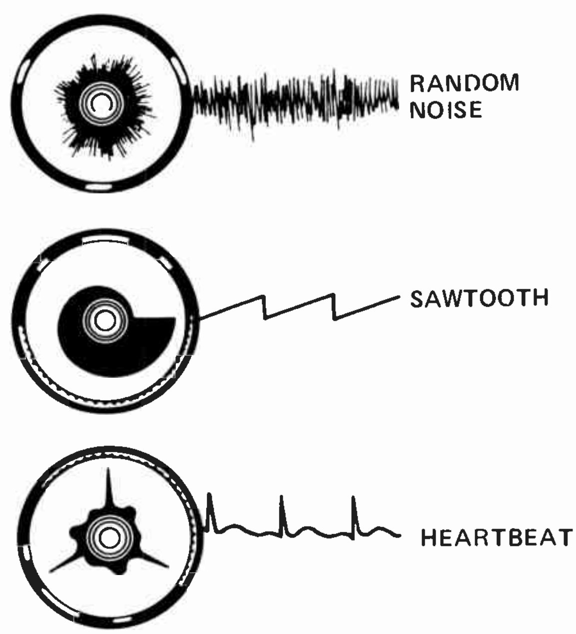


Fig.8. Complicated wave shapes may be generated by opto-mechanical methods. A specially masked disc, as it rotates, is used to vary the amount of light transmitted to a photocell.

ELECTRONICS -it's easy!

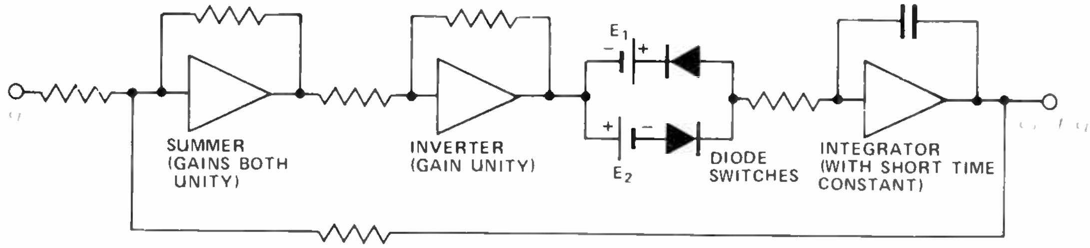
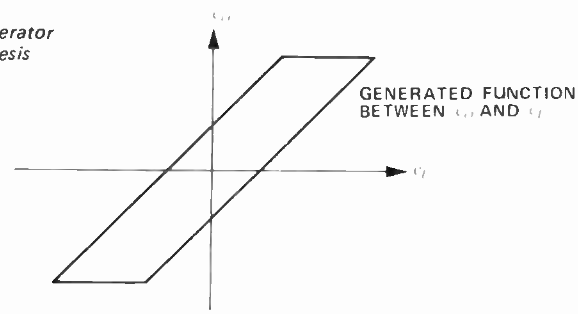


Fig.9. Special diode function generator for simulating backlash or hysteresis characteristics.



process uses two oppositely connected zener diodes placed across the output, as is shown schematically in Fig. 10. If the process is repeated two or three times a quite reasonable square wave results with fast rise-times and clean tops.

A second way, originating from a sinewave, uses a special circuit called a Schmitt trigger. In this circuit, another of the basic family of digital circuits, the output is either low or high depending upon whether the input-voltage level is above or below (respectively) a preset input level. Although the input can exist at any analogue level the output will always be only in one of two states. To produce square waves a sine-wave is fed into the Schmitt trigger which is set to trigger at the point where the symmetrical sinewave passes through zero. The result is a square wave, if the trigger level is exactly at zero, or rectangular if above or below. The advantage of this method is that very low-frequency square waves can be generated.

A typical Schmitt trigger circuit (Schmitt first described this two-state circuit in 1938) is given in Fig. 11. For the values given the output swings from its high value of 12 V to a low value of 1.0 V when the input passes through 1.8 V on the way up. The output swings back again as the input goes through 1.0 V on the way down. The difference between the up and down trigger levels is known as hysteresis (or backlash). Design methods exist that enable the trigger level, backlash and output swing to be set as required. To produce symmetrical square waves from a sine-wave source, with this circuit, the sine-wave would have to have its dc zero placed at 1.5 V. The 150 pF capacitor is added to reduce the impedance of the 1.8 k resistor at high frequencies, that is, whilst the circuit is switching. It is called a "speed-up" capacitor.

As well as being a convenient way to produce square waves, the Schmitt trigger also provides a mechanism whereby a hesitant effect is made positive. Take for example the case where daylight is used to operate a street lamp. As the light falls to around the operating point a relay-switch would chatter on and off with minor changes, until the average light level had fallen below the critical region. By adding a trigger circuit with reasonable amount of backlash, the relay is made to switch on the first time the light falls below the preset level. The relay cannot again change state without a significant rise occurring in light level.

The third way to produce square waves is to generate them using another digital circuit building-block, the free-running multi-vibrator or astable as it is also called.

There are three main types of multi-vibrator – astable, monostable and bistable. The astable automatically switches continuously between two states, thereby producing a square or rectangular wave signal. The monostable is normally in one state, and is triggered by an input signal into its second state. It stays there for a predetermined time before automatically toggling back again thus producing a fixed-length, single, square pulse. The bistable (or flip-flop), toggles from one state to the other with each successive input control pulse. It thus gives one output pulse for every two input pulses.

Each type can be used to produce "square" wave signals – the astable as a free running source, the monostable and the bistable as sources initiated by a train of pulses or changing levels.

Basically each type of multi-vibrator is formed from two common-emitter stages that are coupled together with impedances as shown in Fig. 12. This provides positive feedback from one stage to the other causing the device to always be in one state or the other – never between states for any length of time. This kind of impedance – resistors, capacitors or a mixture – determines the kind of positive feedback applied, and hence which of the three functions is generated.

Free-running astable – here the impedances are identical in both sides and are capacitors. Bias, or charging, resistors are added to each base as shown in Fig. 13. A suitable circuit for generating a 1 kHz square wave signal is given in Fig. 13a, Figure 13b is another circuit that will flash a small lamp at 1 Hz. Astable design is reasonably easy and is fully explained

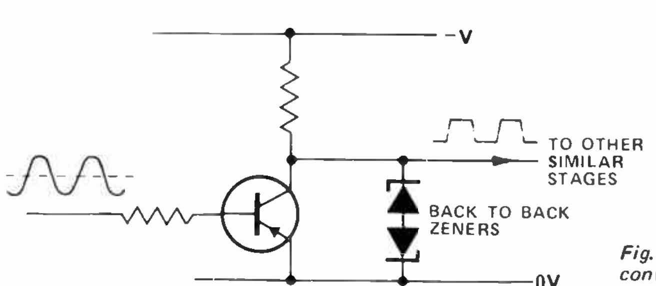


Fig.10. A sinewave may be converted to a squarewave by amplifying and clipping.

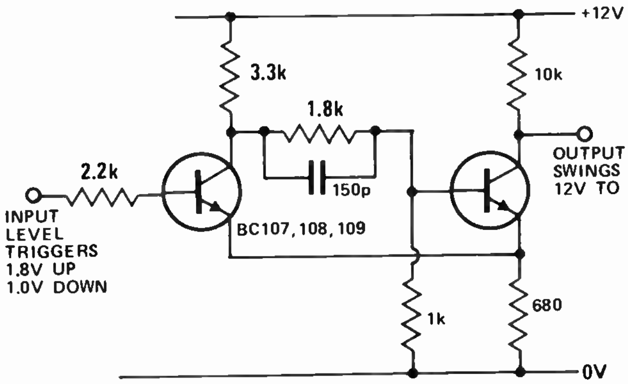


Fig.11. The Schmitt trigger circuit.

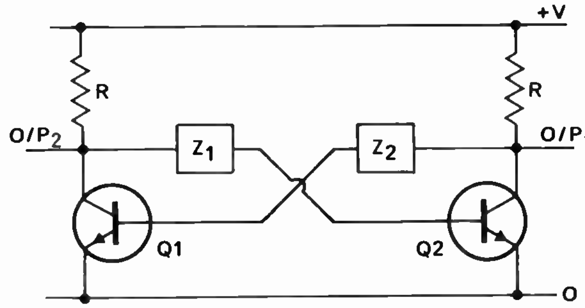


Fig.12. The basic arrangement of the multivibrator family of circuits.

in numerous books, especially those devoted to digital circuitry. The period of the square wave produced is given approximately by T = 1.4RC (refer Fig. 13a) from which the frequency f = 0.7/RC. The other main requirements needed is to ensure that the transistors are capable of handling the current demands of R_L when switched on. The output can be taken from either collector — the two are said to be complementary, that is, when one is high the other is low. Alternately the load can be wired directly into the collector circuit as shown in Fig. 13b.

If the base resistors are fed from an independent source the frequency can be varied by external means. This produces a voltage-to-frequency convertor, or, voltage-controlled oscillator VCO. Referring to the circuit given in Fig.14, the approximate period is

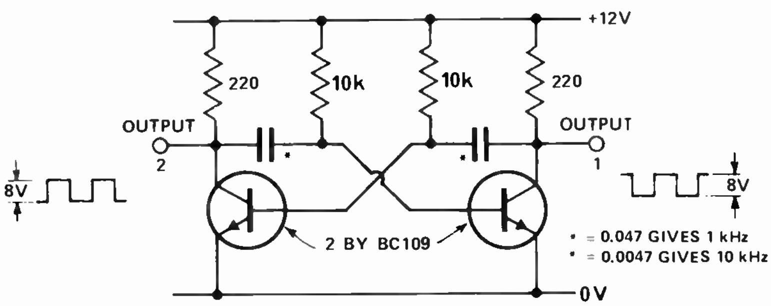
$$T = 2RC \ln \left(\frac{1 + Vcc}{Vbb}\right)$$

If the VCO is fed with a sawtooth signal the frequency output sweeps in synchronism — the well-known police siren sound.

Monostable or one-shot — if the requirement is for a train of pulses of uniform envelope height and width yet of variable repetition rate, then a monostable driven by a pulse train of the required frequency is the answer. A monostable has one transistor base connected as the astable above, the other is resistance coupled. Figure 15 shows a monostable set to provide a 20 μs wide pulse for a very wide variety of pulse inputs. Monostables are often used to reshape pulses back to a standard shape; they also serve to introduce a finite time delay because the initial input pulse can be regenerated later in time from the trailing edge of the monostable pulse. Thus the input pulse is delayed by the time duration of the monostable pulse width. An approximate value for the pulse duration is given by T = 0.7RC.

The circuit given in Fig. 15 features a second voltage rail. This ensures that the off-state transistor, which ever it is at any one time, is adequately switched off. It is, however, possible to design monostables that operate between only two lines — this has been the trend with semiconductor designs.

Bistable or flip-flop — this is the basic element used in digital computer counting as it produces an output pulse for every second input pulse, thereby dividing the input frequency by two. These have the two stages connected with resistors in both sides. Initially the circuit will start in either state — a set voltage is applied to the SET or RESET input thus conditioning the circuit to the initial state required. Input pulses or step voltages applied to both sides will cause the unit to change state at each input pulse. Figure 16 shows a typical simple design of flip-flop. The need for a negative voltage rail has been avoided by adding an emitter resistor.



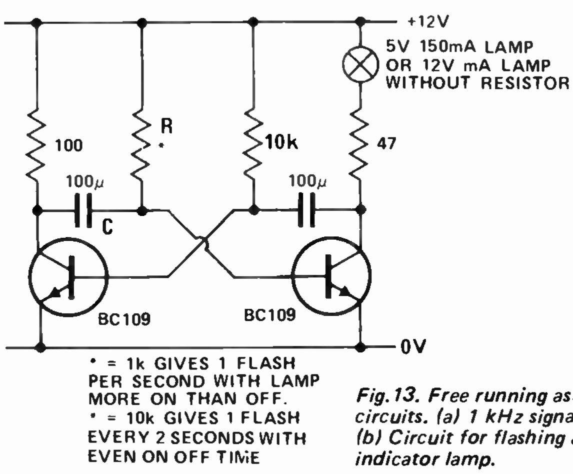


Fig.13. Free running astable circuits. (a) 1 kHz signal source. (b) Circuit for flashing an indicator lamp.

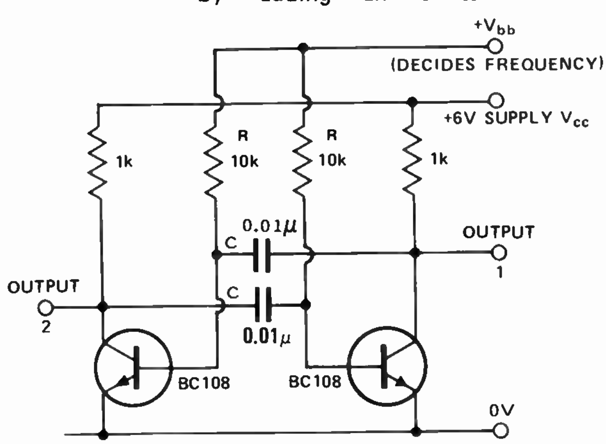


Fig.14. The output frequency of an astable can be varied by altering V_{bb} — it becomes a voltage-controlled oscillator VCO.

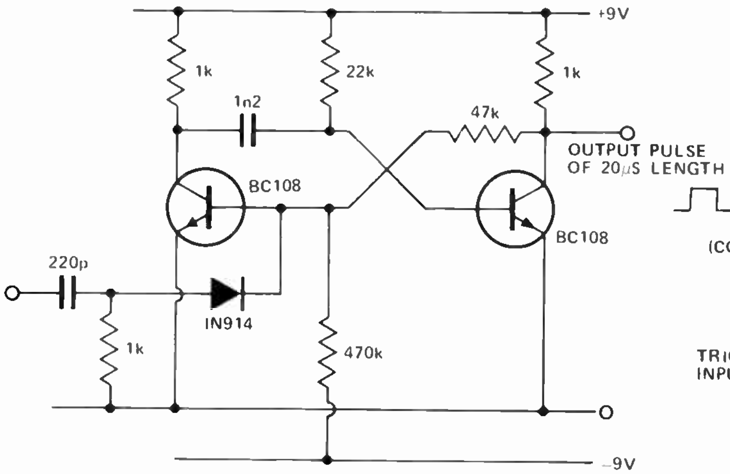


Fig. 15. Simple form of monostable – it produces a 20 microsecond wide pulse for each positive going input pulse.

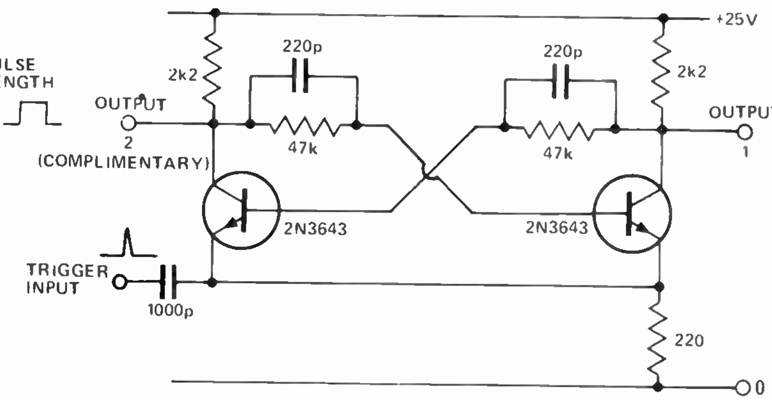


Fig. 16 Basic flip-flop or bistable circuit.

Triggering inputs, not shown, can be arranged to drive into the base, emitter or collector in order to provide the toggle action.

Designs come in two varieties – those in which the on-state holds the transistor well into saturation, and those in which the on-state holds the transistors non-saturated. The latter are capable of faster switching times but need much more careful design. We omit the design of the bistable for that is also well described in texts. It is a rare event, these days, for one to design flip-flops because they are now marketed in IC forms using over 10 transistors to achieve a much more stable and versatile unit at a price less than that of the two discrete transistors needed for the circuit shown in Fig. 16. Monostables and Schmitt triggers are also available in IC form. The latter effect can also be obtained using a linear op-amp with suitable connections.

PULSES

The logical follow on from square-wave generation is that of pulse generation. In Part 7 we described how LR and (more usually) CR networks could produce pulses by differentiating the square wave. The circuits for doing this are shown in Fig. 17. Figure 17d shows the standard differentiation circuit used. It produces signals, as shown in Fig. 17.1, in voltage form from (a) square-wave input waveform. The technique applies equally well for a single pulse requirement. Pulses produced this way alternate in sign. If both pulses are needed it is usually easier to produce two separate trains from anti-phase square waves selecting and combining the pulse polarities needed. This is easier in practice than attempting to invert every second pulse generated by a single differentiator circuit.

GENERATING EXTREMELY HIGH FREQUENCIES

The upper frequency limit for transistor operation is at present just approaching the gigahertz region (10^9 Hz); beyond this quite different techniques are employed. These techniques use devices such as magnetrons and klystrons, millimetre travelling wave tubes, masers and lasers. Figure 18 illustrates the frequency range over which each of these devices is useable.

In the earlier valve era it was very difficult to generate signals for radar needs (300 megahertz to 30 gigahertz) due to limitation of electron transit time, but late in the 1940's special self-resonating structures overcame this problem by using fields combined with valve concepts to 'bunch' electrons in a beam – typical such devices are magnetrons and klystrons. Such devices are still the best where high-power is demanded: microwave cooking ovens use magnetrons to generate kilowatt power levels for heating purposes.

The travelling-wave tube is another special electron device capable of UHF and microwave frequency amplification. In this tube an electron beam interacts with an electromagnetic wave travelling along the tube; again the electron bunching effect overcomes the transit time limitation. The design and use of these forms of generator are very specialised. Circuitry at such high frequencies is

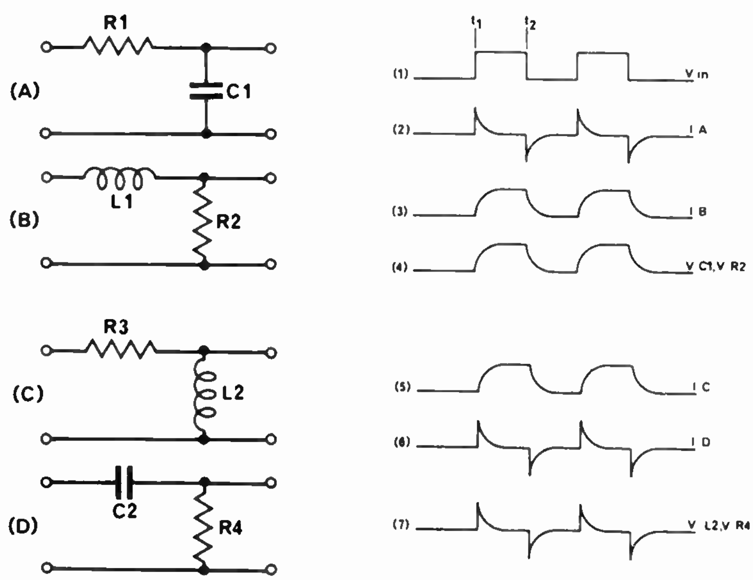


Fig. 17. Differentiation or integration may be crudely achieved by means of LR and CR circuits. Short pulses may be produced from square waves by differentiating with circuit (D).

not accomplished with wires but with waveguides that look more like a piece of precision plumbing than an electronic circuit.

Still higher frequencies can now be generated using various kinds of laser.

COHERENT RADIATION

A proper understanding of what is meant by 'coherent radiation' is essential to understanding why devices such as lasers are so important.

There are plenty of devices which produce radiation at super-high frequencies – eg, a hot soldering iron produces infra-red, an x-ray tube produces x-rays and a tungsten filament lamp produces visible light. But none of these sources produce coherent radiation. That is, their output consists of a multitude of separate packets of radiation which, although they may have the same frequency, have randomly different phase. Thus it is only possible to modulate such sources in bulk amplitude. It is not possible to modulate in frequency or phase on a cycle by cycle basis.

Devices such as lasers do produce *coherent* radiation. That is, the radiation is all in step, in terms of phase, and consequently can be modulated on a cycle by cycle basis.

Lasers can provide signal sources ranging from the far infra-red (10^{12} Hz) right through to x-rays (10^{19} Hz). At present no one device can cover this entire range. Some are tunable over a limited part of the spectrum, but most produce a single frequency within this spectrum.

Many laser sources are still in the exotic class and many problems remain to be solved. A major problem still outstanding is detection of such high frequencies. To date the highest frequency detected on a coherent wave by wave basis (that is, not as an incoherent bundle of energy as do most photo detectors) is 88 376 245 000 000 Hz. This is the frequency of the infra-red emission line of the now well developed helium-neon laser. It is just five times lower than visible light. Above that it is still not possible to detect the individual cyclic changes of the coherent sources that now exist.

REFERENCES

Most books on electronic circuits cover the design of generators. Try "Transistor Manual" – General Electric, 1969.

Application notes for ICs also show how to produce various waveforms.

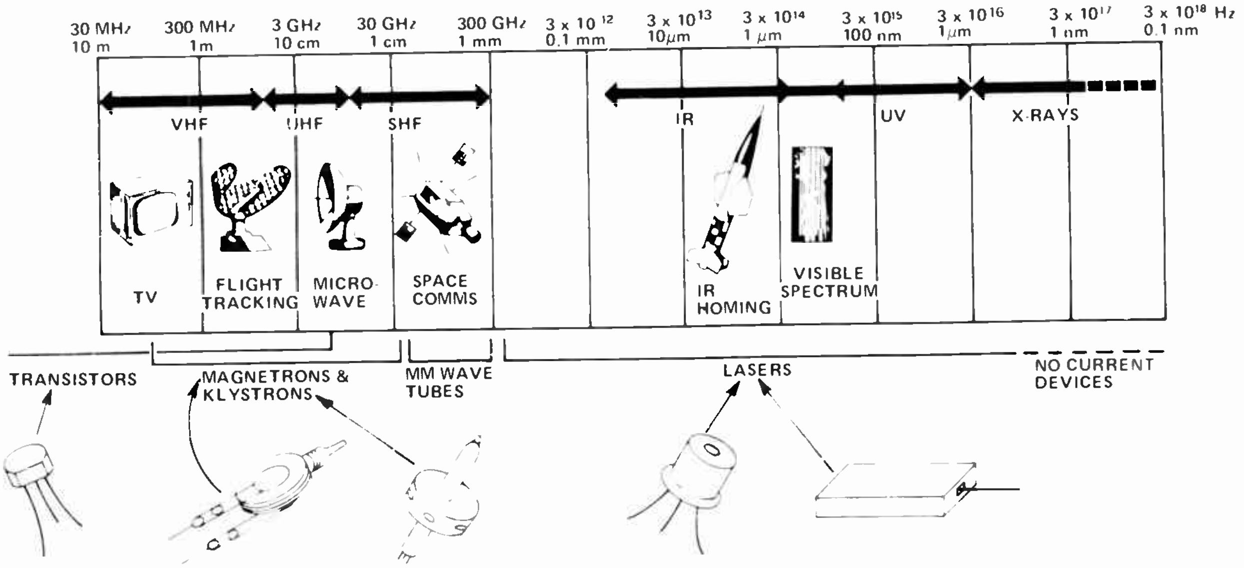


Fig. 18. Chart showing the various regions in the higher electromagnetic spectrum and the devices used in each.

Notes

19 All about electronic filters

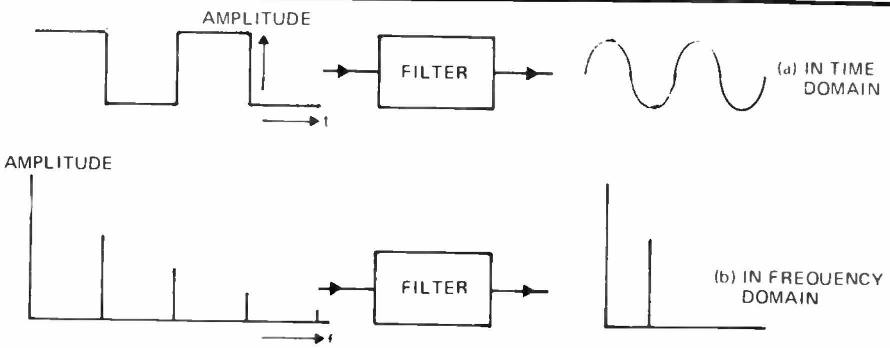


Fig. 1(a). A filter alters the frequency content of a signal. This means the wave shape is changed when displayed as an amplitude-time graph. (b). Using the frequency spectrum form of display the filter removes (or enhances) certain frequencies.

TO SEPARATE peas from boiling water, or dirt from engine oil, one must use an appropriate filter. When the term filter is used, in any discipline, the meaning is always the same — it is a device for separating or selecting something from an available mixture or range of things.

Filters are also extensively used in electronics where they are used to select a desired part of the range of frequencies which make up a particular signal. We have seen many examples of this throughout our course so far. For instance, in our discussion of multiplexed telephone systems (Vol.1) we saw how it is necessary to *separate* the various frequency channels and pass them to individual outlets. We also saw how an LC tuned circuit is used to *select* only one desired radio broadcast station from the many available.

Other examples of the use of filters are the crossover networks used in hi-fi speaker systems, to divide the audio bandwidth between two or more speaker drive units, the compensation stages in instrumentation control systems which improve performance by attenuating or enhancing relevant frequencies or the filters used to correct for the non-linear attenuation versus frequency which occurs with long-line telephone communications.

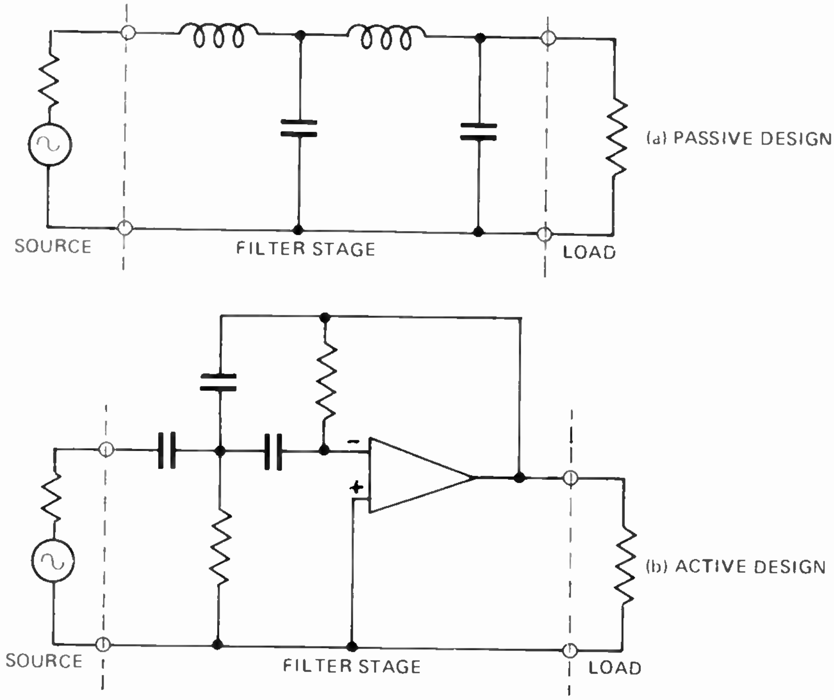


Fig. 2 (a). Passive filters use R, L and C components only. (b). Active filters incorporate active elements with passive elements to great effect.

ALTERING THE FREQUENCY RESPONSE

Electronic filters, in a general sense then, alter the frequency content of signals. Their action can be comprehended first by considering the stage as a unit that alters the amplitude/time shape of an input waveform. This concept is illustrated in Fig. 1a where a square-wave is filtered to remove all but its fundamental sine wave. Alternatively, filters may be thought of as devices that change the frequency spectrum. This is illustrated in Fig. 1b. Both concepts are correct, each finding use to suit different needs.

We generally think of filters as devices which change the amplitude of the signal with frequency. However, filters may also change the phase of the signal. In many applications the phase shift is undesirable and must be considered when making the selection of filter type.

Unlike other circuit blocks which are available as built up units, filters are generally made specifically for the task.

Many filters are extremely simple — varying from two components to (say) ten and the design procedures of most are easily found in texts. This is not, however, to say that filters are trivial and not worth learning about. Filter designs may be grouped into two main classes — those called passive filters (Fig. 2a) that use passive components only — such as resistors, capacitors and inductors; and those called active filters (Fig. 2b) that are based upon an op-amp using single or multiple path feedback loops. Design procedures can be quite complicated but because of the universal need for a few basic types of response, most design is now a matter of applying simple formulae or using graphs to arrive at the component values.

By way of interest the design philosophy of filters — or any network requiring a given frequency response — can proceed two ways. First, one can propose a network configuration and then mathematically analyse it to get the generalised formula. This is called network analysis. The alternative and more modern approach (in the last few decades, that is) is to start with a mathematical expression of the

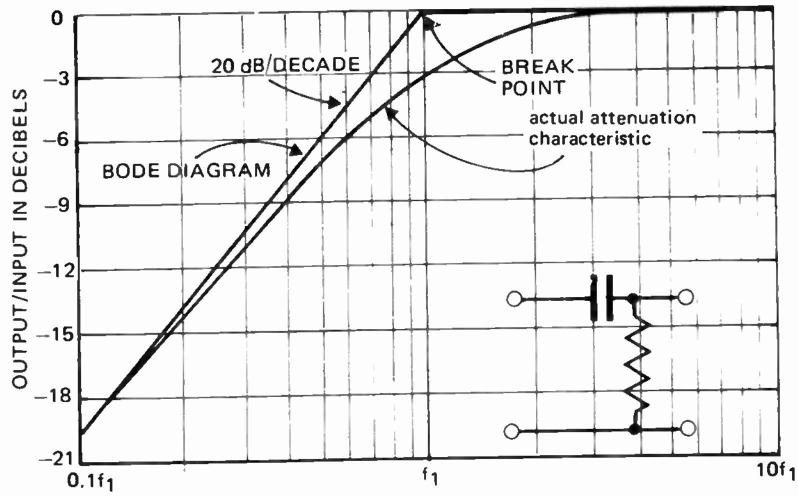


Fig. 3. Bode diagrams usually express amplitude (and phase) variation with frequency in terms of simplified responses consisting of straight lines turning at break points. The actual response will be more gradual near the breakpoint.

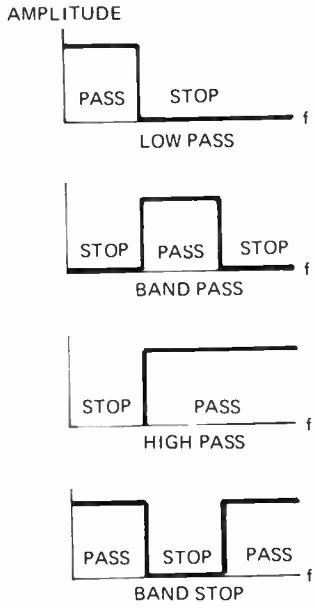


Fig. 4. Idealised responses of various categories of filter.

frequency response needed and, by using appropriate mathematical procedures, create on paper the circuit needed to provide such a response. This is called circuit synthesis. The latter method has a certain fascination because it provides the answer in a more logically direct manner than the cut and dried analysis process (although sometimes one ends up with a requirement for non-realisable circuit needs such as negative frequency!). On the other hand, however, synthesis requires mathematical ability and considerable experience.

In the following sections we will analyse a few of the more common filter stages.

THE BODE DIAGRAM

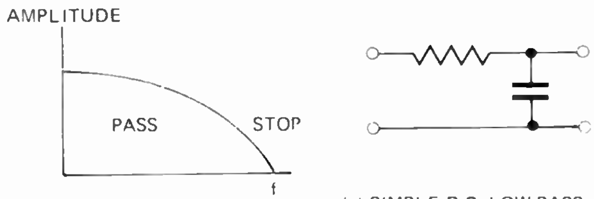
One of the, now classical, works on network analysis is a book "Network Analysis and Feedback Amplifier Design" by H. W. Bode published by Van Nostrand in 1945. Today Bode's work is mostly remembered by the graph which carries his name and relates the amplitude, or phase shift, to frequency for an amplifier, feedback system or a frequency modifying stage such as a filter. There is, at least, in principle, no distinction between the frequency response plots we have discussed to date and the Bode diagram. In practice, however, Bode diagrams are usually mathematical simplifications in that they are drawn with straight lines only, these lines changing direction at what are known as break-points and sloping at known rates.

The Bode diagram exemplifies the behaviour of a circuit as a tool, and is derived from mathematical knowledge of the system, not from actual tests. In truth, the linearization simplification is usually not far from reality, and we will meet Bode diagrams in our study of filters. Fig. 3 shows the difference between a Bode diagram and an actual response plot for an RC filter. The Bode diagram plots signal amplitude in decibels on a linear scale against frequency on a logarithmic scale.

TYPES OF RESPONSE

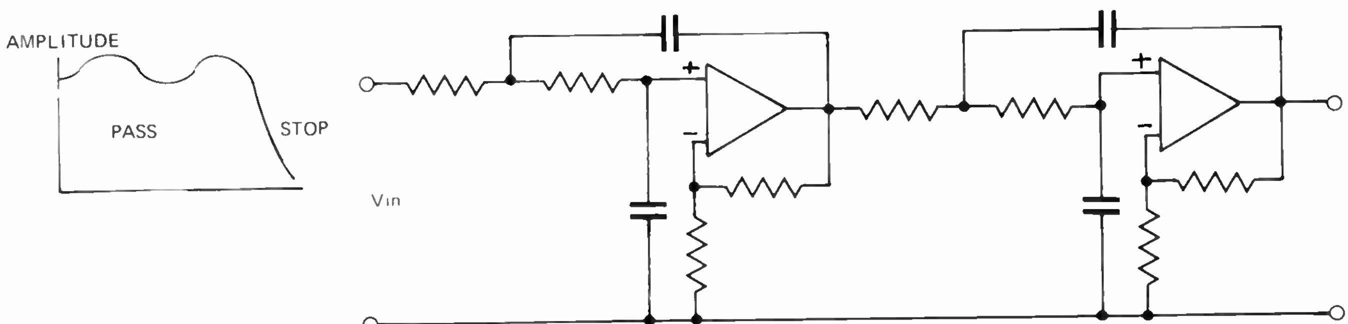
As with amplifiers, filter frequency responses are grouped into low-pass, band-pass and high-pass. Theoretically, ideal filters would have responses as shown in Fig. 4. There is also a constant need in electronic systems for a band-stop stage.

In reality it is impossible to obtain exactly square response curves. The response always rises or falls, within the transition region, with a rate of steepness that depends on the design used. A general rule is that the simpler the design (least number of components) the more gradual will be the transition. Also the more rapid the transition the more likely are effects of "ringing" encountered. Do not confuse these concepts of shape with amplitude-time wave shape graphs: these are amplitude (phase) — frequency curves. To illustrate this concept compare the two extremes given in Fig. 5. Figure 5a is for a most basic RC stage, Fig. 5b is for a response having rapid cutoff — a Chebyshev filter stage.



(a) SIMPLE R.C. LOW PASS FILTER STAGE

Fig. 5. As a general rule the more complex the filter circuit, the sharper the roll-off but the more variable the response in the passband region. (a) RC low pass stage. (b) Advanced Chebyshev stage.



(b) FOURTH ORDER CHEBYSHEV LOW PASS FILTER STAGE

ELECTRONICS -it's easy!

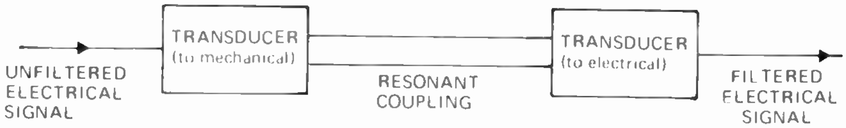


Fig. 6. By converting the electronic signals to mechano-acoustic form it is possible to make use of the extreme sharpness of mechanical resonant systems.

It is also worth noting that no filter is perfect, for frequencies are only attenuated relative to each other. If a signal appears at a high enough level at the input of a filter stage it will appear at a reduced level in the output and could be troublesome. Acknowledging this, the degree of attenuation chosen should be matched to the circumstances expected. It is pointless (and unnecessarily expensive) designing a stage to provide, say, 120 dB reduction of the unwanted frequency if it never reaches more than, say, 10 dB of the wanted frequency, apart from which an unwanted signal which is more than 60 dB down on the wanted one rarely causes problems.

DEFINING THE RESPONSE BANDWIDTH

As realistic filters fall short of being ideal there is no clear-cut point, where the response changes markedly enough, to use as the criteria for defining bandwidth. In some simple filters we could use the apparent position of the breakpoint but this would not hold for all filters.

The convention used is that the cut-off point is defined as where the response power falls to one-half of the passband value. Half power, expressed as a voltage change, is 0.707 of the passband voltage level which is -3 dB in decibel units. (Often called the '3 dB down' point.)

The bandwidth of bandpass (or bandstop) filters is, therefore, the frequency interval between the two cut-off points situated on each side of the bandpass (or stop) region. Bandwidth of a high-pass design has no real meaning as the frequency rises to infinity. Low-pass units have a band-width from zero frequency (dc) to the cut-off value.

In the case of complex designs the stated response often omits what happens at frequencies remote from the usual frequencies of interest. It is wise never to assume that, say, a bandpass filter only passes frequencies between the design points. It may well have "windows" much removed from that region. Additional stages are added in some system designs to exclude these effects.

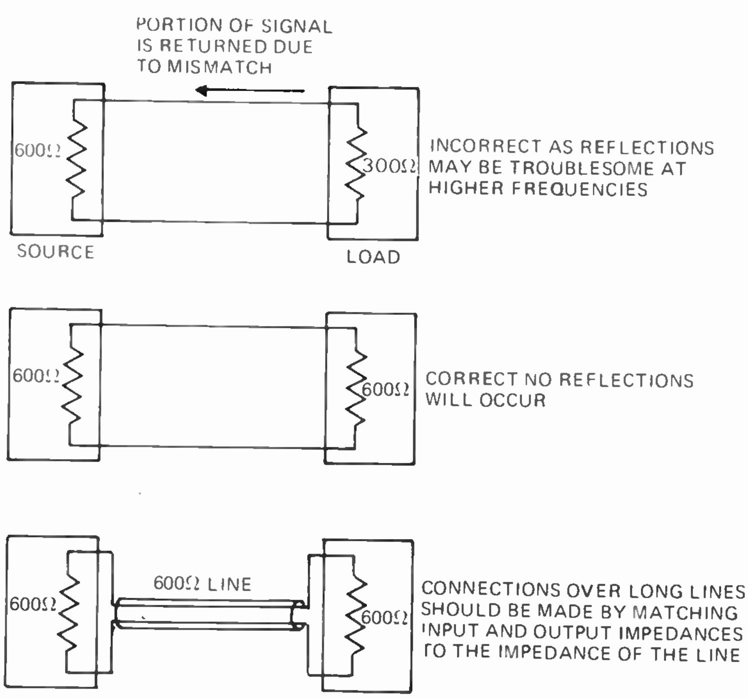


Fig. 7. Reflection of RF energy will arise if stages are not terminated into each other with the same impedance. Filter stages should observe this requirement.

Whereas the majority of filters used in electronic systems are made solely from electronic components there do exist circumstances where transduction to mechanical principles for filtering, and back again to electrical, are advantageous. One example is the use of tuned resonant reed filters, such as is depicted in Fig. 6, which exhibit extremely narrow band-pass characteristics.

Often the response of a bandpass is expressed in terms of its quality factor — that is the Q-factor of the peak. This definition was discussed when we dealt with resonant circuits earlier in the course.

THE EFFECT OF ADDING A FILTER

When the main purpose of adding a filter is to alter the frequency composition of signals it is not unexpected that the other effects brought about by its insertion might be overlooked.

As in any system changed by the addition of a cascaded 'box', the output of the preceding stage and the input of that following must be considered from the loading point of view. It is quite unrealistic to design a stage in isolation, unless the filter stage is adequately buffered, for the impedances connected to its input and output will alter the cut-off points — and hence different values will be required to achieve the designed characteristic.

The term 'Insertion Ratio' will often be encountered, it describes the ratio of output voltage with and without the filter, that is, the voltage Insertion Ratio = $\frac{\text{Vout (no filter)}}{\text{Vout (with filter)}}$

Expressed in decibels of loss we arrive at the term Insertion Loss = 20 $\log_{10}$ (Voltage Insertion Ratio). In practical cases, however, one may well design a stage to provide insertion gain (especially in active filter stages).

When matching a filter into a system it may be important to conserve power, voltage or current. To ensure maximized power transfer the input impedance to the filter must be of the same value as the output impedance of the stage before. Similarly, its output must be terminated into the same value. If voltage levels are to be maximized then the filter input impedance must be much higher than the output impedance of the driving stage. Current maximization requires the reverse relationship.

When the frequency of operation is high another problem becomes significant — that of reflections. When energy is launched into a network containing storage elements — a filter stage is such — some of the energy may be returned to the source which, in turn, may reflect it again, the final situation being that the net sum of all of these travelling waves of energy cause excessive power losses in the line (and distortion). This effect is very pronounced in radio-frequency transmission lines.

The extent to which a reflection occurs is decided by the degree of difference in the impedances seen in both directions at a system block junction. If a filter is terminated into the source with the same impedance in both directions there is no mismatch and no reflection occurs. This concept is depicted in Fig. 7.

As the two impedances differ in magnitude so does the amount of signal reflected. A similar situation applies at the output of the filter.

Mismatch terminations begin to generate noticeable spurious signals this way from megahertz frequencies upwards. This is the reason why wide-bandwidth amplifiers, such as videoamps, must be designed with output impedances that match the feeder cable. Coaxial cable can be shown to have a characteristic impedance set by the ratio of size and spacing of its conductors. It is invariant with length of cable. Typical coaxial cables have impedances of 50 or 75 ohm. Alternatively another kind of cable having two wires with a fixed separation between them may be used. Such transmissions lines have typical impedances of 200, 300 or 600 ohms. Whilst on this subject, one way of locating open-circuit and short circuit faults in cables is to send a sonic pulse (these travel much slower than electromagnetic waves) down the cable – timing the arrival of reflected pulses produced by the gross mismatch that exists at the fault.

Filter stages, as said before, also introduce phase shifts. A sine-wave input will appear at the output shifted in time by some fraction of the electrical cycle. In the compensation networks of feedback controllers phase shift must be carefully controlled, for a wrong value of phase shift may cause the system to become unstable. That is, if the phase shift approaches 180°, the feedback becomes positive, instead of negative, and the system oscillates.

PASSIVE DESIGNS

THE RC FILTER

The simplest passive electronic filter is the RC network set to act as a low-pass or high-pass stage. The two alternatives are shown in Fig. 8. In Fig. 8a it is easy to see that at low frequencies the capacitive reactance is very high and the output is the same as the input, provided the load

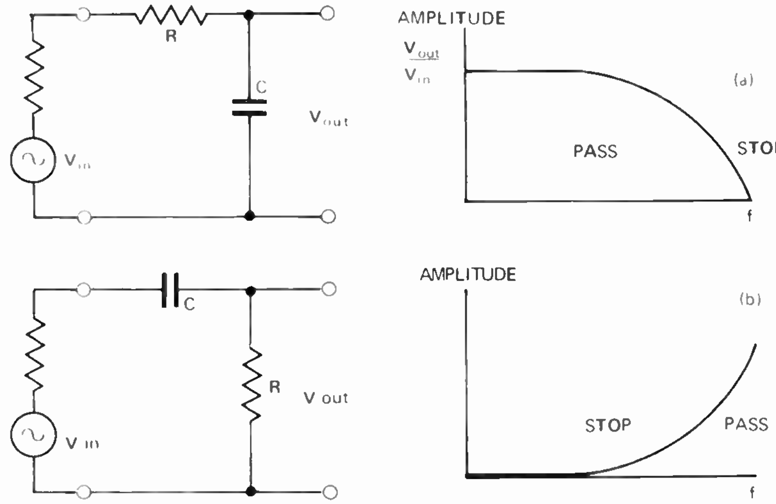


Fig. 8. Basic RC filter stages (a) low-pass (b) high-pass.

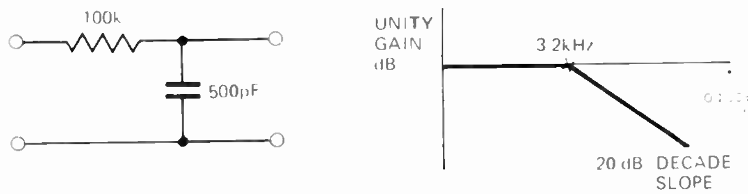


Fig. 9. Bode diagram for low-pass RC filter in which source and load are not significant.

impedance connected is significantly higher than the value R. As the frequency rises X_c decreases, lowering the output voltage. The reverse situation applies for the high-pass unit.

Mathematical analysis shows that the response plot – the Bode diagram – for these can be constructed by recognizing that there is just one break point and that the response falls away at 20 dB/decade change in frequency (ie 6 dB/octave). An octave change corresponds to 2 : 1 frequency ratio; a decade change is a 10 : 1 ratio. The jargon used is that the response rolls-off at the stated rate. Regardless of the values of RC chosen the roll-off rate stays the same. The break point occurs at $f_c = \dfrac{1}{2\pi RC}$.

To illustrate this consider the construction of the Bode diagram for a low-pass filter with R = 100 kilohms and C = 500 pico-farads. The break point occurs at

$$fc = \frac{1}{6.28 \times 100 \times 10^3 \times 500 \times 10^{-12}}$$

and it slopes downward from there at 20 dB/decade to give the plot shown in Fig. 9.

This much may seem almost trivial and, indeed, it is over-simplified. In

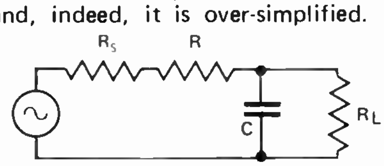


Fig. 10. Practical RC filter designs should allow for source and load resistances.

practice there will be a source and a load impedance connected to the filter terminals. Fig. 10 shows the practical case in general.

It is also not hard to reason out what happens when the source and load impedances are taken into account for Rs is in series with R and R_L is in parallel with C. By expanding our mathematics we find that the formula becomes

$$f_c = \frac{1}{2\pi\ [R_S + R)\ //\ R_L]\ C}$$

Hence, if the stage is not buffered the breakpoint can be quite different from that arrived at from the time-constant of the filter alone. For example if load and source impedances are both 1 k in our previous example the breakpoint changes from 3.2 kHz to 2.66 kHz. Further, the stage will introduce attenuation: the gain in the passband becomes

$$\frac{Vout}{Vin} = \frac{R_L}{R_S + R + R_L}$$

$$\text{for our example} = \frac{1000}{1000 + 5000 + 1000}$$

$$= 0.4$$

By use of appropriate values of source and load resistance it is possible, therefore, to set the attenuation and draw an appropriate Bode diagram.

The high-pass RC filter is considered in the same way – to arrive at

$$f_c = \frac{1}{2\pi (R_S + R\,//\,R_L)\,C} \text{ and}$$

$$V_{out} = \frac{R\,//\,R_L}{R_S + R\,//\,R_L}$$

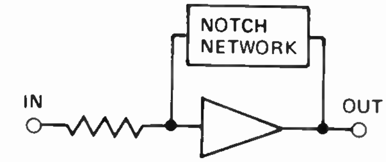


Fig. 12. Notch acceptance can be provided by using a notch-rejection circuit in the feedback of an op-amp.

for the practical case where source and load impedances cannot be ignored.

The observant reader will probably have realised that an amplifier stage with capacitive coupling has an equivalent circuit that is a combined highpass and lowpass filter with gain added between. The high-pass response arises from the coupling capacitor and the stage input impedance, the low-pass response from the output impedance and the stray capacitance existing to ground.

It is possible to combine a low-pass RC stage with a high-pass stage to arrive at a bandpass filter. These, however, are not particularly selective bandpass filters because of the relatively poor roll-off slopes (20 dB/decade). Further, if the bandwidth required is small, the two stages interact producing a non-constant passband gain. To obtain a satisfactory design it is important to ensure that the second stage resistance (the shunt of the high-pass stage) is at least ten times that of the first (the series resistance of the low-pass stage). Also the two break points should be at least a decade apart.

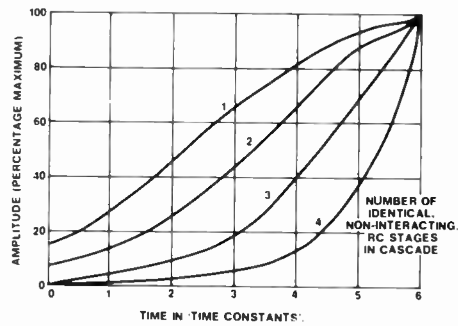
TRANSIENT RESPONSE OF RC STAGES

A sudden signal disturbance to a filter stage will have its frequency content altered giving a different time response at its output from that applied. A step function applied to a single RC stage was discussed in Chapter Six.

There it was shown that the amplitude rises as an exponential characterised by a 'time-constant'. It is shown here in the diagram.

If RC stages are cascaded, and provided each does not affect the operation of the previous one (called non-interacting) the output to a step input will be as shown. Note the effect of cascading is to slow up the initial rate of change.

As the degree of interaction increases, the response becomes even more slow to begin with. It is not possible for the cascaded set to ring.



RC NOTCH FILTERS

Some applications call for rejection of a narrow band of frequencies, the reduction of 50 Hz or 100 Hz noise, for example. A very effective, yet, inexpensive technique makes use of a type of Wheatstone bridge which requires only resistors and capacitors and yet provides very sharp roll-off.

The Twin-T or parallel-T notch filter is such a circuit and is shown in Fig. 11. (It can be redrawn as a more-obvious bridge circuit and comprises two T circuits connected in parallel). At high or low frequencies it is easy to see that the capacitances either go to low or high reactances providing in both instances a virtually unaltered signal level through the stage. At the balance point, of a twin-T bridge, there exists a frequency — the so-called notch — at which the output falls very nearly to zero. This occurs for the circuit of Fig. 11 at

$$f_c = \frac{1}{2\pi RC}$$

Loading will reduce the depth of the notch.

In some applications it is desirable to be able to tune the notch to varying frequency values. In the Twin-T design this requires that all three resistors (or capacitors) be varied simultaneously. A ganged multi-unit potentiometer or capacitance bank is used.

Other forms of bridge filter exist, each having its own particular feature. No simple RC circuits exist that exhibit the reverse characteristic of the notch filter — that is spike acceptance of a particular frequency. This reponse however, can be provided by using a notch-filter as the feedback impedance in an op-amp that is set up as a simple inverter. This is shown in Fig.12. In this way the gain of the stage rises rapidly with increase in effective feedback resistance at the notch frequency.

IMPROVING THE ROLL-OFF

RC filters, apart from notch circuits, cannot provide much selectivity between signals due to their poor 20 dB/decade rolloff. This slope can be improved by cascading stages but this is not a preferred method for there exist other more economical designs.

The next stage of complexity is to use designs combining inductors and capacitors: no resistors are needed. That these provide improved roll-off is to be expected for we have seen earlier in this course that a resonant circuit can provide very sharp responses. By way of example a single stage LC filter can provide at least 12 dB and up to 25 dB/octave rolloff compared with only 6 dB/octave for an RC stage, and furthermore methods have been established (discussed in next part) that enable these to be cascaded without difficulty — a four stage unit can achieve 100 dB/octave rolloff! It is even possible to 'peak up' a specific frequency in the passband. In the next part we will also explain the virtues of adding amplifiers to form active filter circuits.

FURTHER READING:

Electronic Instrumentation Fundamentals" — A.P. Malvino, McGraw Hill, 1967. (This has a fine chapter on RC and LC filters which is not complicated with circuit analysis and mathematical symbology).

"Radiotron Designers Handbook" — F. Langford-Smith, A.W.V. 1955 and subsequent editions. (Covers RC filters and supplies response graphs in a small section).

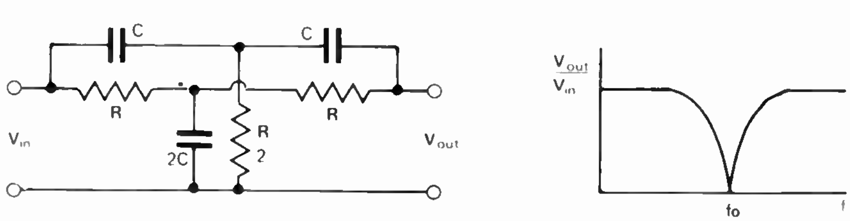


Fig. 11. The Twin-T notch filter provides very narrow rejection of a particular frequency.

20 More about filters

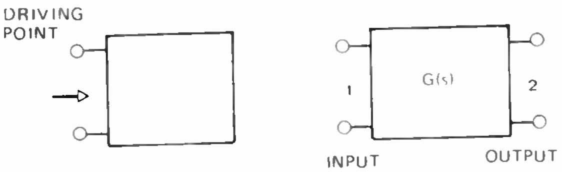


Fig. 1. Basic filter system block diagram.

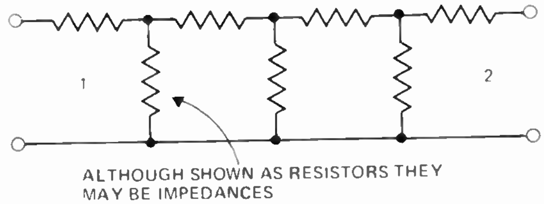


Fig.2. Ladder network.

AS WE saw in the previous section resistor-capacitor filters can only provide roll-offs of 6 dB/octave (20 db/decade). On the other hand combinations of inductors and capacitors can provide much steeper rolls-offs and a response at the turn-over point which can be tailored to a desired shape.

The variety of LC component combinations that can be employed is great indeed and, to the uninitiated, the design of such filters can seem to be very confusing. However, circuit analysts have established design procedures which enable a filter having any practical characteristic to be designed in a logical, formalized manner. The method is based on the use of cascaded basic sections.

TWO-TERMINAL PAIR NETWORK CONCEPTS

As we have seen at various times in the course so far, filters can be circuits having just two terminals – a resonant circuit for example, or they can have two input and two output terminals – the so-called two-terminal pair networks. (The RC filter is of the two-terminal pair kind). The two different types are illustrated as system blocks in Fig. 1. Note that it is conventional to show input on the left and output on the right.

As said before many possible circuit configurations exist for filters, and the designer has to make a compromise between using a simple arrangement of many components that can be easily handled mathematically, or, a few components in a more complex network that cannot be treated by general formulae. Here we will examine the approach based on grouping numbers of simple and similar networks, to obtain the desired response, by the methods originally proposed by Zobel in 1923.

The simplest type of network is the LADDER, as illustrated in Fig. 2, the defining feature being that it has a common line. When the lower line also includes impedances (resistor elements are used to represent what are usually reactances) the network is called a LATTICE; these are much harder to design and are less commonly used. Let us examine how a ladder network is broken down into even more basic structures.

By convention the series elements of a ladder are labelled Z_1, and the shunt elements as Z_2. These elements will be either capacitors or inductors and, it is assumed that the filter is driven from, and drives into, pure resistances.

Within the ladder arrangement, shown in Fig. 2, can be seen three basic building structures – called the L section (inverted L to be absolutely correct), the T section and the π section. The three are shown separated in Fig. 3.

In Fig. 4 we see how standard T or π sections can be connected to provide the same effective ladder network. Conversely a ladder network may be subdivided into standard T or π networks by breaking the values up as shown.

The interesting and quite vital point is that the T or π stages have the same input and output impedance. That is they are *symmetrical*. The L section, however, is *unsymmetrical* in that input and output terminal pairs are not interchangeable. Two L sections in series will produce a T or a π section.

When two identical T or π sections are cascaded they are matched into the same impedance – maximum energy is transmitted and no reflections occur. Each terminal sees an image of itself, this property giving the name *image – parameter* design to this filter design method.

CONSTANT-K FILTERS

Even though a quite simple configuration has been used there can still be a wide range of combinations each with complicated mathematical solutions.

By introducing another assumption we can make some headway toward realising a wide range of characteristics with a reasonable degree of mathematical simplicity. This assumption is that $Z_1 \cdot Z_2 = R_0^2$ where R_0 is a true resistance called the characteristic resistance. (This may seem strange but the multiplication of capacitive reactance with inductive reactance yields just that). Hence Z_1 and Z_2 must be a combination of capacitor and inductor giving us

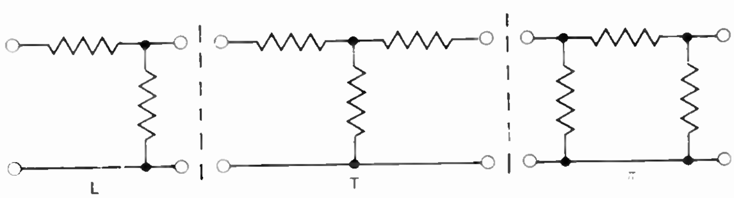


Fig. 3. The basic filter section.

ELECTRONICS -it's easy!

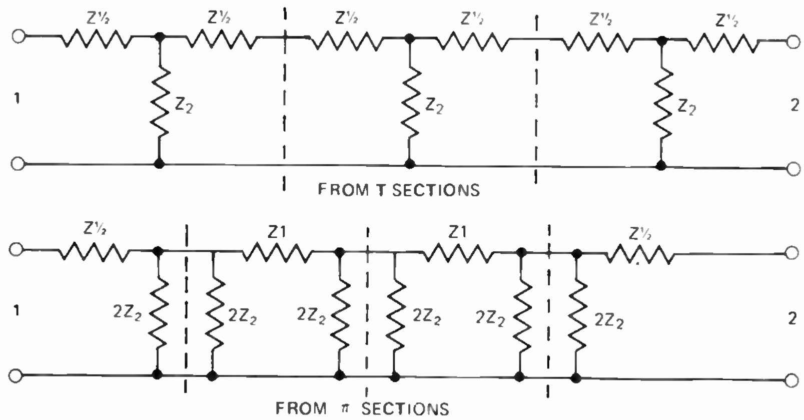


Fig.4. Building up the ladder network from basic sections.

equivalent stages with L and C proportions as shown in Fig. 5. The rule holds true for an L section provided we treat full shunt or series reactance as 2L or 2C.

The name constant-K arose from the original terminology where Zobel, in 1923, used K instead of our now accepted R_0. Filters designed to this rule are hence called *constant-K* filters.

Regardless of whether the stage is designed to be high pass or low pass – the cut off frequency will be the same, that is, at the resonance point of the LC values of the standard equivalent L section.

That is cut-off frequency $f_c = \dfrac{1}{2\pi\sqrt{LC}}$

For example in the π section of Fig. 6 the equivalent L section networks have L of 1 mH and a C of 0.5 microfard.

That is cut-off frequency

$$f_c = \frac{1}{2\pi\sqrt{10^{-3}\text{x}0.5\text{x}10^{-6}}}$$

$$= 7.1 \text{ kHz}$$

Also from $Z_1 \cdot Z_2 = R_0^{\,2}$

characteristic resistance $R_o = \sqrt{Z_1 Z_2}$

However the capacitive reactance must be written as a reciprocal and in Fig. 6 this is Z_2. Hence:–

$$R_o = \sqrt{\frac{Z_1}{Z_2}}$$

$$= \sqrt{\frac{10^{-3}}{0.5 \times 10^{-6}}} = 45 \text{ ohms}$$

Thus we see that the source and load impedances used with this network must be 45 ohms, if maximum power is to be transferred, and the network is a low-pass stage having a cut-off frequency of 7.1 kHz.

If L and C were reversed the filter would have identical R_0 and f_c but it would now be a high-pass stage.

An important feature of image-parameter design is that image-matched stages can be cascaded without altering the cut-off frequencies or the characteristic resistances. Each additional stage improves the roll-off, thereby giving a powerfully reliable way to obtain the desired rapidity of attenuation without having to re-design the whole system as extra stages are added.

It can be shown that the attenuation, a, in the stop band, expressed in decibels, is a dB = 9.7 n α where n is the number of standard –T (or standard – π) sections cascaded, and α is 2 Cosh^{-1} f/f_c. Cosh^{-1} means the cosh function (a hyperbolic trigonometric expression) whose ratio is f/f_c. As most readers will not be familiar with the coshine function Fig. 7 gives the relationship between values of α and frequency ratios normally encountered. Note that either f/f_c or fc/f is used depending on whichever gives a value greater than one.

The following example shows how a constant-K filter is designed to given response requirements.

The basic design formulae are:

$$L = \frac{R_o}{2\pi f_c} \quad C = \frac{1}{2\pi f_c R_o} \quad n = \frac{a \text{ dB}}{8.7\,\alpha}$$

The values given at the start will be R_0, f_c, α and a dB. We need to establish, in the synthesis situation, the values of L, C and n. The necessary configuration is established by logical deduction of the appropriate placement of components in the sections.

Example : Design a high-pass filter having a cut-off frequency of 10 MHz and a signal attenuation of 100 dB at 5 MHz. The characteristic resistance is to be 50 ohms in order to match the existing system into which the filter is to be fitted.

$$L = \frac{R_o}{2\pi f_c} = \frac{50}{2\pi\, 10\, 10^6} = 0.769\ \mu\text{H}$$

$$C = \frac{1}{2\pi f_c R_o} = \frac{1}{2\pi\, 10\, 10^6 \cdot 50} = 318\text{pF}$$

To determine α

$$\frac{f_c}{f} = \frac{10.10^6}{5.10^6} = 2$$

From the chart α = 2.64.

Number of stages required, $n = \dfrac{a \text{ dB}}{8.7\alpha}$

$$= \frac{100}{8.7 \times 2.64} = 4.35$$

We cannot however have 0.35 of a stage and therefore must use five stages to obtain at least 100 dB attenuation at 5 MHz.

The formulae for L, C are for the basic section so we have to halve values accordingly, giving us the circuit of Fig. 8. We could just as correctly divide the system into a π rather than a T configuration. Design of a low pass stage proceeds in just the same way.

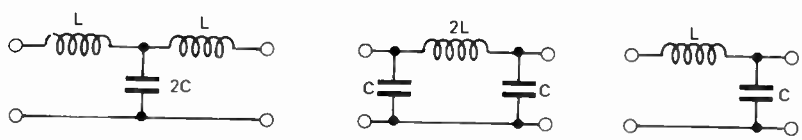


Fig.5. Convention used in this explanation of LC filter design.

Fig.6

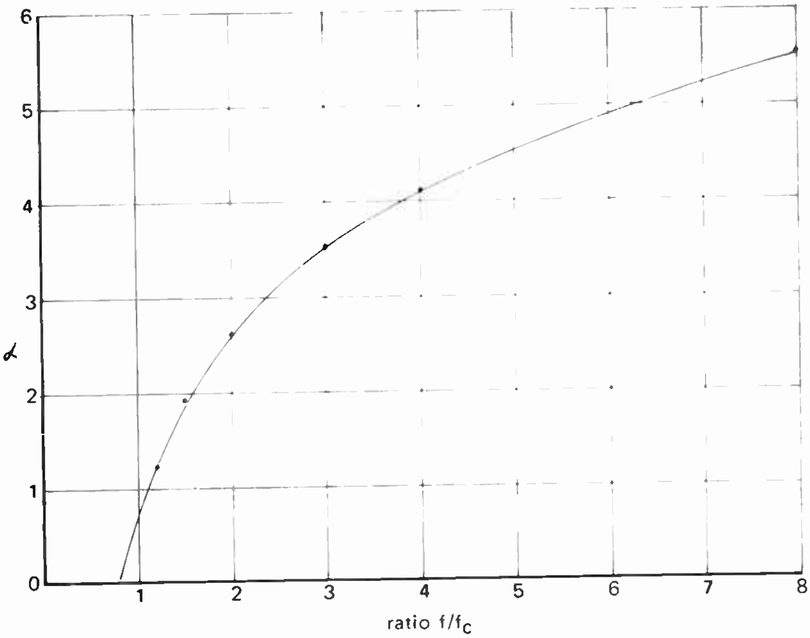


Fig.7. Chart relating value of ∝ and frequency ratio.

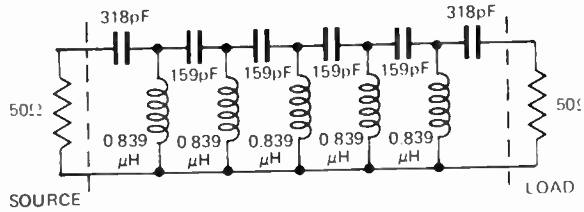


Fig.8. 5 section, constant -K, high-pass filter with cut-off of 10 MHz.

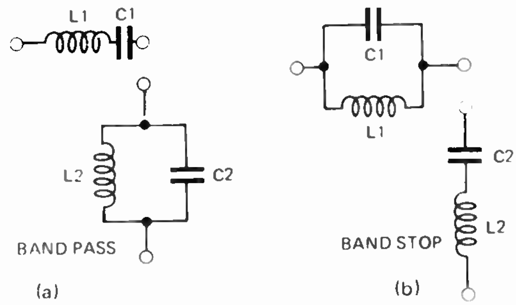


Fig.9. Basic elements of band-pass constant -K filters.

The design of band-pass and band-stop stages is more complicated going beyond the scope of this course. Suffice to say that the components in the arms now become series or parallel resonant combinations. The basic L-section for a constant-K band-pass is shown in Fig. 9a and the basic L-section for a band-stop in Fig. 9b. Readers who wish to pursue these can obtain guidance from the reading list.

M-DERIVED SECTIONS

As can be expected the simplifying assumptions made in the constant-K design, to obtain a reasonably straight forward mathematical procedure, also create practical disadvantages. The first defect is that the image impedance does not remain constant and varies in such a way that noticeable reflections occur near the cut-off points. The second defect is that the roll-off is slow just near the cut-off point: it is adequate further away from that point.

Zobel's concept to overcome this involved additional cascaded stages that, in effect, flatten out the passband response and sharpen up the cut-off point attenuation. These extra stages are called M-derived sections: one is usually added on each end of the ladder designed by the constant-K method.

We can only give an example circuit to illustrate this — Fig. 10. Although the formulae for arriving at the values are simple they must be applied with great care, the user having adequate experience in order to know the correct procedures. Again we must leave it to the reader to take this up in more specialized texts. The design of a full M-derived system requires extensive effort and training and is much more the task of a professional circuit designer than the reader for which this course is designed. The most extensive application of M-derived filters has been in communications engineering — telephones, telegraphy and multiplexed radio links. Voluminous books have been compiled that list tables giving values for chosen designs. Special computer programmes have also been developed to provide automatic constant-K and M-derived section filter designs.

ACTIVE FILTERS

The basic active RC building blocks

Passive filter designs had reached their present sophistication as much as 50 years ago and in the absence of anything markedly better they continued to be the most used design until the mid 1950's. Amplification was added to make up for the attenuation that usually is experienced with passive designs.

With the introduction of reliable and less power-thirsty solid-state amplifiers in the late 1950's came the so-called active-RC filters. These combine an operational amplifier with passive RC components thereby producing filtering action more efficiently than the more obvious passive network followed by an active stage. One very valuable feature is that the effective value of, say, a capacitor can be multiplied up many times on its actual value thereby saving space and enabling designers to build circuits needing large effective values. It is also possible by active filter design to avoid the need for inductors in filter circuits. Inductors are best left out, if possible, for they are usually bulky, expensive and very lossy — they are nowhere as "ideal" as capacitors. They also are non-linear in operation and can be saturated by excessive current.

The basis of an active RC network is more often than not a reasonable quality operational amplifier set up to provide one of the following four basic circuit concepts.

1. The high gain (60 dB or more)

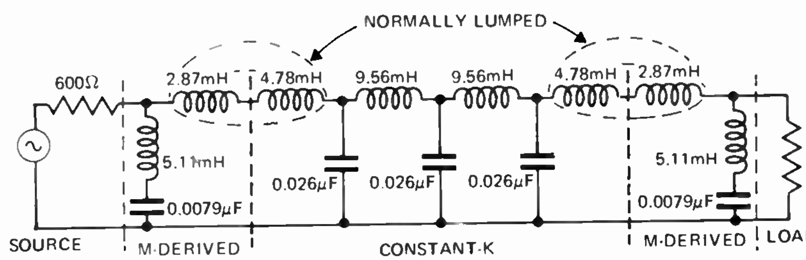


Fig.10. Example of constant -K filter with M-derived end sections.

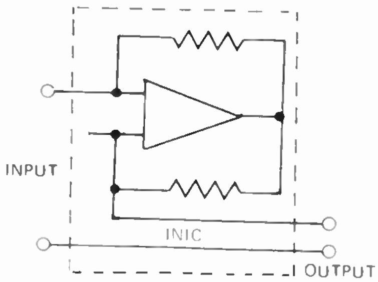


Fig. 11. Realising an INIC with an op-amp.

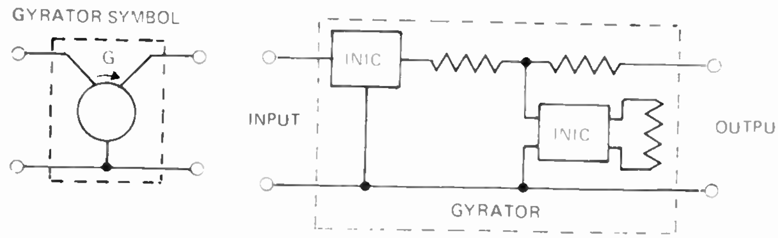


Fig. 12. The gyrator.

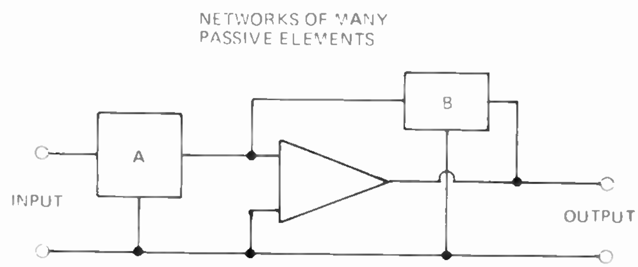


Fig. 13. Single-loop feedback active filter schematic.

voltage amplifier with close to infinite input impedance and almost zero output impedance – in short, the normal mode of an op-amp as we have discussed previously.

2. The low gain (20 dB or less) voltage amplifier, also referred to as a voltage-controlled voltage source or just VCVS.

3. The negative-emittance or negative impedance converter NIC. This is a most interesting system block for it enables positive value capacitance or resistance (that obtained with normal capacitors and resistors) connected at its input to appear as negative value capacitance or resistance at its output. It enables circuit designers to physically build circuits requiring non-physical negative capacitors and resistors. (INIC indicates an ideal current-inversion NIC, and VNIC indicates an ideal voltage – inversion NIC). A typical realisation is shown in Fig. 11.

4. The Gyrator. This is another intriguing unit for the output appears as the reciprocal of any impedance connected to its input. Thus a capacitor at its input appears as an inductor at the output. The gyrator, therefore, eliminates the need to use physical inductors and what is more, can provide more "ideal" inductors than real units. It can be realised using op-amps as shown in Fig. 12.

With these four basic possibilities available the circuit designer is rarely restricted by having synthesised a circuit needing non-physical components.

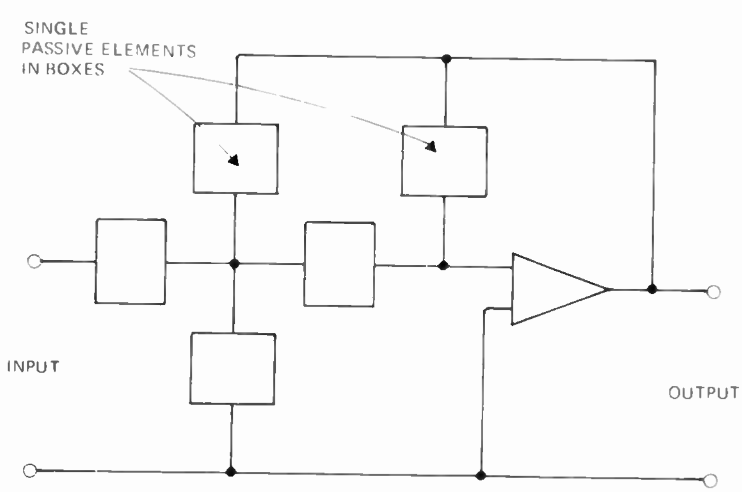


Fig. 14. The multiple-feedback-path active filter.

CHOOSING AN ACTIVE FILTER DESIGN

Given the above four system blocks it is possible to produce an incredible variety of active filters. As with advanced passive designs, few people have enough training to be expert active-filter designers. Here we can only give a guide that provides the necessary awareness of what to look for, along with words of caution as to what it is reasonable to expect from an actual active-filter design.

The voltage amplifier can be used in its simplest conceptual way with a *single-loop feed-back path* (SFP) as shown in Fig. 13 – remember how we have already seen that an op-amp integrator acts as a low-pass filter and how a notch-rejection filter, introduced into the feedback path, produces a notch-acceptance response instead.

Alternatively, we can make use of *multiple feedback paths* (MFP) as depicted in a general sense in Fig. 14, the design using minimum component count. These, somewhat surprisingly, use fewer passive elements than single-loop circuits. For this reason this form of active filter is the configuration most often used.

The other options open to us are to use an op-amp set up as either a controlled source with added elements – see Fig. 15, or as the negative-impedance convertor shown schematically in Fig. 11. These can offer certain advantages over the voltage-amplifier designs but suffer some disadvantages. NIC devices, for instance, do not give the ideal zero output impedance. Stages must be buffered to retain designed

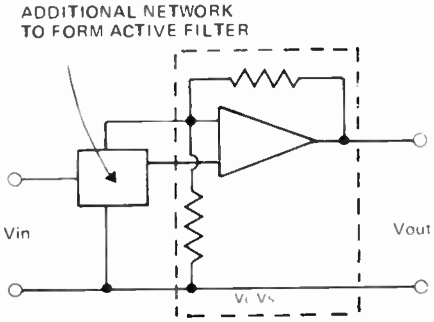


Fig. 15. Active filter using controlled-source approach.

	Realization Technique			
Property	Infinite-Gain Single-Feedback	Infinite-Gain Multiple-Feedback	Controlled Source	Negative-Immittance Converter
Minimal number of network elements	-	+	+	+
Ease of adjustment of characteristics	-	0	0	+
Stability of characteristics	+	+	-	-
Low output impedance	+	+	+	-
Presence of summing input	+	-	-	-
Relatively high gain available	+	-	+	+
Low spread of element values	+	-	+	+
High-Q realizations possible	+	-	+	+

\+ indicates the realization is superior for the indicated property

0 indicates the realization is average for the indicated property

\- indicates the realization is inferior for the indicated property

Fig. 16. Comparison table for various kinds of active filter realisations (from Burr-Brown handbook).

performance, tor example, when they are cascaded to obtain higher orders. On the good side is the small number of passive elements needed. Fig. 16 compares the four alternatives showing that no one type is exclusively the best choice.

At this stage we can only suggest that details of designs can be found in the many text books and application notes now available. Very few people would attempt (or even could) design an active filter from basic theory today. There are now available many well-prepared circuit design guides. The Burr-Brown "Handbook of Operational Amplifier Active RC Networks" has an excellent one. This book is, however, now out of print but many copies were distributed by Burr Brown agents. It contains twelve basic circuits, along with quite manageable design procedures for each, in which desired values are put in for formulae to arrive at circuit values for low-pass, band-pass and high-pass requirements.

FILTER CHARACTERISTIC TERMINOLOGY

The ideal edge on a filter characteristic is usually a sharp "square" response with attenuation occuring instantly as the frequency passes through the corner point. It should also have a constant response level at all points in the pass-band regions. As well as the rudimentary RC filter characteristic which falls off at 20 dB/decade from a breakpoint, two other kinds of response are commonly encountered. These are Butterworth and Chebyshev responses. Both derive their names from persons who developed the mathematics involved — (Butterworth designed filters around 1930, Chebyshev developed certain mathematical theory in his study of steam-engine linkages around 1850).

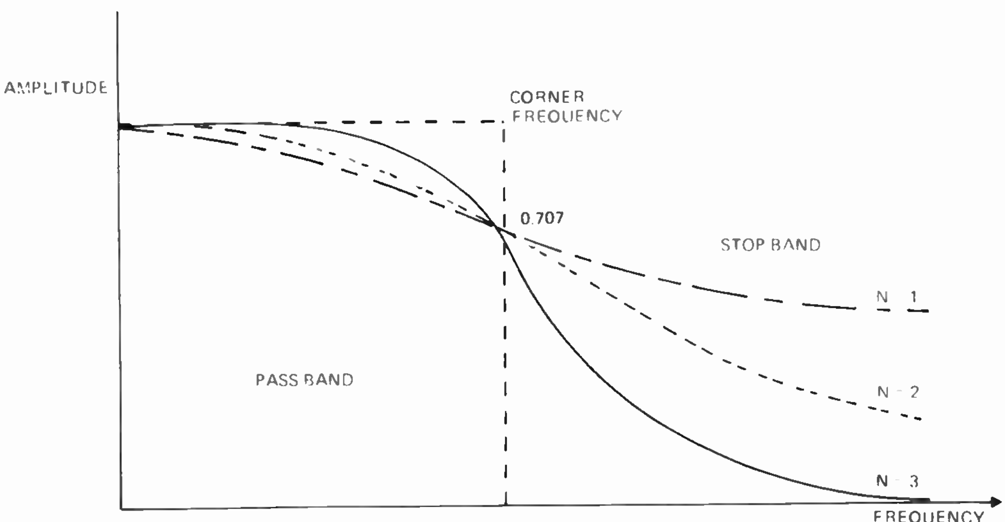


Fig. 17. Butterworth filter responses for various orders used.

The Butterworth response is said to be maximally flat (that is as flat as possible) in the pass-band region. It has the optimum constancy possible with a given number of available peaking resonances (the complex passive or active filter circuits can be regarded as a group of staggered-tuned resonating sections, each arranged to peak just aside of the others, thereby, providing a broadened response band and a reject region). Fig. 17 shows the kind of Butterworth responses obtainable. Note that each passes through the 3 dB, down half power, point. The order (a mathematical term denoting the number of resonances available) of the filter is denoted 'n' in the chart. A typical roll-off rate is 20 n dB/decade so a fourth-order Butterworth response filter (which can

ELECTRONICS -it's easy!

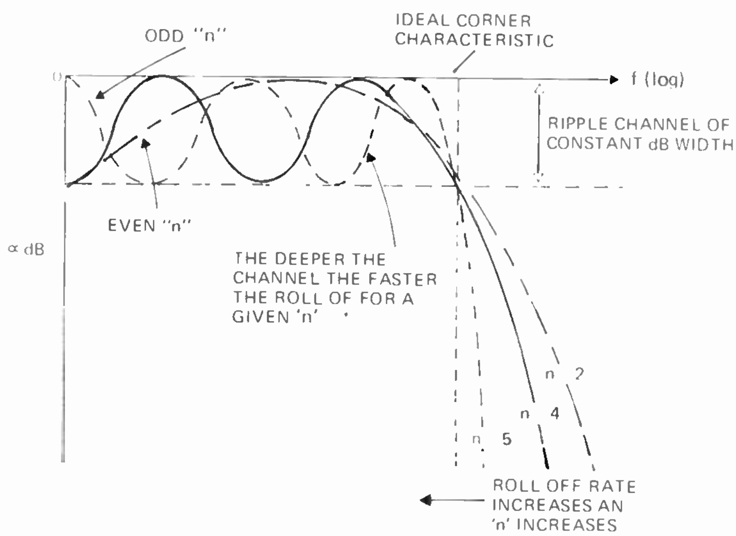


Fig. 18. Chebyshev filter responses for various orders.

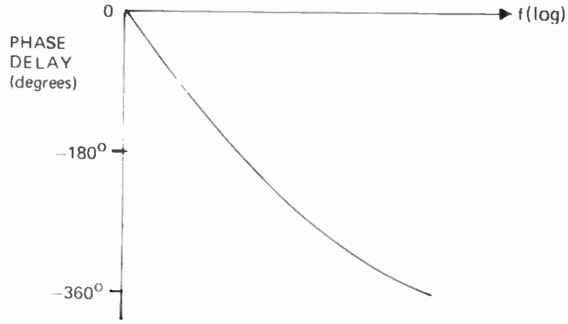


Fig. 19. Phase response of phase-linear all-pass filter.

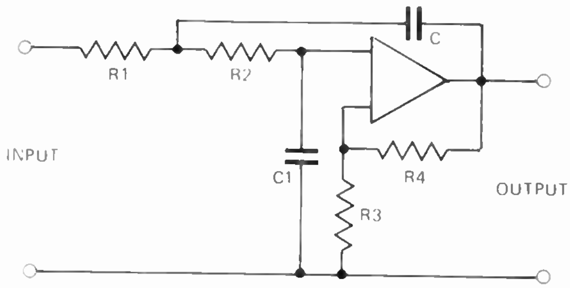


Fig. 20. General configuration of constant time-delay Bessel all-pass filter.

be realised by either passive or active methods) will attenuate at around 80 dB/decade.

Whereas the pass-band response is reasonably constant, the rate of roll-off is not as good as can be obtained if the resonating sections are staggered differently. Other criteria of staggering the resonances can provide higher roll-off rates but only by introducing "ripples" in the pass-band response. When these individual ripples have equal amplitude across the pass-band response Chebyshev polynomials describe the shape, thus giving the name to an alternative response situation. As with Butterworth designs the higher the order the better the roll-off rate as can be seen diagrammatically in Fig. 18. The depth of ripple that can be tolerated also influences the roll-off rate – the smaller the variation that can be allowed the less the roll-off rate. (This can be readily seen by sketching in the required number of ripples of given depth at the appropriate scale).

Normally Butterworth or Chebyshev response filters will be of order 1 to 4 but higher orders are possible. These two forms are not the only sophisticated filter responses available: other mathematical criteria could be used to set up workable mathematical equations for designing other networks. These two will, however, meet most demands required and all filter design, as we have seen, is dominated by need to compromise between what is needed and what can be handled mathematically.

PHASE SHIFT AND DELAY FILTERS

These act to provide a phase shift to a signal without selectively attenuating the frequency content. They are sometimes called all-pass filters. The amount of phase shift of practical circuits, however, usually varies with the frequency of the signal even though the amplitude response is invariant. Constant time-delay or linear-phase filters have a reasonably straight (linear) phase response as shown in Fig. 19. The so-called Bessel filter approximates this response using a workable mathematical formulation. Fig. 20 gives the general configuration of such a method realised as an active filter design.

COMPONENTS TO USE

Resistors – In non critical applications the normal 20% tolerance carbon composition resistor may be acceptable. If tighter filter characteristics are needed then one must resort to more expensive resistors such as 5% or closer tolerance carbon composition. Even better, use metal-film or wirewound types. It is sometimes permissible to hand choose values from wide tolerance groups in order to produce specific values, but it must not be forgotten that wide tolerance resistors often lack the same

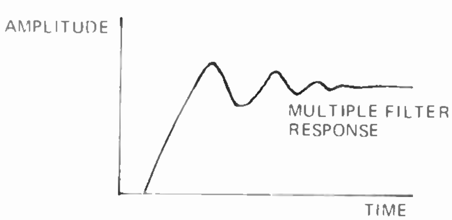


Fig. 21. Multi pole filters resonate with transient excitation.

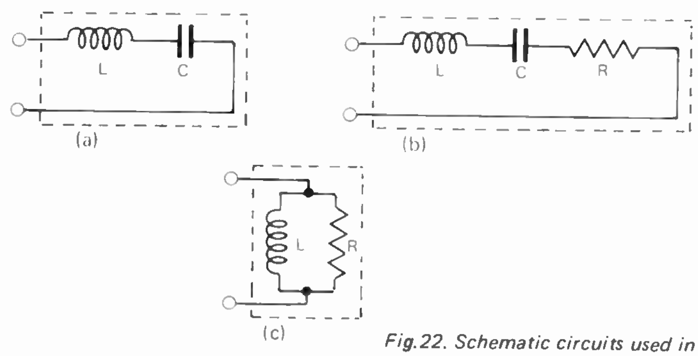
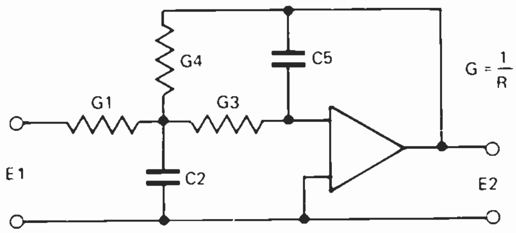


Fig. 22. Schematic circuits used in examples.

degree of time and temperature stability as the more expensive types.

Capacitors – Ceramic disk capacitors can be employed but they are best avoided. Nylon film, polystyrene and Teflon capacitors are much the better to use. When especially long-period filters are needed the capacitance value will be large. In such cases the leakage current due to losses in the dielectric is extremely critical and this rules out, in the majority of cases, using electrolytics.

Op-amp – It is easy to assume all op-amps will provide good active filters but this is not so. The main factor is a low offset current, this being especially important in long-period filters. As a general rule the more critical the need the better the op-amp should be. When op-amp filters also add gain they should have an open-loop gain at least 50 times the filter gain. Many active-filter design procedures enumerate the requirements of the op-amp.



$$\frac{E2}{E1} = \frac{-G_1 G_3}{S^2 C_2 C_5 + SC_5 (G_1 + G_3 + G_5) + G_3 G_4}$$

Fig.23. Typical low-pass active filter with its transfer function in s-notation form.

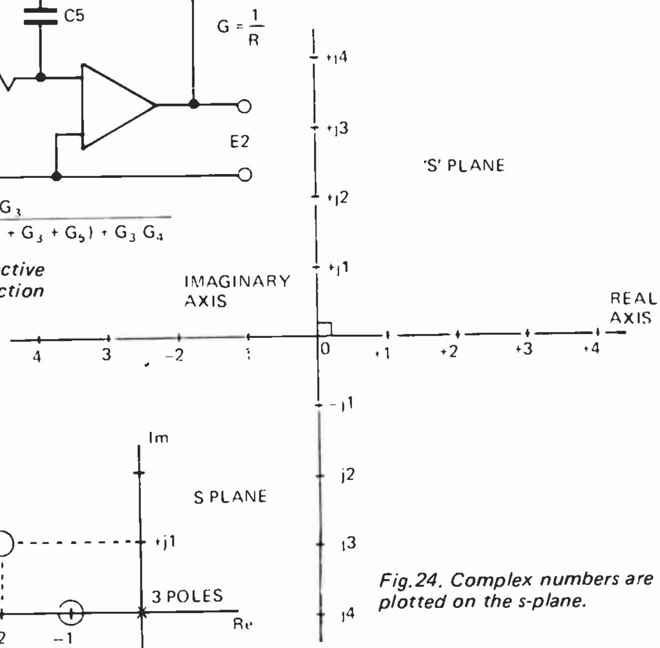


Fig.24. Complex numbers are plotted on the s-plane.

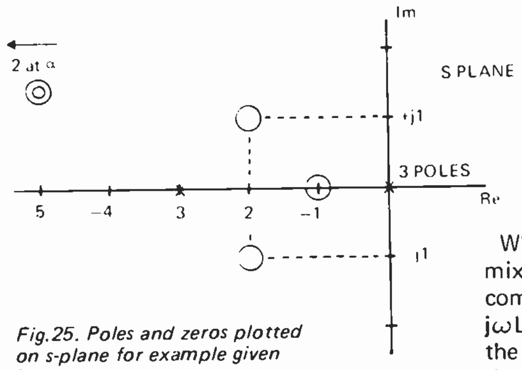


Fig.25. Poles and zeros plotted on s-plane for example given in text.

RESPONSE TO TRANSIENTS

Filters of second order and higher invoke the characteristics of resonating circuits for their operation. In passive filters we can readily identify the inductance and capacitance; in active circuits these may not be so obvious, the mathematical expression showing that resonances do occur.

When a step change in signal is applied to a resonant circuit, the circuit 'rings', that is, the output rises rapidly but then oscillates with decreasing amplitude to the final value as indicated in Fig. 21. The extent to which a resonant circuit rings is decided by the damping provided – the higher the Q of the resonant configuration the greater the ringing effect.

It is not hard to see that higher order filters, therefore, will tend to ring more than the lower order designs when transient signals appear at their input terminals. Transients occur in practice as noise spikes, switching spikes, sudden signal appearance and departure.

THE S-PLANE, POLES AND ZEROS
(For the advanced reader)

S-Notation

The above study of filters can only act as a guide to filter selection. From there one must turn to the many articles and books available for details. To make good use of such material it is necessary to have a basic understanding of the mathematical methods used. This section is given to assist the more advanced reader. It is possible to get by without this information, provided a suitable configuration and design procedure can be located. Therefore do not be concerned if you are unable to understand this section.

Scanning through even basic, well-organised books on filter (and feedback amplifier design) the terms transfer function, s-plane, poles, zeros and root-locus will be encountered. Sadly, most books omit to provide the background explaining what this is all about. The concepts are not difficult to grasp, any confusion arising almost certainly from the number of synonomous terms used and the fact that the concepts are, perhaps, quite alien to begin with.

We have seen how reactive elements (capacitance and inductance) have apparent resistances of $2\pi fL$ for inductance and $1/2\pi FC$ for capacitance. These terms, however, do not provide information about the phase changes produced with these reactance elements.

Electronic circuit designers use the operator symbol j (mathematicians use i) to denote a phase change of 90° hence, $j2\pi fL$ represents both the reactance value *and* the phase cnange. Furthermore $j = \sqrt{-1}$. For capacitive reactance the complete notation is $-j2\pi$ Fc, as the capacitor introduces a 90° phase shift of opposite sign to inductance. Resistance, having no phase shift, nor being frequency dependent is merely R. We can be a little more basic still and use ω instead of $2\pi f$, ω is the angular frequency being expressed in radians $\cdot$ sec^{-1} (There are 2π radians in one cycle).

When reactance and resistance are mixed we represent the value as a complex number as, for example, $R + j\omega L$. The left-hand part is known as the Real part, the other (that after j) the Imaginary part, the whole forming what is called a complex number.

Where the circuit element is only reactive the complex number representing the impedance reduces to $j\omega L$ or $-j\omega C$ for which the symbol 's' is used instead of $j\omega$. (In some books 'p' is used instead of 's'). A trap can occur here for the $-j$ of $-j\omega C$ indicates a 180° phase shift over j, not a negative quantity in the normal way. To avoid confusion we rewrite $-j\omega C$ as $1/j\omega C$ (which is valid – it comes from multiplying both numerator and denominator $-j\omega C$ by j. Hence we obtain sL and 1/sC as the shorthand way of writing inductive and capacitive reactance in which frequency dependency and phase information are both retained.

Once these terms and concepts are mastered it becomes much more straightforward to write down the transfer function for a frequency dependent network. For example, consider finding the impedance presented by a series, lossless, resonant circuit shown in Fig. 22a.

$$Z = sL + 1/sC = \frac{L\,(s^2 + 1/LC)}{s}$$

(The individual components of the expression are put on a common denominator, dividing out to get the s^2 terms with unity coefficients).

For the series resonant lossy circuit of Fig. 23b.

$$Z = sL + 1/sC + R = \frac{s^2 + R/L\cdot s + 1/LC}{s \cdot 1/L}$$

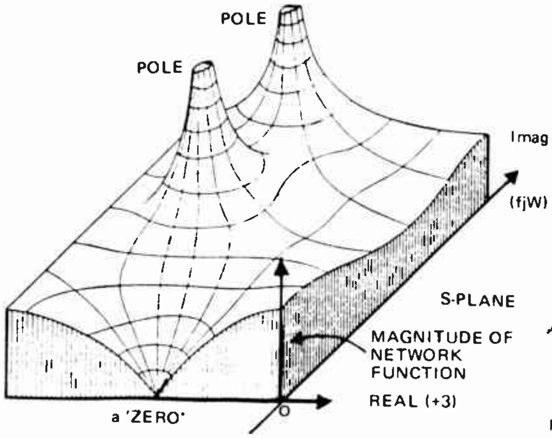


Fig.26. Topographical representation of poles and zeros in s-plane.

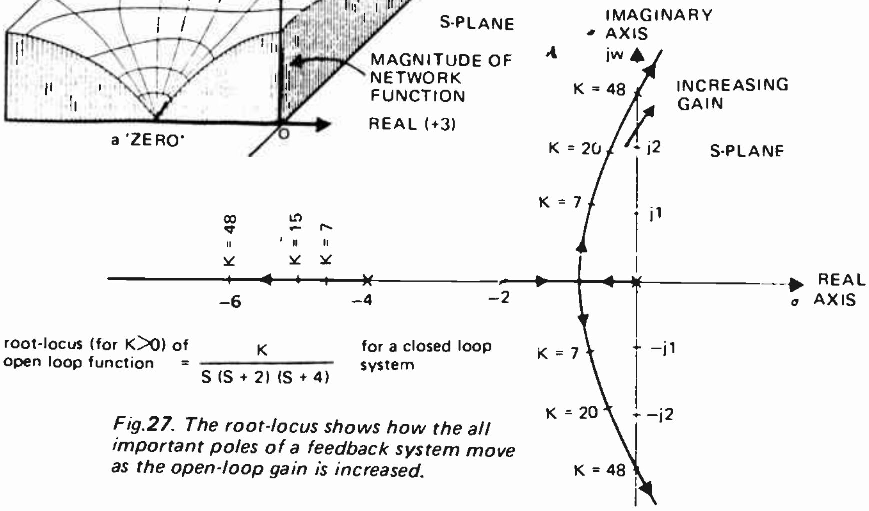


Fig.27. The root-locus shows how the all important poles of a feedback system move as the open-loop gain is increased.

Again, for the parallel L and R circuit of Fig. 22c.

$1/Z = 1/sL + 1/R$ from which

$$Z = \frac{R \cdot s}{(s + R/L)}$$

It is these forms of expression that are quoted in circuit design books. The form of expression is not restricted to two terminal networks – it applies for all frequency dependent reactive networks. Fig. 23 gives the circuit of a low-pass multiple feedback active filter along with its derived transfer function expressed in 's' notation form.

As these complex numbers possess two parts we must use a two-axis graph to represent them in which the two axes are mutually perpendicular. Thus complex-number quantities need a plane rather than a line to depict a unique number. This plane is known as the s-plane (see Fig. 24). The two axes are usually labelled Re, R or σ for the Real axis and Im, I or $j\omega$ for the Imaginary axis, each pair being used respectively.

POLES AND ZEROS

We have seen above how a network of passive elements (active designs also apply) produces a mathematical expression in terms of s notation. As s merely represents $j\omega$ and j denotes only phase information we can, whenever s appears, substitute ω (or 2π f) to see how the expression varies in magnitude with varying frequency.

Consider the case where a function is given by the numerical example:

$$\frac{(s + 1)\ (s + 2 + j1)\ (s + 2 - j1)}{s^3\ (s + 3)\ (s + 5)}$$

When s = −1, −2 -j1 or −2 + j1, the numerator becomes zero for one of the bracketed terms becomes zero. Hence at each of these frequency values the expression becomes zero. We say it has 'zeros' at these points. Zeros also exist when the singular s term goes to infinity in the denominator. When s = 0 (three times, as it is from $s^3 = s.s.s.$), −3 or −5, we get the reverse situation for at all of these values of ω the denominator goes to zero making the function rise to infinity. These frequency points are called 'poles'.

Thus the poles and zeros express the peaks and hollows of the function. The position of these can be plotted on the s-plane diagram as shown in Fig. 25. 0 is used for zeros, a cross X for poles. In realisable networks there must be as many poles as zeros – including those at zero and infinity.

Another way to imagine the network characteristic is to draw a topographical representation giving relative height to poles and zeros on the s-plane placed horizontally as shown by the example of Fig. 26. This makes the terms poles and zeros more meaningful in a physical sense.

In the numerical example we avoided, in that case, using a quadratic or higher order term such as $s^2 + 4s + 5$. When these are encountered they must be factorized by finding the roots of the expression – giving the two terms s + (2 + j1) and s + (2 – j1) in this case. These are the individual roots, i.e., poles and zeros, of the expression. Note that quadratic elements involving an Imaginary part form mirror image pole or zero pairs – called a conjugate pair. If these are lossless (no Real part) they lie on the imaginary axis, if lossy (with Real part) they will be displaced out into the s-plane depending upon the resistive value. Positive values of resistance result in displacement into the left-hand plane, negative resistance gives poles or zeros in the right-hand plane, these halves being denoted LHP and RHP respectively.

Mathematics of complex numbers show that resonant systems with roots lying in the LHP are stable systems, their oscillations die down because to be in the LHP they must contain resistive damping. If the roots lie on the Imaginary axis itself the system is marginally stable – transients will undoubtedly create unstable situations at times even though the system is not absolutely unstable. Note that this situation only arises if the resistive component occurs as negative resistance – oscillators create this condition by the use of an active element.

ROOT LOCUS

When considering the behaviour of feedback systems, such as amplifiers, controllers and active filters, it is highly valuable to plot the changes in position on the s-plane of the closed-loop poles of the system transfer function as the open-loop gain changes. The path traced by the movement of the poles in this way is called the root-locus. These are often referred to in amplifier and other feedback-mechanism designs and it is, therefore, helpful to at least appreciate what they are. It is, however, not a simple matter to produce them from an original expression; lots of experience is vital.

By way of example the root-locus for a relatively simple transfer function is given in Fig. 27. This tells us that an open-loop gain in excess of 48 places some of its poles in the RHP establishing an unstable situation. The value of the root-locus is that we can "see" the behaviour of the system as the gain is increased and, more importantly, what we should do to the position of the poles most influencing an unstable situation. By altering the transfer function we can place the locus in more favourable situations. This is done by altering original component values where possible or by adding other networks that reduce the effect of the dominant poles – those lying close to the RHP.

Volume 1 of this two-part edition covers basic electronic theory. It deals generally but not exclusively with analogue technique.

Volume 2 takes the reader into digital theory and technique. Subjects covered include: an introduction to digital systems, logic algebra, logic ICs, digital sub-systems, digital displays, D-A and A-D conversion, digital instruments and computers, computer peripherals, transmission links, oscilloscopes and chart recorders.
See inside back cover for further details.

Index

Both volumes of this book are available from all major newsagents, technical bookshops and electronics suppliers, or direct from the publishers, Electronics Today International, 15 Boundary Street, Rushcutters Bay, NSW 2011 (price $5.95 per volume plus $1.00 per volume post and packing).
Electronics it's Easy is also available in two-volume hard-cover form. It may be obtained from technical bookshops and electronics suppliers or direct from the publishers at the above address (price $12.95 per volume plus $1.00 per volume post and packing).

Acknowledgement

Electronics it's Easy was first published as a monthly series in Electronics Today International. It was republished as three separate volumes during the period 1975-1976.

A second edition was published in 1977-1978 and a revised version of Volume 1 was produced especially for the TAFE School of Electrical Engineering in Sydney in 1980.

This, the third edition of the complete series, has been totally revised and new material added where relevant. At the same time the material has been re-arranged so that the complete work now appears in two volumes rather than the original three. Apart from a card-cover edition, the work is also available in hard-back form — details at the end of both volumes.

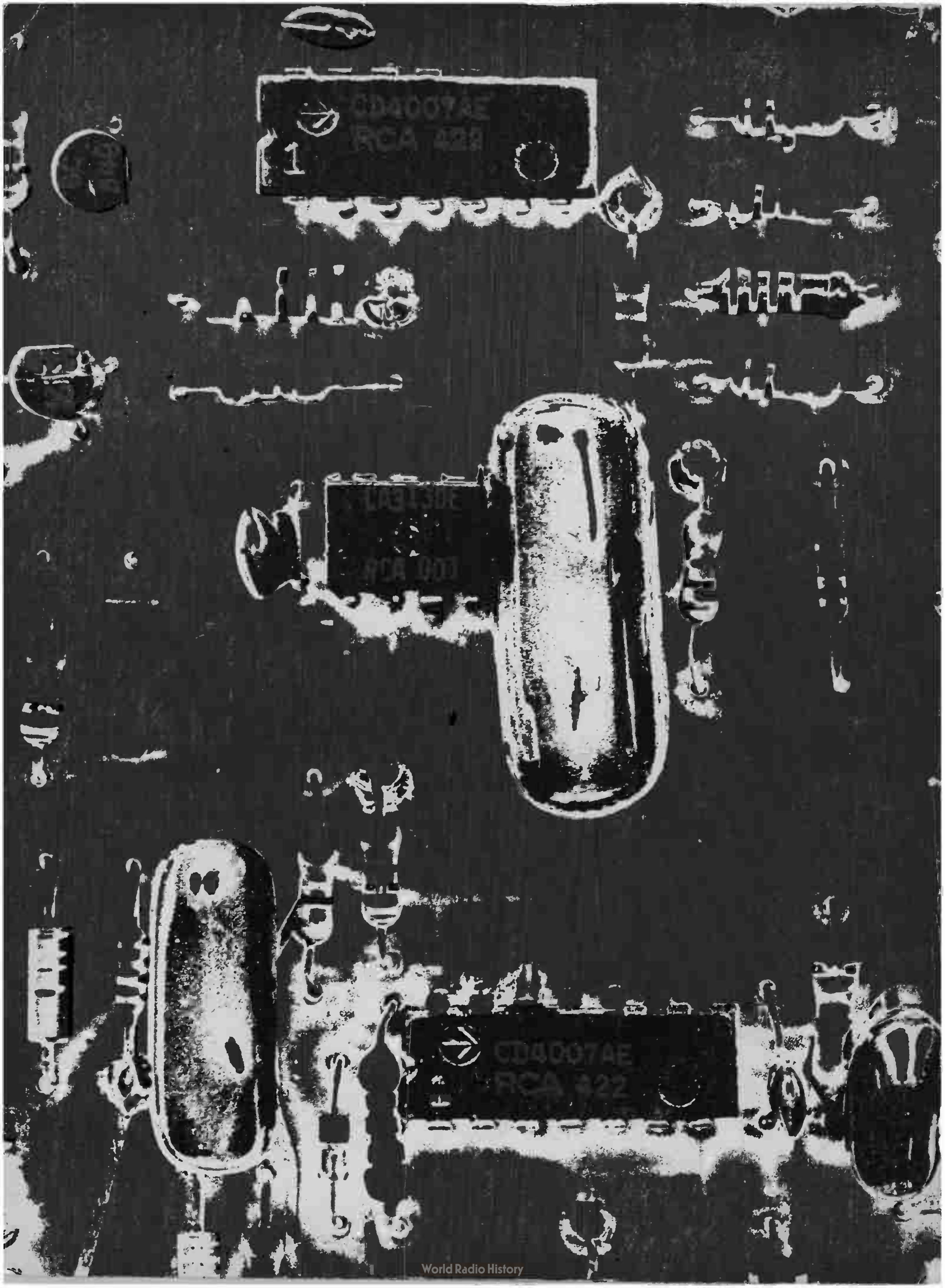
The author would like to thank Brian Chapman (technical editor of Electronics Today International from 1973-1976) for editing the original edition. Thanks also to Ron Farley and Maree Breen for turning hundreds of my rough pencil sketches into clear and concise engineering drawings.

The very considerable work of turning my original individual articles into book form, and of editing and producing this latest revised edition, has been done by Janet Vernon B.A., special publications editor of Electronics Today International.

I wish to thank the following for use of illustrations:—
Philips; BWD; Science Museum (London); Advance Instruments; Electronics Australia; Hewlett-Packard; Teledyne Philbrick; Siemens; Tektronics; Non-Linear Systems; John Fluke; Heidenhein; Lloyds Bank; AEI Associates; Kirklands; Novus; Gerber Scientific; Honeywell; Motorola; National Semiconductor; Levers Rich; Andrew Antennas; Vu-Data; Data-Lab; Marconi; Australian Post Office; Kent Instruments' Record Electrical Company; Goerz, Hathaway Instruments; Devices; and Kepco.

P Sydenham

Peter H. Sydenham. Adelaide, April 1981.


CD4007AE
CA3130E
CD4007AE
RCA 922