McGRAW-HILL SERIES IN
INTUITIVE IC ELECTRONICS

John Watson

Mastering Electronics

THIRD EDITION

MASTERING ELECTRONICS

THIRD EDITION

JOHN WATSON

McGRAW-HILL, INC.

New York St. Louis San Francisco Auckland Bogotá
Caracas Hamburg Lisbon London Madrid
Mexico Milan Montreal New Delhi Paris
San Juan São Paulo Singapore
Sydney Tokyo Toronto

Library of Congress Cataloging-in-Publication Data

Watson, John (John Richard)
Mastering electronics / John Watson. – 3rd ed.
p. cm.
Includes index.
ISBN 0-07-068483-9. – ISBN 0-07-068482-0 (pbk.)
1. Electronics. I. Title.
TK7816.W35 1990
621.381 – dc20 90-6257
CIP

1 2 3 4 5 6 7 8 9 0 DOC/DOC 9 5 4 3 2 1 0

ISBN 0-07-068482-0

ISBN 0-07-068483-9 {PBK}

The sponsoring editor for this book was Daniel A. Gonneau and the production supervisor was Thomas G. Kowalczyk.

Printed and bound by R. R. Donnelley & Sons Company.

First published in Great Britain by
Macmillan Education Ltd. in 1983.

First American edition published by McGraw-Hill in 1986.

CONTENTS

CONTENTS

CONTENTS

III. DIGITAL ELECTRONICS

CHAPTER 1

INTRODUCTION TO ELECTRONICS

Only a few years ago, you would not have found the study of 'electronics' on the curriculum of any university or college. The subject was not considered important enough to be given a category of its own, and was studied as a rather specialist branch of electricity. Today, this relatively new branch of technology has made an impact on almost every aspect of our lives. Electronics has made possible a range of inventions, from televisions to spacecraft. Many of these inventions—or the effects of them—have revolutionised the way we live.

The growth of electronics as a branch of technology has been unprecedentedly rapid. Never before has such a completely new technology been developed so quickly or so effectively, and so universally. Electronic techniques are now applied to all branches of science and engineering. A study of electronics is therefore central to any science or engineering course, and more and more people, in all walks of life, are going to be needing a basic knowledge of electronic technology.

It is true that electronics developed from the study of electricity. Early ideas about the way electric current could flow through conductors and through a vacuum led to the development of useful radio systems. It was possible to send messages around the world using what was, by today's standards, incredibly simple and crude equipment. The Second World War provided an urgent requirement for more sophisticated communication and other electronic systems. The invention of radar (radio direction finding and ranging) required a big step forward in theory and an even bigger step forward in engineering. The study of electronics gradually became an important subject in its own right, and the radio engineer became a specialised technician.

The development of television led to one of the most massive social changes that has taken place. Many households now possessed televisions, radios and record players. In some branches of industry, electronic systems became useful—but electronic devices not directly concerned with the wire-

less transmission of sound or pictures were still something of a curiosity.

Only in the early 1960s did electronics technology 'come of age' thanks to the work of three scientists, Bardeen, Brattain and Shockley, in the Bell laboratories in the USA. In 1957 they assembled the first working transistor.

1.1 THE MICROELECTRONIC REVOLUTION

In order to understand just how much impact the invention of the transistor was to have, it is necessary to realise that every electronic machine required the use of *valves*. Valves will be described in Chapter 5, but for the time being it is enough to know that valves are rather inconvenient devices for handling electrons. Valves are rather large, difficult to produce on a large scale, and extremely wasteful of power. The transistor, on the other hand, can be made quite small and uses relatively little power. It is much cheaper to make than a valve and lends itself to mass-production techniques. More importantly, the transistor can be made *very* small. There is a theoretical limit to the size of a transistor, but this limit is astonishingly tiny. Using transistor technology it became possible to make electronic circuits very complicated and very small. It is quite difficult to get an idea of the difference in scale between electronic devices using valves and electronic devices using microelectronic technology. It does not give a very valid comparison to look at a radio receiver of the 1930s and compare its size with a modern transistor radio. The size of the transistor radio is not limited by technology but rather by the size of the person who is expected to operate it. Transistor radios can be made extremely small, but the lower limit of smallness is reached when it becomes impossible to operate the controls!

A more useful comparison can be made by looking at computers. One of the first working, large-scale computers was made in the late 1940s. It occupied an area about equal to that of an hotel suite, and used as much power as a medium-sized street of houses. It was vastly expensive, and vastly unreliable—on average a valve had to be changed once every ten minutes.

I am fortunate enough to own a computer of substantially more power than that of the giant of the 1940s. It cost me less than one week's pay, and fits neatly into my jacket pocket. It will operate for approximately 300 hours on two tiny batteries and I would be surprised if it ever went wrong.

All this gives an indication of the scale of change that has taken place as a result of the development of electronics, and of the 'microelectronic revolution' sparked off by the invention of the transistor.

1.2 ELECTRONICS TECHNOLOGY TODAY

Most of the development in electronics has taken place since 1960. Inevitably there have been large changes in the way the subject is taught, and indeed in what is taught. Solid-state technology has even led to a change in the way we look at the physics of the atom. Models of the atom which work well for electronics based on the relatively crude 'valve' technology are often not sufficient for a useful description of modern microelectronic components.

This book is divided into three parts. Part I is a revision of **basic electricity**. This part of the book goes rather faster than the other two parts. It will provide you with sufficient knowledge of electricity to enable you to understand the rest of the book. It is not, however, a complete study of the subject but will be a useful memory-jogger for anyone who has studied electricity at school but who needs to be reminded of what was learnt there.

Part II deals with **linear electronics**. Broadly, this is a study of electronic components and systems that deal with continuously varying quantities and represent these quantities by means of an electrical analogue. Electronic systems that involve this technology include television and radio, audio and video reproduction equipment, and a number of electronic instruments.

Part III deals with **digital electronics**. In digital electronics, systems deal only with numbers, usually represented by the presence or absence of an electrical signal. Digital electronic systems include, of course, computers and calculators, and a number of electronic instruments. Digital electronics is probably the most rapidly developing branch of electronics technology. Various functions previously carried out by linear systems are now being dealt with digitally—one major example is studio-quality audio recording and reproduction. A sound signal can now be represented as a stream of numbers, and can be recorded and replayed in this form to give a totally reproducible and high-fidelity audio signal. Most commercial sound-recording studios now use digital techniques for recording.

1.3 ELECTRONIC SYSTEMS

It used to be necessary for students of electronics to understand at a basic physical level the way in which electronic systems worked. This is still true in so far as a student must have a thorough understanding of the physics of electronic devices. However, it is no longer necessary—or indeed possible—for students to understand the detailed workings of individual circuits. A microelectronic counter, for example, may be extremely com-

plex in its actual operation. If a manufacturer has some novel circuit, he may be very reluctant to publish details of the way the device works. In the case of very complex microelectronic systems—such as microprocessors—it would be a major study to examine in detail the operation of just one such device. The technologist today is concerned with *systems*, and with the operating parameters of the individual circuit components. An individual circuit component in this context may be a complete amplifier or counter. The manufacturer will provide very complete data, and provided the user is fully aware of the device's characteristics, it is no longer necessary to understand the precise details of a device's internal functioning. For reasons outlined above, such information may be rather difficult to obtain.

PART I

ELECTRICITY

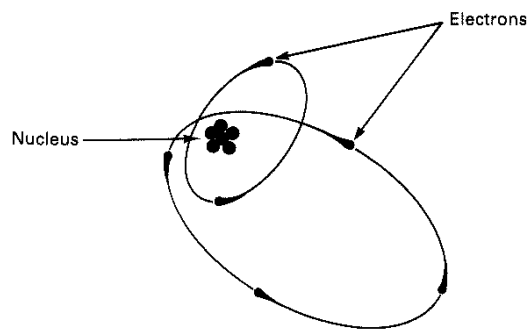
CHAPTER 2

ELECTRICITY

What is electricity? To begin to answer this question, we have to look at the composition of matter. All matter is made up of atoms, but it is far from simple to describe an individual atom. Atoms are of course far too small to be seen under the most powerful microscope, and we have to infer their construction from the way they behave and from the way they affect various forms of radiation beamed at them. Because nobody quite knows just what an atom really looks like, physicists construct *models* of atoms to help them explain and predict atomic behaviour.

One of the simplest and most straightforward models of the atom was proposed by Niels Bohr, a Danish physicist, in 1913. In Bohr's model of the atom it is assumed that tiny electrons move in orbit around a central, much more massive, nucleus. The electrons are confined to orbits at fixed distances from the nucleus, each orbit corresponding to a specific amount of energy carried by the electrons in it. If an electron gains or loses the right amount of energy, it can jump to the next orbit away from the nucleus, or towards it. A typical atom, drawn according to Bohr's theory, is shown in Figure 2.1. Electrons towards the outer edge of an atom are

fig 2.1 *a model atom, according to Bohr's theory*



held in orbit rather loosely compared with those towards the centre. Electrons in the outermost orbits can easily be detached from the atom, either by collision or by an electric field. Once detached, they are known as *free electrons.* All materials contain some free electrons.

Materials that are classed as *insulators* contain only a very few free electrons, while materials that are classed as *conductors* (notably the metals) contain large numbers of free electrons. These free electrons drift around in the material more or less at random. However, under certain circumstances the electrons can be made to flow in a uniform direction. When this happens, they become an *electric current.*

Let us look at one such example, that of a dry cell. A study of the workings of the cell is not part of a book on electronics, but details can be found in textbooks on electricity. The illustration in Figure 2.2 shows a much simplified diagram of a typical dry cell. For our purposes, the important thing is that the internal chemistry of this cell causes the outer casing of the battery to accumulate an excess of free electrons. The centre terminal, on the other hand, has almost no free electrons. If the casing and centre electrode of the battery are connected by a conducting wire—as shown in Figure 2.3—electrons will migrate from the casing to the centre electrode by flowing down the wire. This current flow is explained by the

fig 2.2 *a single dry cell*

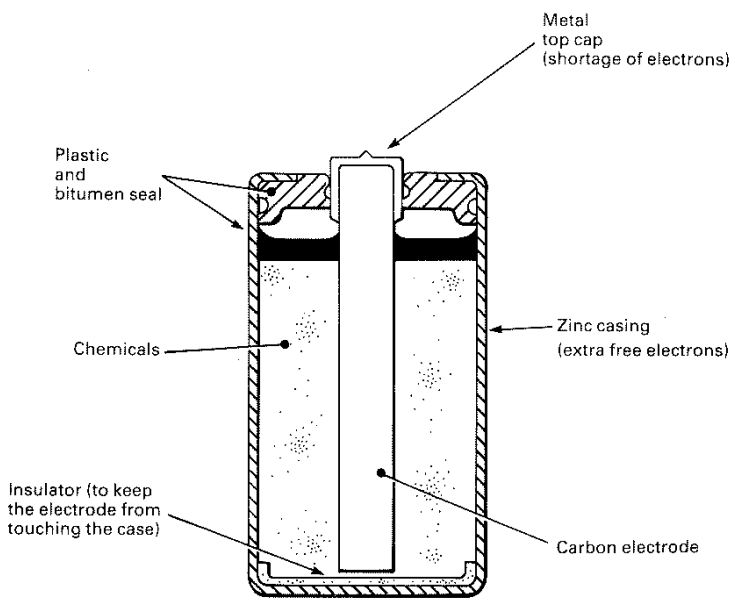
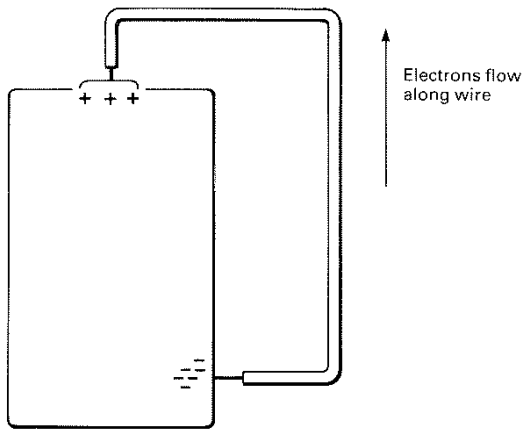


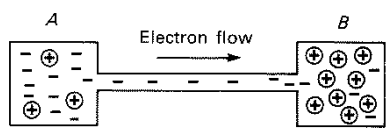
fig 2.3 *electrons flowing in a circuit*



fact that all electrons carry a single unit of negative electric charge. The electric charge is a form of energy, and may in some ways be regarded as analogous to magnetism. In the same way that similar poles of a magnet repel each other (north repels north, south repels south), so do electrons repel each other. In a normal atom the charges on the electrons are exactly balanced by positive charges on the nucleus. An atom in this state is said to be *neutral*. Free electrons are repelled by other free electrons, and, as you might expect, are attracted towards areas with a net positive charge. This is illustrated diagrammatically in Figure 2.4. Area *A* has a large number of free electrons, far more than are necessary to balance the positive charges on the nuclei of the atom. Area *B* has very few free electrons, and indeed a number of the atoms may be missing electrons from their outer orbits and may therefore exhibit a net positive charge. The electrons are repelled by the charge in area *A* and flow along the wire towards area *B*.

The above is something of a simplification, but illustrates the general principle.

fig 2.4 *electrons move from an area having net negative charge to an area with net positive charge*



2.1 UNITS OF ELECTRICITY

It is possible to count (indirectly) the number of electrons flowing along the wire. In order for there to be a useful current, a very large number of electrons must flow. Accordingly, the basic unit of electric current flow is equivalent to around 6.28×10^{18} electrons per second. This unit of electricity is called the *ampere*, after André Marie Ampère, a French physicist who did important pioneering work on electricity and electromagnetism. The ampere, often abbreviated to 'amp', is a practical unit of current for most purposes.

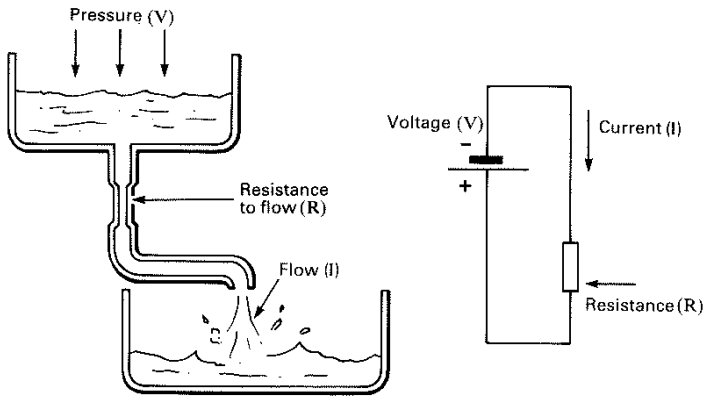
In addition to measuring the current flow, we also need to measure another factor, the *potential difference* between two points. The potential difference between two points is the energy difference between them. Look at Figure 2.4 again, and imagine that there is a much larger number of free electrons on side *A*. This would have the effect of increasing the amount of negative charge on *A* and increasing the amount of energy available to push the electrons along the wire. Similarly, if side *A* had only a very few more free electrons than side *B*, the energy available to push electrons along the wire would be much less. There is a parallel here between the potential difference and water pressure—in fact we can use a 'water analogy' to illustrate the way an electric circuit works. If current flow is the equivalent of a flow of water along a pipe, the potential difference is the amount of pressure available at the end of the pipe. But whereas we might measure pressure in terms of kilograms per square metre, potential difference is measured by a unit called the *volt*. The volt is named after Alessandro Volta, an Italian physicist who, like Ampère, pioneered work in electricity—he produced the first practical battery. As you might imagine from the water analogy, the current flow is affected by the potential difference. All else being equal, the higher the potential difference—or voltage—the greater will be the current flow.

There is one other factor that would affect the flow of water in the plumbing system: that is the size of the pipe. There is a corresponding factor in the electric circuit, and that is the *resistance* of the wire. Resistance is a property of all materials that reduces the flow of electricity through them. The higher the resistance, the more difficult it is for current to flow. The three factors affecting current flow are illustrated—with the water analogy—in Figure 2.5.

The relationship between current, potential difference and resistance was first discovered by Georg Simon Ohm in 1827. It is called *Ohm's Law* after him. Ohm's Law states that the current, *I*, flowing through an element in a circuit is directly proportional to the voltage drop, *V*, across it. Ohm's Law is usually written in the form:

$$V = IR$$

fig 2.5 *a plumbing analogy: voltage, current and resistance all have their equivalents in the water system*



The unit of resistance is called the *ohm*, for obvious reasons. From the formula above we can see that a potential difference of one volt causes a current of one ampere to flow through a wire having a resistance of one ohm. Given any two of the factors, we can find the other one. The formula can be rearranged as:

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

or

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

This simple formula is probably used more than any other calculation by electrical and electronics engineers. Given a voltage, it is possible to arrange for a specific current to flow through a wire by including a suitable resistor in the circuit. Similarly, it is possible to determine voltage and resistance, given the other two factors. For some applications in electrical engineering, and for many in electronic engineering, the ampere and volt are rather large units. The ohm, on the other hand, is a rather small unit of resistance. It is normal for these three units to be used in conjunction with the usual SI prefixes to make multiples and submultiples of the basic units. These are given in the chart in Figure 2.6. When using Ohm's Law in calculations, it is vital that you remember to work in the right units. You cannot mix volts, ohms and milliamps!

fig 2.6 *SI prefixes*

Prefix	Symbol	Meaning	Pronunciation
tera	*T*	$\times 10^{12}$	tare-ah
giga	*G*	$\times 10^{9}$	guy-ger
mega	*M*	$\times 1\,000\,000$	megger
kilo	*k*	$\times 1\,000$	keel-oh
milli	*m*	$\div 1\,000$	milly
micro	μ	$\div 1\,000\,000$	—
nano	*n*	$\div 10^{9}$	nar-no
pico	*p*	$\div 10^{12}$	peeko
femto	*f*	$\div 10^{15}$	femm-toe

2.2 DIRECT AND ALTERNATING CURRENT

So far we have been considering electric current that flows continuously in one direction through a circuit. However, in many applications we will encounter electric current that flows first in one direction and then in the reverse direction. Current that flows first one way and then the other is called *alternating current*, and a voltage source supplying such a current is called an *alternating voltage*. Alternating voltages and currents occur in various circuits, particularly in radio and audio, and in power supplies. Alternating current is used for all mains power supplies because a.c. permits the use of *transformers* to increase and decrease the voltage, so that the power can be transmitted from place to place using efficient high-voltage transmission lines.

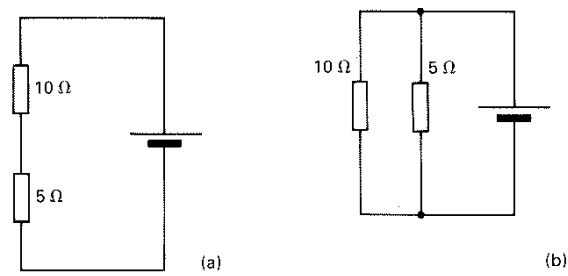
The rate at which the voltage reverses is the *frequency* of the alternation. One complete *cycle*, from one direction to the other, and back to the original direction, is the measure: one cycle per second is given a unit, the *hertz* (symbol Hz) after Heinrich Rudolf Hertz who first produced and detected radio waves. The hertz is equal to one cycle per second, and can be used with the usual SI prefixes; thus 1000 Hz is usually written as 1 kHz, etc.

2.3 RESISTORS IN SERIES AND PARALLEL

An electrical circuit that draws power from a source of electricity is known as a *load*. It is possible to connect two loads in an electrical circuit so that all the current flows through both of them. Equally, it is possible to connect them so that current flow is divided between them. The two possibilities are illustrated in Figure 2.7.

Figure 2.7a illustrates *series* connection. In this illustration the loads have a resistance of 10 Ω and 5 Ω respectively. To work out the combined resistance of the two loads, for use in calculations involving Ohm's Law, the two values can simply be added together. This is true of any number

fig 2.7 *resistors in series (a) and parallel (b)*



of resistive loads connected in series—just add the values together, making sure that all the resistances are expressed in the same units (once again, you cannot mix ohms and kilohms).

The second case, illustrated in Figure 2.7b, is less easy to calculate. The formula for working out combined parallel resistance is:

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

R is the combined resistance, where R_1, R_2, etc. are the individual parallel resistances. Working this out numerically for the example in Figure 2.7b, we get:

$$\frac{1}{R} = \frac{1}{10} + \frac{1}{5}$$

$$\frac{1}{R} = 0.1 + 0.2$$

$$\frac{1}{R} = 0.3$$

$$R = 3.333\,\Omega$$

Thus the combined resistance is around 3.3 Ω. It is important at this stage to consider what degree of accuracy is required. My calculator shows that the reciprocal of 0.3 is actually 3.33333333. This seems to be a very high degree of accuracy, but the accuracy is apparent rather than actual. We must consider how accurately the true values of the two resistive loads are known. In order for the calculator's answer to be realistic, we would need to know the resistance of the load to a far greater accuracy than could be measured without the most specialised equipment. We should also consider how accurately we *need* to know the answer. For most

purposes—and bearing in mind the very large number of variables that usually exist within any electric or electronic circuit—it is quite sufficient to say that the combined resistance is equal to approximately 3.3 Ω.

In practice, quite complex combinations of series and parallel resistances may be met with. Figure 2.8 illustrates such an instance. The rule for dealing with part of a circuit like this, and arriving at the combined resistance of all the loads, is to work out any obvious parallel or series combinations first, and progressively simplify the circuit. There is an obvious parallel combination in this figure, so we can work out the combined resistance of the two 100 Ω and two 50 Ω resistors first. Using a calculation like the one above, we arrive at a total resistance of about 16.7 Ω for the parallel combination of the four resistors. Figure 2.8 has now become more simple—the simplified version is illustrated in Figure 2.9a.

fig 2.8 *a complex series-parallel network*

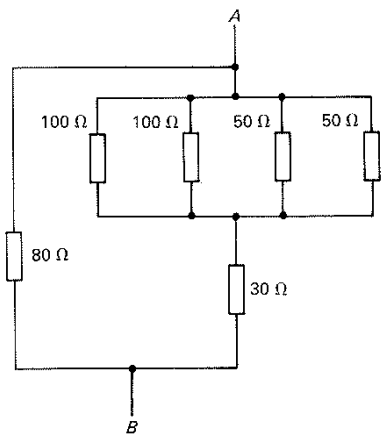
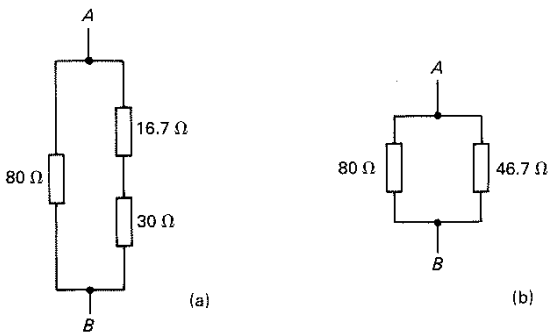


fig 2.9 *progressive simplifications of figure 2.8*



The series resistors, 16.7 Ω and 30 Ω, can simply be added together to give a total of 46.7 Ω. This leads us to Figure 2.9b, which is a simple parallel combination which can easily be worked out.

The total resistance between points *A* and *B* in Figure 2.8 is therefore 29.5 Ω–or, in the real world, about 30 Ω.

Change some of the values in the example given in Figure 2.8 and work through the calculations yourself, to make sure that you are completely happy with these simple calculations. These simple calculations will in fact be a lot simpler if you use a pocket calculator–preferably one that can give you reciprocals. Calculators are inexpensive and any student of electronics should always keep one handy.

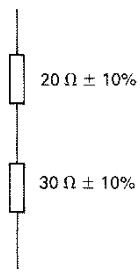
2.4 TOLERANCES AND PREFERRED VALUES

It should be clear that a load having a resistance of, say, 20 Ω is unlikely to have a value of *exactly* 20 Ω if you measure it accurately enough. Components having a specified resistance, such as *resistors* used in electronic circuits, may be marked with a specific value of resistance. But inevitably the components will vary slightly and not all resistors stamped '20 Ω' will necessarily have a resistance of exactly this value. Manufacturing tolerances may be quite wide.

When designing circuits, it is important to bear in mind that components may not be exactly what they seem. As we saw above, this affects the calculations we make, and in any complex circuit we would have to specify circuit values of resistance, voltage and current in terms of a range of values. A resistor marked 20 Ω might, for example, have a resistance of 18 Ω, or 22 Ω. These two limits represent a departure from the marked value of around ± 10 per cent. This is in fact a typical manufacturing tolerance for resistors.

But let us look at the effect of two such resistors connected in a series circuit. Figure 2.10 shows this. We would expect, using simple addition,

fig 2.10 *the effect of tolerance in values*



that the combined resistance of these two resistors would be 50 Ω. We might base our circuit design on this, and might, for example, predict that this very simple 'circuit' would draw 100 mA from a 5 volt supply. However, let us consider two extremes. If *both* resistors happen to be at the limits of their tolerance range, then the combined resistance might be 55 Ω. Our circuit would be drawing only 90 mA. If the resistors happened to be at the low end of the tolerance range, then a combined value of 46 Ω could give us a 108 mA current drain from our 5 volt supply. According to the function of the circuit, this might or might not be important.

In some cases we need to be able to specify a resistor more accurately than ± 10 per cent. Manufacturers of these components therefore produce different ranges, according to the degree of tolerance permitted. The closer the tolerance, the more expensive the resistor, so it is commercially unwise to specify component accuracy greater than the circuit requires. Because of the manufacturing tolerance, and because of the impossibility of producing *all* possible values of resistor, manufacturers restrict themselves to an internationally standardised range of values. Figure 2.11 lists the available values. Intermediate values are not generally available, but can be made by combining preferred values. It is important to be cautious when doing this, and to bear in mind the tolerance range of the resistors you are using. There is little point, for example, in adding a 3.3 Ω resistor

fig 2.11 *preferred values of resistors*

E12 series	E24 series
10	10
	11
12	12
	13
15	15
	16
18	18
	20
22	22
	24
27	27
	30
33	33
	36
39	39
	43
47	47
	51
56	56
	62
68	68
	75
82	82
	91

Preferred values of resistance

E12 series is available in all types of resistor

E24 series is available in close-tolerance or high-stability only

All values are obtainable in multiples or submultiples of 10, e.g. 2.2 Ω, 22 Ω, 220 Ω, 2.2 kΩ, 22 kΩ, 220 kΩ, 2.2 MΩ, 22 MΩ

Resistance values less than 10 Ω and more than 10 MΩ are uncommon, and may not be available in all types of resistor

to a 47 Ω resistor in an attempt to produce something that is exactly 50 Ω if the tolerance of the combination is likely to be ± 5 Ω. Manufacturers in general restrict themselves to three or four ranges of tolerances. Typically, small resistors would have a tolerance of ± 5 per cent, or in some cases ± 10 per cent. Close-tolerance components of ± 2½ per cent or even ± 1 per cent are available, but expensive. Components made to this accuracy would normally be used only in accurate measuring equipment or in particularly critical parts of some circuits.

2.5 CIRCUIT DIAGRAMS

I have already used circuit diagrams, and it is usually fairly clear how circuit diagrams are laid out. You will see many in this book, and should get a good idea of the way circuit diagrams should be drawn. It is normal to produce diagrams with vertical and horizontal lines, and as few diagonal or angled lines as possible. It is usually convenient to lay the circuit out so that the potential difference in the circuit (or voltage) varies from the top to the bottom of the page: this means arranging the diagram with one power supply line at the top and the other at the bottom in most cases. Circuit symbols are used to represent different components and different types of components. We have already seen the circuit symbol for a resistor and the circuit symbol for a battery. Other component symbols will be dealt with in context, and explanations will be given where necessary. There are rules about lines in the diagrams crossing, and these must be obeyed rigidly if the diagram is not to be misinterpreted. Figure 2.12 shows some of the conventions. Figure 2.12a illustrates wires crossing, but not connecting. Figures 2.12b and 2.12c show wires that are connected—the little dot is used to indicate a connection.

Sometimes where wires are crossing but not connected the illustrator will use a 'bridge' to emphasise the fact that the wires are not connected. This is illustrated in Figure 2.12d; this particular convention is now disappearing. Figure 2.12e illustrates something which should *never* appear in circuit diagrams. In all cases where wires are crossing and connected the illustration in Figure 2.12c should be used. The reason for this is that, if the circuit diagram is printed, the point where the wires cross may 'fill in' in the printing or photocopying process and make it appear that they should be connected when in fact the circuit designer intended that they should not. As you work through this book you should look at the way the circuit diagrams are laid out, and try to draw neat, *clear* diagrams yourself. Often, this may mean redrawing—but it is worth the trouble.

In the next chapter we will be looking at further 'passive' components. We have already considered resistors, but we shall look at them in more

fig 2.12 *circuit diagrams*

(a) *wires crossing but not connected*

(b) *wires joined*

(c) *wires crossing and connected*

(d) *a non-preferred convention for wires crossing but not connected*

(e) *something that should not appear on any circuit diagram*

detail—and in the context of the way they are used in circuits. We shall also look at the other two main classes of passive components, *capacitors* and *inductors.*

QUESTIONS FOR CHAPTERS 1 AND 2

1. Describe the difference between *linear* and *digital* electronic systems.
2. What is the unit of electric current? What is the unit of electrical potential?
3. Calculate the value of the combined value of the following resistors connected in series: 33 kΩ, 18 kΩ, 4.7 kΩ. If a battery supplies a current of 215 μA through the resistors when they are connected to it, what is the battery's voltage?
4. What is the combined resistance of the following resistors connected in parallel: 220 Ω, 100 Ω, 470 Ω?
5. A 10 kΩ and a 5.6 kΩ resistor are connected in series. If both resistors have a manufacturing tolerance of ± 10 per cent, what, approximately, are the maximum and minimum values of resistance we would expect to measure across the combination?

CHAPTER 3

PASSIVE COMPONENTS AND POWER SUPPLIES

The components used in electronic circuits are divided into two main classes. These are referred to as 'active' and 'passive' components. Active components use the movement of electrons either through a vacuum or through a semiconductor to perform some function in the circuit. Passive devices, on the other hand, do not control the movements of the electrons, but rather allow the electrons to flow through them. It could be argued that the simplest of all passive devices is a length of wire!

3.1 RESISTORS

Resistors are components intended to insert a known amount of resistance into an electrical circuit. As we saw in the last chapter, they are available in a range of different values and tolerances. It will also be evident to anyone who has looked at an electronic circuit that they are also available in a range of *sizes*. The sizes relate to the amount of power that the resistor can dissipate. An electric current passing through a resistor will always cause a certain amount of energy to be dissipated as heat. The amount of energy dissipated can be calculated by the simple formula:

$$P = VI \text{ watts}$$

where P represents power.

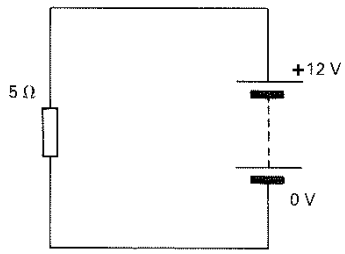
Let us consider the case of a 5 Ω resistor connected across a 12 volt battery. This is illustrated in Figure 3.1. We can use Ohm's Law to calculate the amount of current flowing through the resistor. This is given as:

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

$$I = \frac{12}{5}$$

$$I = 2.4 \text{ A}$$

fig 3.1 *power: a simple circuit like this will dissipate heat*



It is clear that the voltage, when measured across the resistor, must be equal to the battery voltage, provided we neglect any resistance in the wires. We can now use the formula to calculate the amount of power that the resistor is dissipating as heat:

$$P = VI$$
$$P = 12 \times 2.4$$
$$P = 28.8 \text{ watts}$$

28.8 watts is a respectable amount of heat—the above calculation might be made for a car's rear-window heating element. It is quite obvious that a tiny component such as a resistor one might find in an electronic circuit could not possibly dissipate this much heat without damaging itself. In fact, a resistor designed to dissipate around 30 watts would be quite large —perhaps 5 cm long. It would also be made in such a way that the body of the resistor could become quite hot. In practice, the designers of most electronic circuits try to avoid such large heat outputs which can cause problems in cooling the equipment.

Resistors used in electronic circuits are designed with power dissipations of 0.25, 0.5 or 1 watt. Larger sizes are available, but are less commonly used.

The same formula as the one above is used to calculate the power dissipation of series and parallel combinations. Figure 3.2 shows such circuits. Figure 3.2a illustrates two resistors connected in series. It is quite clear that the same current flows through both resistors. However, when measured across each resistor, the voltage will be different. The total potential difference is given as 5 volts. We can find the voltage across each individual resistor by using Ohm's Law. We know that the total resistance of the circuit is 3.2 kΩ and can use Ohm's Law to calculate that the current flowing through the circuit is about 1.562 mA. Using the form of Ohm's

Law $V = IR$, we can calculate the voltage drop across each resistor. For the first resistor:

$$V = 1000 \times 1.562\,\text{mV}$$
$$V = 1.562$$

And for the other resistor:

$$V = 2200 \times 1.562$$
$$V = 3.44$$

Now that we know the voltage dropped across each resistor we can calculate the amount of power that each resistor is dissipating. For the 1 kΩ resistor this is:

$$P = 1.562 \times 1.562$$
$$P = 2.43\,\text{mW}$$

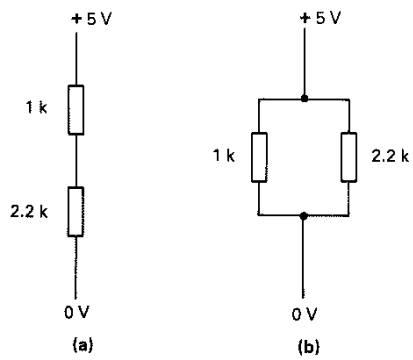
And for the 2.2 kΩ resistor the power dissipated is given as:

$$P = 1.562 \times 3.44$$
$$P = 5.37\,\text{mW}$$

Look carefully at the units used in these calculations! The amounts of power dissipated are very small, typical, in fact, of the sort of thing we will become used to in electronic circuits. The smallest commonly available size of resistor (0.125 watts) is substantially larger than is necessary to dissipate this amount of power.

Exactly the same kind of calculation is used to work out the power dissipations in resistors in parallel. In this case—illustrated in Figure 3.2b—the voltage across both resistors is the same (5 V), but the current is different. We can use Ohm's law to ascertain the current flowing through

fig 3.2 *resistors in series and in parallel*



each resistor, and arrive at 5 mA and 2.27 mA for the 1 kΩ and 2.2 kΩ resistors respectively. We can now use our power calculation to determine that the 1 kΩ resistor is dissipating 5 x 5 = 25 mW. The 2.2 kΩ resistor is dissipating 2.27 x 5 = 11.35 mW. If you are in any doubt about this calculation, work it out on a piece of paper. Another, equally valid way of working out power dissipations in series or parallel combinations of two resistors is to calculate the total dissipation of the pair (very simply done) and then use a ratio calculation to determine how much power is dissipated by which resistor. Note that in *series* combinations it is always the resistor with the *higher* value that dissipates the most power. In *parallel* combinations, it is the resistor with the lower resistance that dissipates more power.

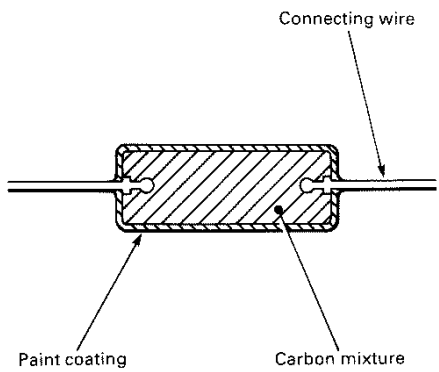
Before leaving resistors, we should look at the physical structure of the three main types of resistor. These are all electrically comparable—there is only one kind of resistance—but they are mechanically different, being suitable for slightly different applications.

Types of resistor

The most usual type of resistor is the solid carbon resistor (see Figure 3.3). Its structure is very simple, consisting of a small cylinder of carbon which is mixed with a non-conductor. A connecting wire is fixed into each end, and the resistor is given a coat of paint to protect it from moisture which might alter the resistance.

Resistors are always marked with a *colour code* to indicate the value. The colour code consists of three or four coloured bands painted round the resistor body. This system is used because it makes the resistor's value

fig 3.3 *a solid carbon resistor (this component might be anything from a few millimetres to a few tens of millimetres long)*



visible from any direction—a printed label could be hard to read with the component in place on a crowded board. Also, a painted or printed value could easily get rubbed off, whereas painted bands are relatively permanent.

The first three bands of the colour code represent the value of the resistor in ohms. Bands 1 and 2 are the two digits of the value, and band 3 represents the number of zeros following the first two digits. Figure 3.4 gives the resistor colour code; the standard is international. The fourth band is used to indicate the tolerance of the resistor's stated value.

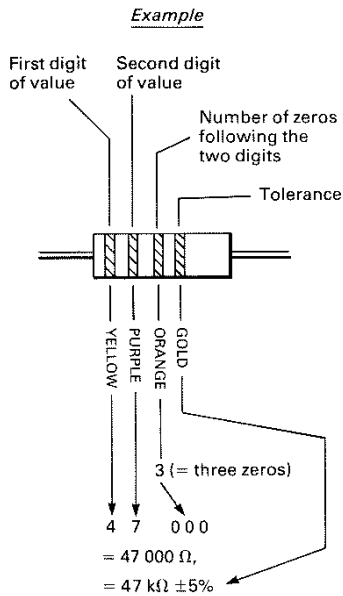
fig 3.4 *the international standard resistor colour code*

Values

0	black
1	brown
2	red
3	orange
4	yellow
5	green
6	blue
7	purple
8	grey
9	white

Tolerances

±20%	no band
±10%	silver
±5%	gold
±1%	pink

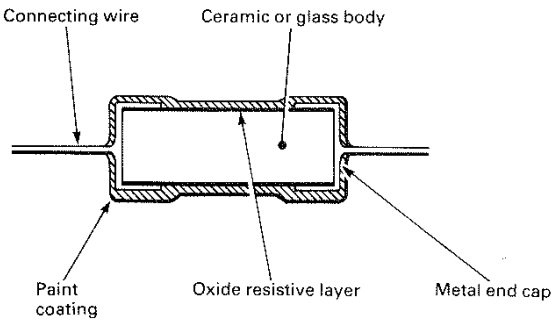


Where values of resistance less than two digits are required, the two bands are followed by a gold band. Thus 4.7 Ω would be represented as yellow, purple, gold, plus a tolerance band.

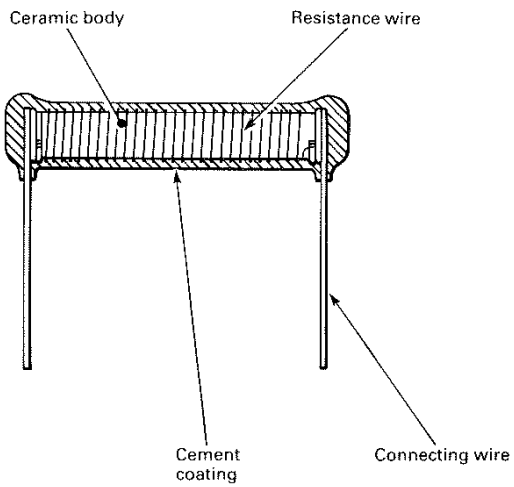
The next kind of resistor is the *metal oxide* or *metal glaze* resistor. This looks rather like the carbon resistor from the outside, but the internal structure is different. Figure 3.5a shows a cross-section.

Metal oxide resistors can be made to closer tolerances than carbon resistors, and change their resistance less with changes in temperature. For this reason they are sometimes called *high-stability* resistors. The resistance of a metal oxide resistor changes by approximately 250 parts per million

fig 3.5



(a) *a metal oxide resistor*



(b) *for high power applications, a wire-wound resistor is used*

per °C. This compares with about 1200 parts per million per °C for carbon resistors.

Both carbon and metal oxide resistors are made in a range of stock sizes, from about 0.125 W dissipation to about 3 W. Sometimes it is necessary to have resistors that can cope with higher powers; for this, *wire-wound* resistors are used. The wire-wound resistor is shown in Figure 3.5b.

It is easy to make wire-wound resistors in low resistance values, down to fractions of an ohm. High resistance values use wire of low conductivity,

requiring many turns of fine gauge wire as well. The maximum practical value for a wire-wound resistor is a few tens of kilohms, at least for components that are a reasonable size. Power ratings range from 1–50 W in the stock sizes; there is no limit to the size in practice, and larger special-purpose wire-wound resistors are common.

It is possible to make *precision wire-wound resistors* in which the resistance is specified to a very close tolerance, within ± 0.1 per cent. Such resistors are expensive and would be used only in measuring equipment.

Variable resistors

For applications such as volume controls and other 'user controls' in electronic equipment, it is often necessary to have a resistor that can be altered in resistance by means of a control knob. Such resistors are called *variable resistors*, or *potentiometers*. They consist of a resistive *track*, made with a connection at either end. A movable *brush*, generally made of a non-corroding metal, can be moved along the track; an electrical connection to the brush allows a variable resistance to be obtained between either end of the track and the brush.

Three forms of variable resistor are shown in Figure 3.6b. Note that the shaft (or slot) that moves the brush is generally connected to the brush; it is important to know this for safety reasons.

Variable resistors are available in various shapes and sizes, with power dissipations from around 0.25 W upwards. Tracks are either carbon, conductive ceramic ('cermet') or wire-wound. Resistance ranges are available between fractions of an ohm and a few megohms.

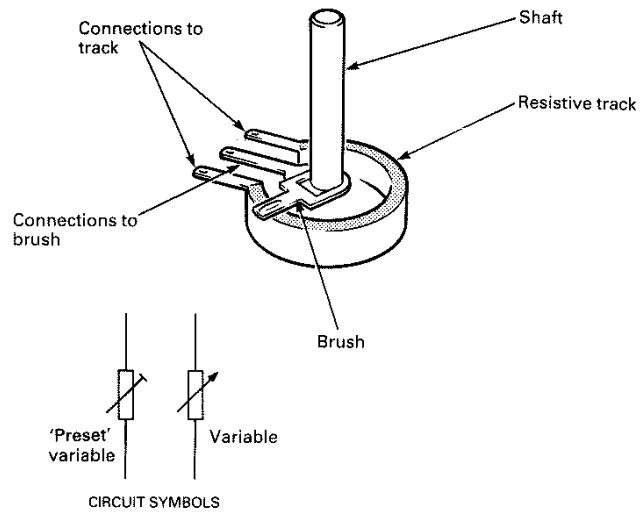
The track can be made in such a way that the resistance increases smoothly along the track—the usual, *linear*, type of potentiometer. For some audio uses (such as volume controls) *logarithmic* potentiometers are made in which the resistance increases according to a logarithmic law rather than a linear law. For most purposes you need only remember that the way logarithmic potentiometers control volume approximates the way the human ear responds to sounds of different loudness. If you use a linear potentiometer for a volume control, the effect of the control seems to be 'all at one end' of the scale.

3.2 CAPACITORS

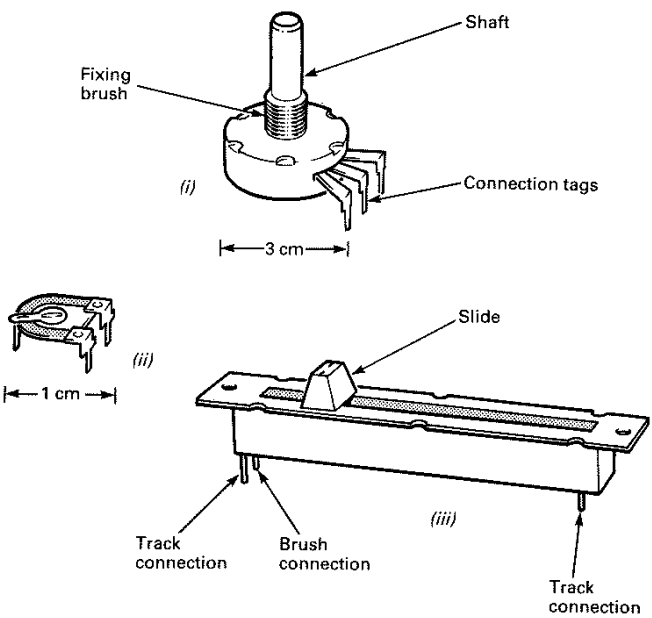
Next to resistors, the most commonly encountered component is the *capacitor*. A capacitor is a component that can store electric charge. In essence, it consists of two flat parallel plates, very close to each other, but separated by an insulator (see Figure 3.7).

When the capacitor is connected to a voltage supply, a current will flow through the circuit (see Figure 3.8). Electrons are stored in one of the

fig 3.6



(a) *principle of the potentiometer or variable resistor*



(b) *three different types of potentiometer*

fig 3.7 *schematic view of a capacitor*

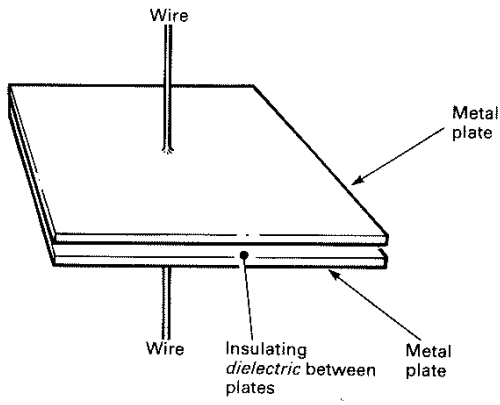
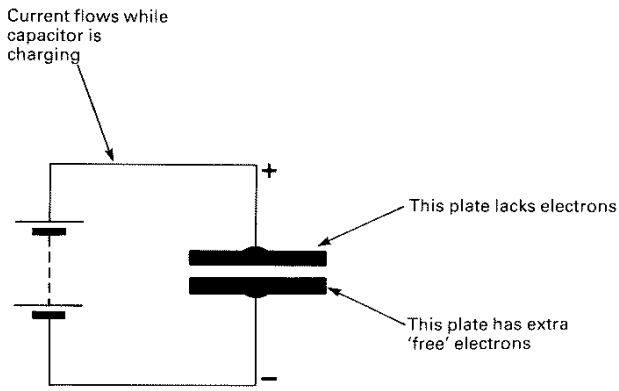


fig 3.8 *a capacitor connected to a power supply; current will flow until the capacitor is charged*



plates of the capacitor; in the other there is a shortage of electrons. In this state the capacitor is said to be *charged*, and if it is disconnected from the supply, the imbalance in potential between the two plates will remain.

If the charged capacitor is connected in a circuit, it will, for a short time, act as a voltage source, just like a battery. This can be demonstrated quite nicely with nothing more than a large capacitor, a 9 V battery such as PP9, and a capacitor rated at 10 V, 1000 μF (see below regarding *units of capacitance*). Note that the capacitor will be an *electrolytic* type (again, see below) and must be connected to the battery with its + terminal

towards the positive (+) terminal of the battery. The demonstration is shown in Figure 3.9.

Try the experiment again, this time allowing two minutes between charging the capacitor and discharging it into the bulb. Try waiting progressively longer, and you will see that the charge gradually leaks away on its own. This is due to the imperfection of the insulator separating the plates, allowing a tiny *leakage current* to flow.

A graph of the current flowing through a charging and then discharging capacitor is compared with the voltage measured across it in Figure 3.10.

The unit of capacitance is the *farad*, named after Faraday. One farad is a very large unit of capacitance, in the context of electronic circuits, and the smaller derived units of capacitance are used. The largest practical unit is the *microfarad*, symbol μF. Although values of capacitance in the range

fig 3.9 *a simple experiment for demonstrating the way in which capacitors can store power; a large capacitor is first charged up from a 9 volt supply, and then discharged into a small lamp*

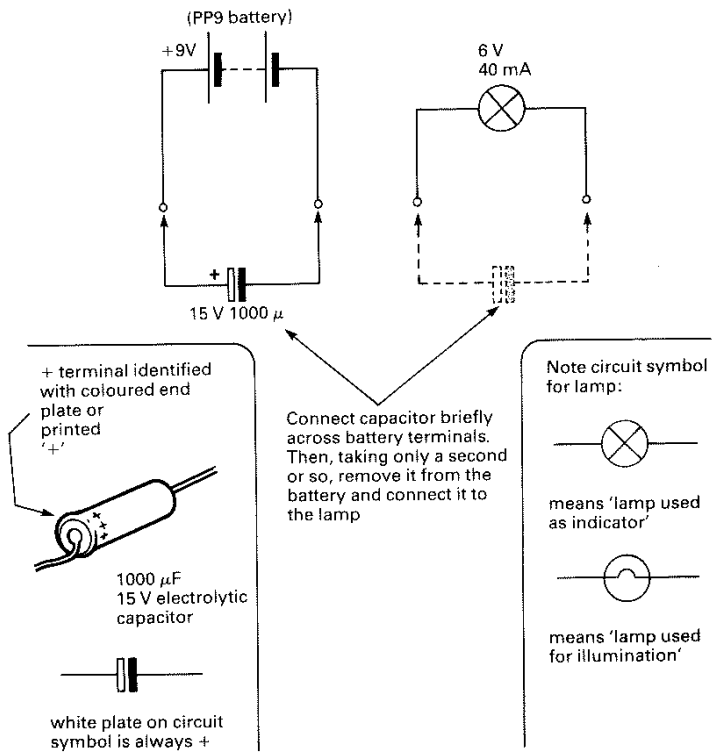
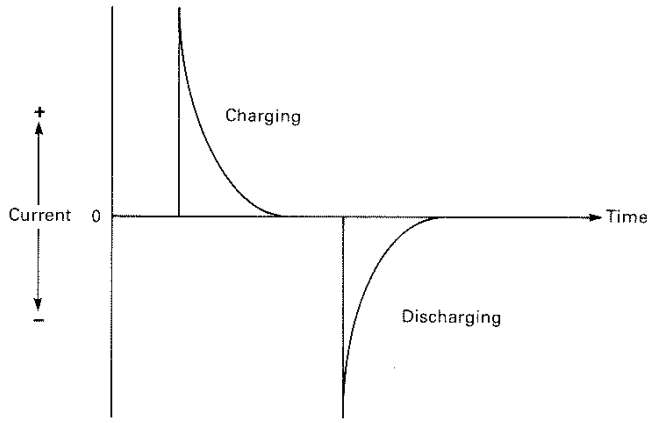


fig 3.10 *a graph showing a capacitor charging and discharging*



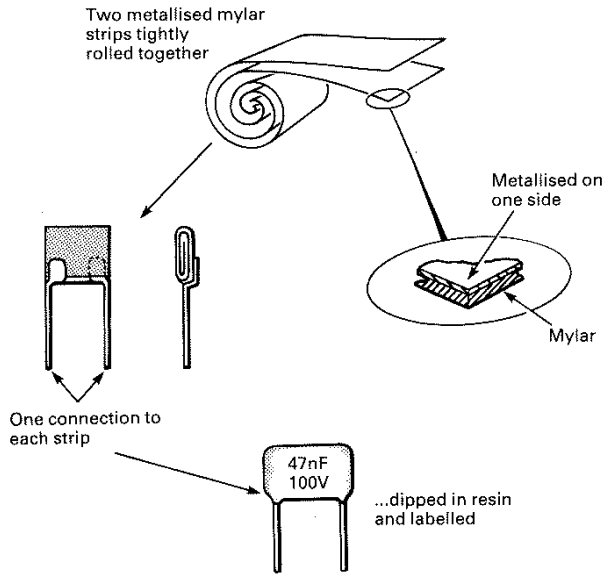
of thousands of microfarad are sometimes used, the millifarad is never encountered—a capacitor would be marked 10 000 μF, for example, not 10 mF. If you do see a capacitor marked 'mF', it is almost certainly meant to be μF, with that manufacturer (Far Eastern?) using a non-standard abbreviation. In addition to the μF, the nF and the pF are commonly used units.

There are many different types of capacitors, according to the use to which the component is to be put, and also to the operating conditions. The capacitance of the component is determined by three factors: the area of the plates, the separation of the plates, and the insulating material that separates them, known as the *dielectric*. For the same dielectric material, the closer and larger the plates, the greater the capacitance. Another factor determines the thickness of the dielectric, and that is the maximum voltage to which it is going to be subjected. If the dielectric is very thin, it will *break down* with a relatively low voltage applied to the plates of the capacitor. Once the dielectric has been damaged, the capacitor is useless. High voltage capacitors need thicker dielectrics, to withstand the higher voltage. To produce the same capacitance, the plates have to be larger in area—so the component is bigger.

The simplest type of capacitor uses a roll of very thin aluminium foil, interleaved with a very thin plastic dielectric such as *mylar*. A physically smaller capacitor can be made by actually plating the aluminium on to one side of the mylar (see Figure 3.11). For higher voltages, *polyester, polystyrene* or *polycarbonate* plastic material is used as the dielectric.

Ceramic capacitors are used where small values of capacitance and large values of leakage current are acceptable—the ceramic capacitor is inexpen-

fig 3.11 *a typical capacitor, used in electronics work—the 'metallised film' type*



sive. A thin ceramic dielectric is metallised on each side, and coated with a thick protective layer, usually applied by dipping the component.

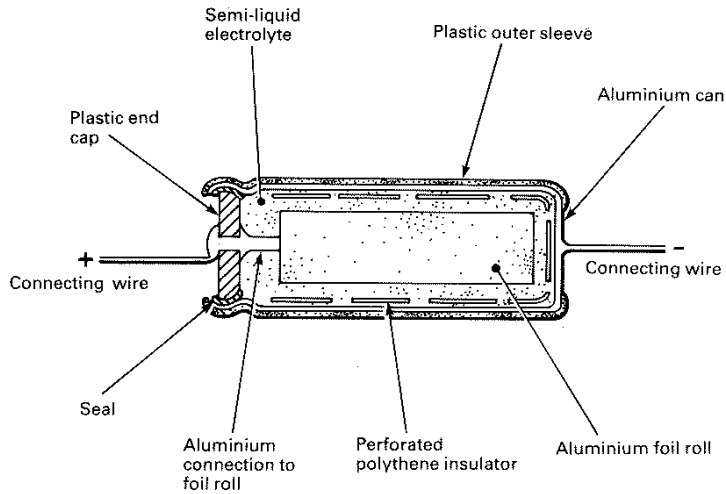
Both plastic film and ceramic capacitors are available in the range 10 pF–1 μF, though plastic film types may be obtained in larger values.

Where really high values of capacitance are needed, *electrolytic capacitors* are used. Electrolytic capacitors give very large values of capacitance in a small component, at the expense of a wide tolerance in the marked value (−25 to +50 per cent) and the necessity for connecting the capacitor so that one terminal is always positive.

The most commonly used type of electrolytic capacitor is the *aluminium* electrolytic capacitor. The construction is given in Figure 3.12. After manufacture, the capacitor is connected to a controlled current source which electrochemically deposits a layer of aluminium oxide on the surface of the 'positive' plate. The aluminium oxide makes an excellent dielectric, with very good dielectric strength (resistance to voltage applied across the plates). Since the layer is chemically deposited it is very thin.

The electrolytic capacitor must not be subjected to voltages applied in the wrong direction, or the aluminium oxide layer will be moved off the positive plate, and back into the electrolyte again.

fig 3.12 *construction of an aluminium electrolytic capacitor*



Variable capacitors

Capacitors are made that can be varied in value. Either air or thin mica sheets are used as the dielectric, so the capacitance of variable capacitors is usually low. *Rotary* or *compression* types are made. In the former, the capacitance is altered by changing the amount of overlap of the two sets of plates; see Figure 3.13a. In compression trimmers, the spacing between plates is altered, as in Figure 3.13b.

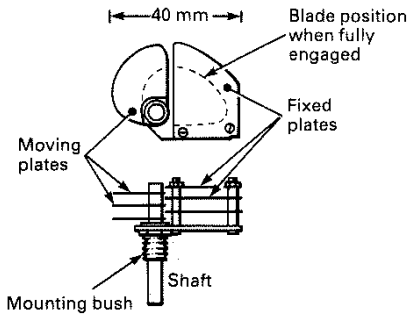
Variable capacitors can be obtained with maximum values from 2 pF to 500 pF.

Capacitors as frequency-sensitive resistors

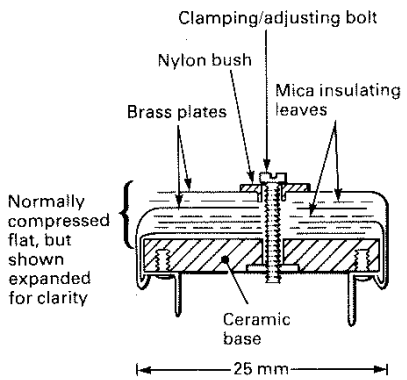
Once a capacitor has charged up, it will not pass current when connected to a direct voltage supply. However, if we *reverse* the polarity of the supply, the capacitor will permit a current to flow until it has charged up again, with the opposite plates positive and negative. The capacitor will then, once again, block the flow of current. If the frequency of an alternating voltage applied to a capacitor is high enough, the capacitor will actually behave as if it were a low value resistor, as it will never become charged in either direction. At lower frequencies, the capacitor will appear to have a higher resistance, and at zero frequency (d.c.) it will, as we have seen, have an infinitely high resistance if you disregard the leakage current.

The precise value of resistance shown by the capacitor will depend upon the capacitance, the applied voltage and the frequency. The property

fig 3.13



(a) *a variable capacitor of the 'air space' type (this component would be used for tuning a radio receiver, for example)*



(b) *a 'compression trimmer', a form of variable capacitor used to preset circuit values (adjustment is by means of a screw)*

is known as *reactance*. It enables the circuit designer to use the capacitor to block or accept different frequencies in a variety of different circuit configurations, many of which you will meet later in this book.

Capacitors in series and parallel

Capacitors can be used in series and in parallel. When used in series, the working voltage is the sum of the two working voltages, so two capacitors intended for a 10 V maximum supply voltage could be used, in series, with a 20 V supply.

The calculation of capacitor values in series and parallel is similar to the

calculation used for resistors in series and parallel, but in the opposite case. Thus for capacitors in *parallel* the formula:

$$C = C_1 + C_2 + C_3 + \ldots$$

is used, simply adding the values together. For capacitors in *series*:

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

gives the total value of capacitance.

Unfortunately, there is no internationally agreed colour code for capacitors, so manufacturers usually stamp the value on a capacitor, using figures in the normal way. A few capacitors are marked with coloured bands, but it is wise to look at the maker's catalogue to check what the code means.

3.3 INDUCTORS

The third major passive component is the inductor. This is a coil of insulated wire that may, or may not, be wound over a ferrous metal former. When a coil is connected in a circuit, as in Figure 3.14, the flow of current through the coil causes an electromagnetic field to be created around the coil. Building the field absorbs energy, and the electromagnetism produces its own current in the coil, opposing the direction of the applied current. The effect is that, when current is first applied to the inductor, it seems to have a high resistance, reducing the current flow through it. Once the electromagnetic field is established, the resistance drops. A graph of the current flowing through an inductor, and the voltage measured across it, is given in Figure 3.15.

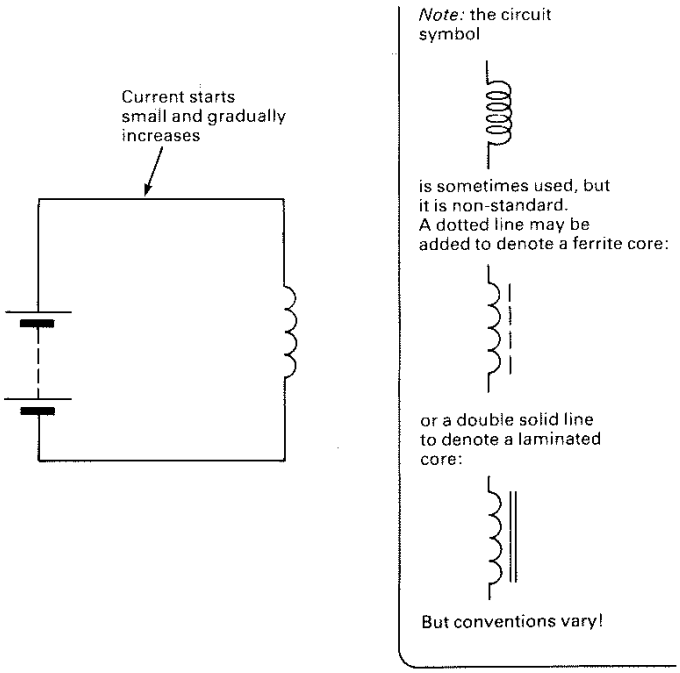
Compare this graph with the one in Figure 3.10, and you will see that the inductor is the 'opposite' of a capacitor! Once the current flowing through the inductor is constant, the only opposition to current flow comes from the resistance of the coil wire—usually small compared with the apparent resistance (called inductive *reactance*) when the current is building up.

The opposing force that prevents current flowing right away is *induced current*, trying to flow the opposite way. We can make great use of this phenomenon in constructing a *transformer*.

The unit of inductance is the *henry*, symbol H. In electronics, mH and μH are commonly found, but the henry, like the farad, is rather a large unit.

Inductors are available as a range of components, in small inductance values from a few μH to a few mH. However, there are so many variable

fig 3.14 *an inductor connected in a simple circuit; the current rises to a fixed level, determined by the resistance of the inductor's windings*



factors that a 'stock range' of all required types would be too large for a supplier to produce. Where inductors are needed (and they are relatively rare in modern circuits) they may well be specified in terms of wire thickness, number of turns, core type, etc.

The unit used to measure inductance is the *henry*, symbol H. The henry is a descriptive unit, and is the potential difference generated across an inductor when the current flowing through it is changing at the rate of one ampere per second.

3.4 TRANSFORMERS

Transformers make use of *mutual inductance*, in which a current flowing in a coil produces an electromagnetic field, which in turn induces a current to flow in a second coil wound over the first one.

The construction of a transformer is shown in Figure 3.16. The iron

fig 3.15 *graphs showing voltage and current in the circuit of Figure 3.14*

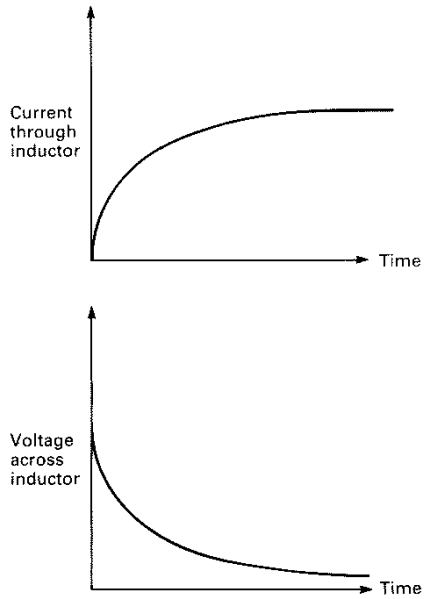
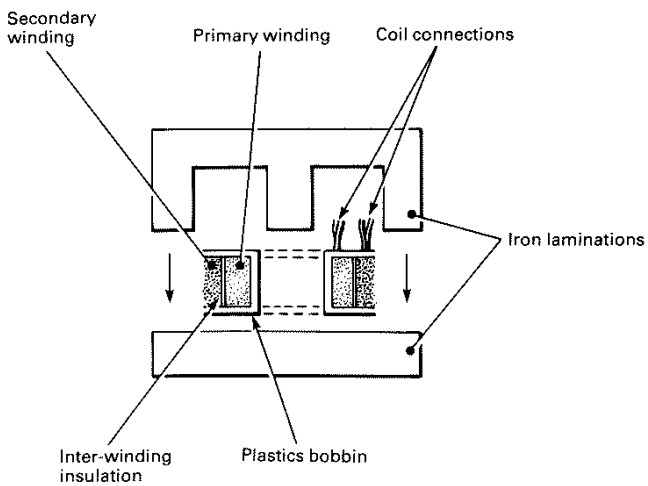


fig 3.16 *a typical small transformer, showing the iron laminations*



laminated core is used to concentrate the electromagnetism and thus improve the efficiency. It is important that the iron laminations are insulated from each other; if they are not, the core itself will behave as if it were a one-turn coil, and a current will be induced in it. The current would be very large, and cause the transformer to overheat.

The ratio of the *input voltage* to the primary winding to the *output voltage* from the secondary winding depends on the turns ratio of the coils, and is approximately equal to that ratio. Thus a transformer with 2000 turns on the primary and 1000 turns on the secondary will have an output voltage that is half the input voltage.

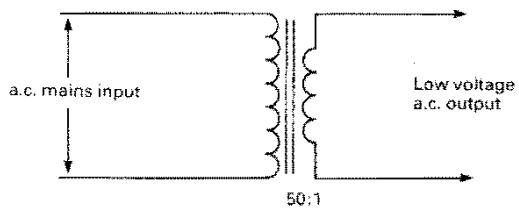
Since currents are induced only by *changing* magnetic fields, the transformer will operate only if the input voltage is constantly changing. The transformer could be used with alternating current (such as the mains supply), and one of the main uses for a transformer in electronics is to reduce the voltage of the a.c. mains to a lower level, suitable for electronic circuits. This is illustrated in Figure 3.17.

Transformers are rated according to the turns ratio, the power handling capability (in watts), and the type of application. *Power transformers*, intended for power supplies, usually have mains voltage primaries, and a high standard of insulation between the primary and secondary coils, for safety.

The secondary winding may have a range of connections, or *taps*, so the transformer can be used in different applications. It is important to realise that the transformer simply transforms voltages; there is no net power gain, always a small loss. If a transformer has a 200 V primary and a 1000 V secondary, the voltage will be increased by a factor of five times, but it will require more than five times the current in the primary than will be available from the secondary. You can trade current for voltage and vice versa, but the *power* will always be slightly less at the output, due to various losses in the transformer (dissipated, as usual, in the form of heat).

Any component having a coil carrying current will have an inductive characteristic, though it may be swamped by other factors (resistance or capacitance). This may sometimes be important. For example, a *relay* is

fig 3.17 *circuit symbol for a transformer with a laminated core*



an electromagnetic switch, the coil being used to operate a mechanical switch. However, the coil has inductance. A relay operating from a 6 V supply can, when suddenly disconnected, produce an output pulse of 100 V or more as the current flowing through it drops very rapidly to zero. This voltage may be high enough to harm semiconductor components, and precautions should be taken to ensure that this 'spike' of high voltage is made harmless.

3.5 POWER SUPPLIES

I am going to introduce, prematurely, a semiconductor component that is described fully in Chapter 7. This is the *diode*, a sort of electronic one-way street that will let current pass through it in one direction but not in the other direction.

Diodes are used in power supply circuits, in association with transformers. A transformer is used to reduce the voltage of the a.c. mains to something more usable in electronic circuits, say 10 V in this case. However, the output is still a.c., and most electronic circuits require a direct current supply. The transformer's output can be converted, or *rectified*, by one or more diodes. Figure 3.18 shows a transformer operating a circuit with a bulb. The waveform of the voltage across the bulb is shown also.

Now compare the waveform with the one in Figure 3.19. The diode has blocked current flowing in one direction, so the top connection of the bulb is always positive. Only half the power reaches the bulb, so this is not a very efficient design. If four diodes are used, in a configuration called a *bridge rectifier*, the whole of the current available from the transformer can be passed through the bulb. Follow the current path through Figure 3.20, first for one direction of output from the secondary, and then from the other.

fig 3.18 *waveforms at the input and output of a transformer driving a lamp*

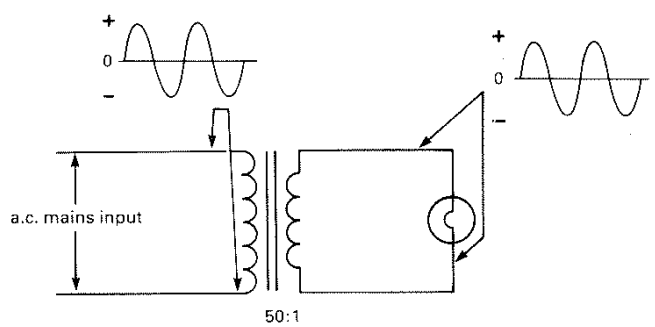


fig 3.19 *waveforms for half-wave rectified direct current, supplied by the secondary of the transformer, through a diode*

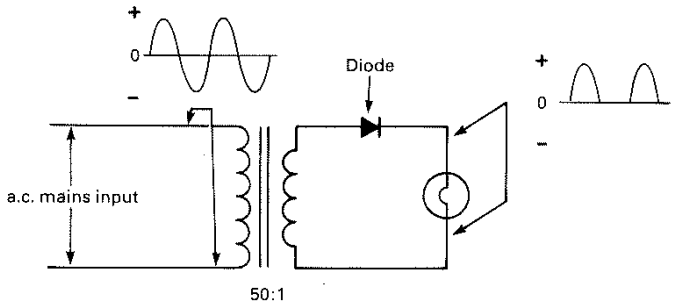
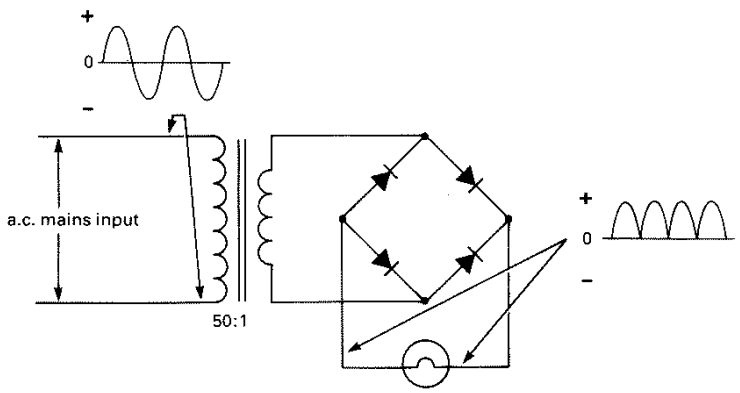
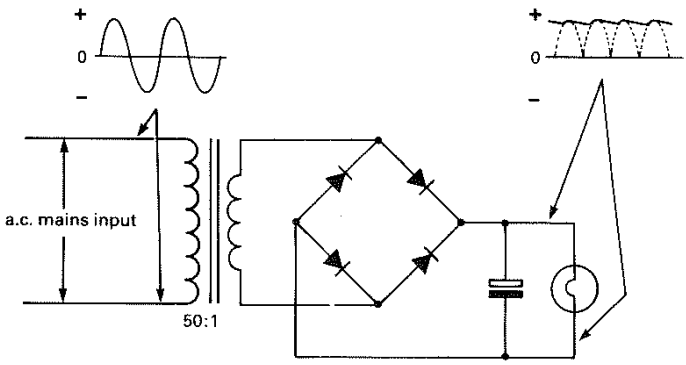


fig 3.20 *the diode 'bridge rectifier' configuration, used to provide full-wave rectification of an a.c. supply (although both halves of the a.c. input waveform pass through the lamp, the current flow through the lamp is not smooth—and this would upset the operation of many circuits)*



The supply to the bulb is now d.c., but constantly changing in voltage. The voltage supplied to the bulb (or any other circuit) can be *smoothed* by means of a large-value capacitor connected in parallel with the load (the bulb in this case). The effect of the capacitor is to supply the required current when the voltage drops during each half-cycle of the mains supply. When the voltage rises, the capacitor is recharged, so that it has energy available to fill in the 'gaps' in the supply. Figure 3.21 illustrates this.

fig 3.21 *the bridge rectifier with smoothing capacitor; a large capacitor is used to provide power during the 'gaps' in the rectified waveform*



This circuit is the basis of almost all power supply systems used in electronic circuits. Although much more sophisticated 'regulated' power supplies are often used—as you will discover when you get to Chapter 20—the transformer, bridge rectifier and smoothing capacitor feature in all the designs.

Before leaving the subject of diodes and circuits, I will use this opportunity to mention an important fact that electronics engineers, secretly, are a little ashamed of.

The current flowing through the power supply circuits above is flowing in the direction of the arrowhead of the diode symbol. The arrowhead to indicate current flow occurs over and over again in circuit diagrams. Current, we say, always flows from positive to negative.

Actually, it doesn't.

Electrons move the other way, from negative to positive, opposite to what has been called 'conventional current'. This is an odd fact, but one that only rarely complicates circuit design and analysis. The reason for it, frankly, is a Mistake. When the study of electricity was in its infancy, there was a need to standardise (in books and notes) on the direction of current flow. In the absence of our knowledge of electrons and atoms, someone sat down to perform an experiment that would settle the problem of current flow once and for all. It might even have been Volta, or Faraday, I can't remember. The experiment consisted of setting up equipment that would make a very long electric spark, and then repeatedly watching the spark. Eventually the experimenter made up his mind that the spark always jumped from *here* to *here*. It was at this point that the Mistake was made.

When, years later, it was discovered that (i) electrons go from negative to positive, and (ii) electric sparks jump too fast to see which way they are going anyway, 'conventional current' was established in all the textbooks, and it would have been disastrous to have tried to change it.

So we *still* say that 'conventional current' flows from positive to negative, except when we are discussing semiconductor physics, where electron flow is what matters.

3.6 **PRINTED CIRCUITS**

Before leaving passive components, mention must be made of that most ubiquitous of passive components, the *printed circuit board*. It used to be necessary to build a *chassis* for every piece of electronic circuitry, with insulating strips carrying rows of tags for component connections. The printed circuit board (PCB), however, has made possible small, lightweight circuits, easy to construct and repair.

The PCB consists of a strong plastic laminate or fibreglass laminate board, with a thin sheet of copper bonded to one (or sometimes both) side(s). The sheet is drilled with holes that enable the circuit components to be pushed into place with their connecting leads through the holes. The underneath (copper) side of the PCB is made in a pattern which corresponds to the required wiring pattern. The components are soldered in place, and the PCB provides both support for the parts and the necessary electrical interconnections. A completed PCB can be seen in Figure 15.23. Appendix 1 shows some examples of PCB patterns.

In production, the PCB is produced using automatic processes. The required pattern is produced as a photographic negative, and this is printed on to the PCB, which has been coated on the copper side with a photographic emulsion. The PCB is then developed like a photograph, leaving emulsion only on the parts that were exposed to light.

Next, the PCB goes into a bath of a powerful chemical that will dissolve copper—ferric chloride is often used. The ferric chloride etches away all the copper, *except* where it is protected by the emulsion. The emulsion is cleaned off, leaving the copper pattern. Holes are drilled or punched as required, and the board is ready for assembly (although the copper is often coated with a thin layer of solder by a process known as *roller tinning* to improve ease of soldering.

Anyone can make PCBs. For single PCBs, the circuit is simply drawn on to the copper with an etch-resistant ink (special pens are available) or with rub-on transfers. The rest of the process is the same as that used commercially.

Be careful with the strong ferric chloride solution it is corrosive, poisonous, and will gallop through most metals. Use safety goggles, rubber gloves, plastic dishes, plastic tongs, and *don't* work on a stainless steel drainer!

* * * * * * * * * * * * * * * * * * *

Electricity and Passive Components

Chapters 2 and 3 have been intended as revision only. If you want to know more about electricity, there are many good textbooks available. If, however, you have found the first three chapters of this book fairly easy going, you will be able to understand the rest without too much trouble.

QUESTIONS

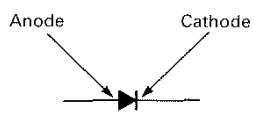
1. State the value and tolerance of the following resistor colour codes, as read from the resistors in a circuit board in a radio receiver: *brown, black, red, silver*; *brown, grey, yellow, silver*; *red, black, black, silver*; *gold, red, blue, green.*
2. What is the largest commonly used unit of capacitance?
3. Three capacitors are connected in parallel. Their values are 4.7 μF, 10 μF and 3.3 μF. What is the combined capacitance? If the tolerance of each is +20 per cent and −10 per cent, what are the maximum and minimum values of capacitance to be expected for the combination?
4. The following two capacitors are connected in series: 3.3 nF, 2.2 nF. What is the combined capacitance?
5. A transformer is taking a current of 500 mA from a 220 V a.c. supply. If the output of the transformer is at 440 V, what is the maximum current the transformer could deliver from its secondary winding?

THE pn JUNCTION DIODE

The special properties of semiconductors are utilised to make a wide variety of electronic components, ranging from diodes to microprocessors. A detailed understanding of one of the simplest semiconductor devices, the *pn junction diode*, is an important prelude to a study of more complicated devices, and it is here that we begin.

A diode is basically the electrical equivalent of a one-way valve. It normally allows electric current to flow through it in one direction only. The symbol for the diode is given in Figure 7.1.

fig 7.1 *circuit symbol for a diode*



The arrowhead in the diagram shows the permitted direction of conventional current flow (remember, electron current goes the other way). Diodes can be made to handle currents varying from microamps to hundreds of amps—a selection of types is shown in Figure 7.2.

A 'perfect' diode would probably have infinite resistance in one direction, the 'reverse' direction, and zero resistance in the forward direction. A real diode exhibits rather different characteristics, the most notable of which is a *forward voltage drop*, which is constant regardless (almost) of the current being passed by the diode. Compare Figures 7.3 and 7.4. In Figure 7.3 the diode is *reverse-biased* and does not conduct. The lamp remains out and the meter shows the whole battery voltage appearing across the diode. In Figure 7.4 the diode is *forward-biased*, the lamp lights, and the meter shows 0.7 V across the diode. By using bulbs of different wattages to pass different currents through the circuit we can show that the 0.7 V is constant. The diode used in this demonstration is a silicon

fig 7.2 *a selection of diodes of various power ratings, from 1 A to 20 A*

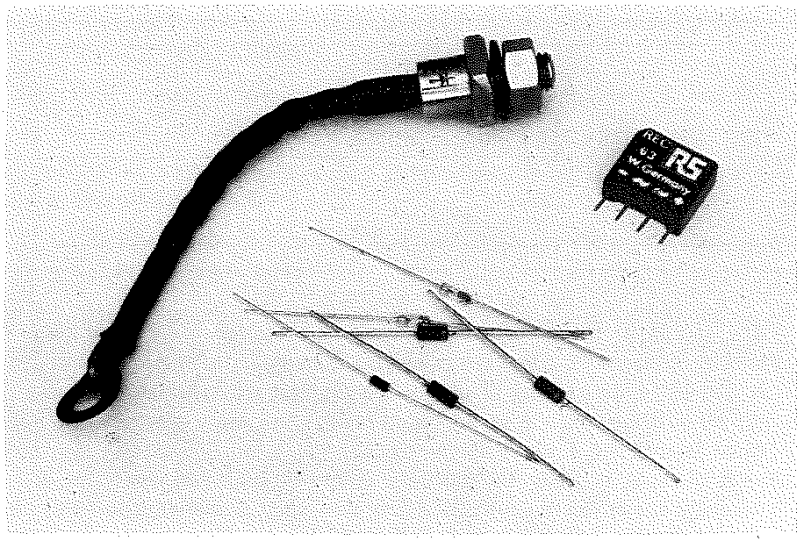
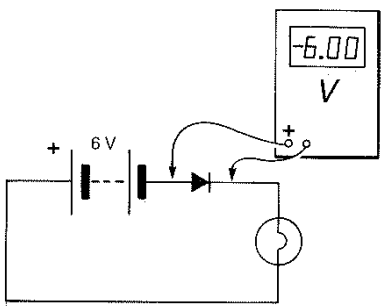


fig 7.3 *a circuit to reverse-bias a diode; note the meter is connected with its '+' to the battery '—', which results in a negative reading*



diode. If we were to replace it with one using a germanium semiconductor, the forward voltage drop would be constant at 0.3 V. What causes the voltage drop? To find out we must look at the way the diode functions at an atomic level, and go back to the energy-level diagrams.

Figure 7.5 shows an energy-level diagram for a piece of p-type silicon and a piece of n-type silicon; the two materials are not in contact.

fig 7.4 *a circuit to forward-bias a diode: the forward current is limited by the lamp*

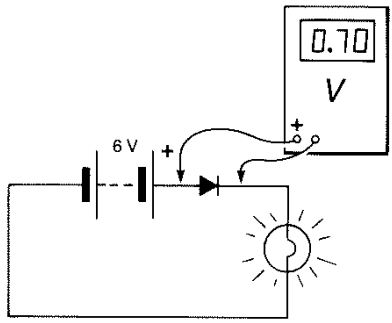
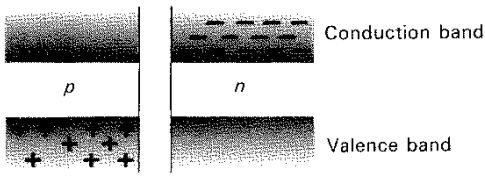


fig 7.5 *energy levels for two separate pieces of semiconductor*

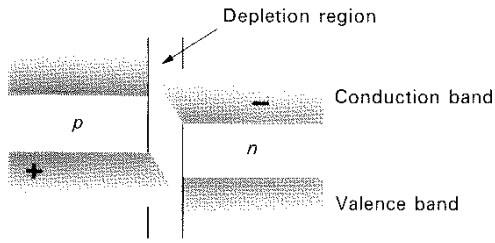


Now we place the two pieces of silicon in intimate contact (fuse them together for best results) and the energy-level diagram looks like Figure 7.6.

A few moments thought will show what has happened. The *p*-type material has extra holes and the *n*-type material has extra electrons. When the two types of semiconductor are put in contact the 'extra' electrons in the *n*-type material flow across the junction to fill up the holes in the *p*-type material. The electrons drift from the *n*-type to the *p*-type through the conduction band, and then fall down into the valence band of the *p*-type material to annihilate one hole for each electron. The holes also move, drifting from the *p*-type to the *n*-type semiconductor in the valence band.

Eventually an equilibrium is reached—but the *n*-type material has lost electrons to the *p*-type, and lost energy in the process. The *p*-type, on the other hand, has gained energy and the result is the state of affairs represented by Figure 7.6. The 'hill' in the diagram represents an area in the junction between the two semiconductor materials about 0.5 μm wide, called the *depletion region*, or sometimes the *transition region*. In this

fig 7.6 *energy levels for two pieces of semiconductor, one p-type and one n-type, in contact with each other*

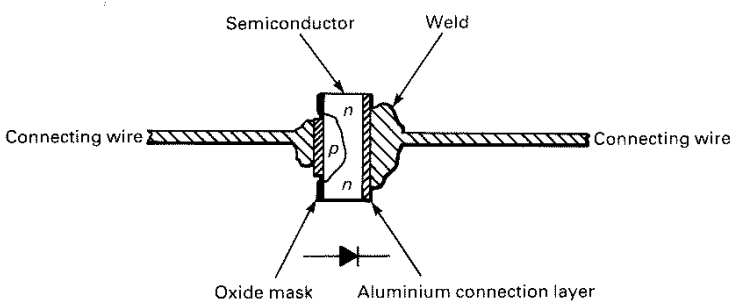


region there are no holes or free electrons. The 'height' of the hill, i.e. the difference in energy levels, can be measured in volts: for silicon it is 0.7 V, and for germanium it is 0.3 V.

The junction between the *p*-type and *n*-type semiconductor is, in fact, the active part of the *pn* diode. A typical low-power *pn* junction diode might be constructed as shown in Figure 7.7.

Details of the way the diode is manufactured can be found in Chapter 11, but for the time being it is enough to regard the diode as a block of *p*-type semiconductor in contact with a block of *n*-type semiconductor.

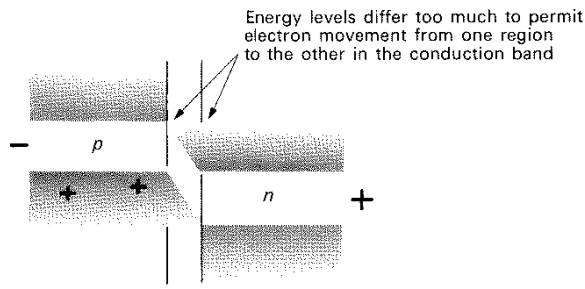
fig 7.7 *construction of a low-power diode (the circuit symbol underneath indicates the direction of current flow)*



7.1 REVERSE-BIASING THE *pn* DIODE

Assume the diode is connected as in Figure 7.3, so that it is reversed-biased (non-conducting). The energy-level diagram for this state is shown in Figure 7.8.

fig 7.8 *energy levels in the reverse-biased diode*



The energy of the *p*-type material is increased by the negative battery potential, while that of the *n*-type is decreased. Look at the thickness of the conduction and valence bands and at their relative positions in the two different materials, on either side of the depletion region. No parts of the conduction bands or the valence bands have the same energy levels. Electrons cannot move from the *p*-type to the *n*-type material through the conduction band because even the least energetic electrons have too much energy. Similarly, holes in the valence band of the *n*-type material all have too little energy to drift into the valence band of the *p*-type. The energy levels of the bands on either side of the depletion region are too different to allow either electrons or holes to move from one type of material to the other.

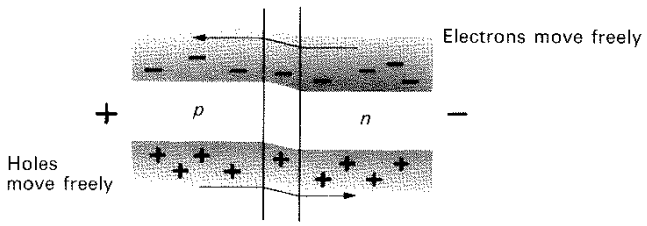
It follows that if electrons and holes are unable to cross the depletion region, there can be no current flow, for electric current is a movement of electrons (or holes) along a conductor.

7.2 FORWARD-BIASING THE *pn* DIODE

Let us reverse the situation, and connect the diode as in Figure 7.4, i.e. to forward-bias it. The energy-level diagram for this state is given in Figure 7.9. This time, the battery potential decreases the energy of the *p*-type semiconductor, and increases the energy of the *n*-type. When the potential applied is equal to or more than 0.7 V (assuming we are using a silicon

diode and not a germanium one) the two types of material are at the same energy level, and the conduction and valence bands coincide. Both electrons and holes can move freely between the two types of semiconductor, and electric current can flow through the diode.

fig 7.9 *energy levels in the forward-biased diode*



The reason for the 0.7 V forward voltage drop of the silicon diode should now be clear. A potential of 0.7 V must be applied to the *pn* junction to equalise the energy levels, and this is effectively subtracted from the potential available to push electric current through the diode.

7.3 POWER DISSIPATION IN THE DIODE

It is the forward voltage drop that sets an upper limit on the amount of current we can get through a diode. The lamp in Figure 7.4 might use, for example, 1 A, and the battery might be 12 V. In such a case a current of about 1 A flows in the circuit—but some of the power is 'wasted' in the diode as a result of the forward voltage drop, and is dissipated from the junction as heat. The amount of power that the diode dissipates can be calculated very simply using the formula:

$$P = V_f I$$

where P is the power (in watts), I is the current flowing (in amps), and V_f is the forward voltage drop of the diode in question. With 1 A flowing, the diode will dissipate 0.7 x 1 watts, or 0.7 W. When a diode is used in a circuit its power rating must exceed the expected dissipation, preferably by a comfortable margin.

In practice, the formula is complicated by the fact that the diode gets hot. The intrinsic conductivity of semiconductors increases with temperature (the bands get thicker) and this has the effect of lowering the forward

voltage drop as the temperature rises. For both silicon and germanium the forward voltage drop decreases at the rate of about 2.5 mV/°C temperature rise.

Since the junction temperature of a silicon diode may exceed 100°C with the diode operating near its maximum current, temperature effects can become quite significant. The 'normal' forward voltage drop of 0.7 V is reduced to 0.45 V for a 100°C temperature rise. For most purposes this effect can be discounted, but in circuits where the voltage drop across the diode is important, we have to bear it in mind. It is not, incidentally, possible to reduce the forward voltage drop to zero—at very high temperatures, the junction melts. Silicon is much more tolerant of temperature than germanium. For this reason, and also because the reverse leakage current is much lower (see below), silicon diodes are the norm and germanium diodes are relatively rare—they are used only in applications where a low forward voltage drop is essential.

7.4 REVERSE LEAKAGE CURRENT

Diodes are 'imperfect' in another way. When connected in a circuit that reverse-biases it (Figure 7.3 again) a diode allows a very small current to flow. This current is called the *reverse leakage current* and is partly due to thermally generated electron-hole pairs and partly to leakage across the surface of the diode. The reverse leakage current is very small for silicon diodes, in the region of 2 to 20 nA. For almost all purposes this minute current can be ignored. The reverse leakage current for germanium diodes is much higher, typically 2 to 20 μA, and this may be important in some circuits.

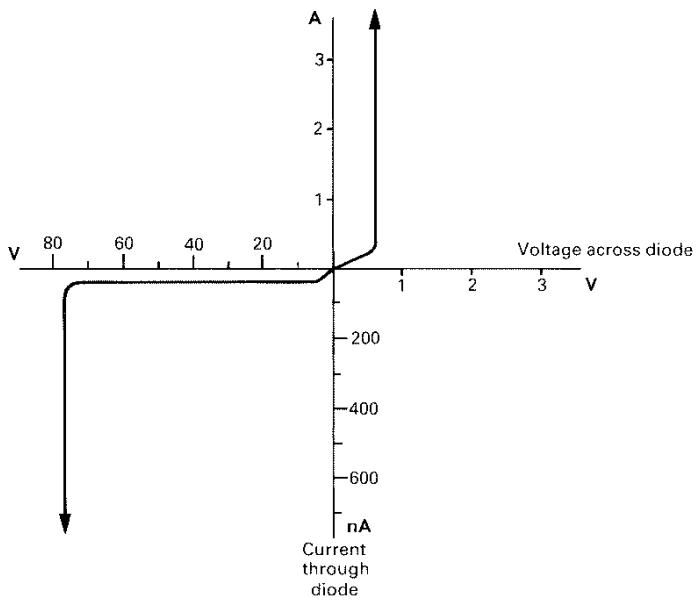
The reverse leakage current is also temperature-dependent and for both silicon and germanium it roughly doubles for each 10°C rise in temperature.

7.5 REVERSE BREAKDOWN

As the voltage applied to a diode to reverse-bias it is increased, there comes a point at which *reverse breakdown* occurs. In normal *pn* diodes, this breakdown is catastrophic and destroys the diode. Depending on the construction of the diode, the point at which reverse breakdown occurs can be anywhere from fifty to a few hundred volts. Manufacturers always specify the safe reverse-bias voltage, which is normally referred to as *peak inverse voltage*. Thus a 50 p.i.v diode can be reverse-biased at 50 V without danger of breakdown taking place.

We are now in a position to draw a graph of a typical diode's conduction characteristics at room temperature (see Figure 7.10). Note that the

fig 7.10 *conduction characteristics of a silicon pn junction diode*

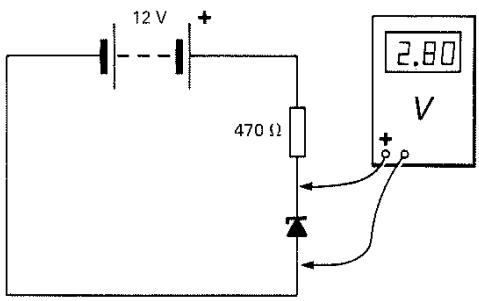


four scales of the graph are different, to show different aspects of the characteristics. On the right, the voltage across the diode is shown for different values of current through the diode. The graph continues straight up until the current reaches a value that causes overheating ($P = V_f I$) and melts the *pn* junction. On the left, the reverse leakage current is shown for increasing voltage; it remains steady at a few nanoamps until reverse breakdown occurs (at a substantially higher voltage than the 50 p.i.V manufacturer's maximum for this device) and the current increases to a value limited by factors other than the characteristics of the *pn* junction, which at this point will have just ceased to exist.

7.6 THE ZENER DIODE

There is one particular type of diode that makes use of a form of reverse breakdown to provide a constant 'reference' voltage. The circuit of Figure 7.11 shows the symbol for a Zener diode and also a suitable circuit to demonstrate its properties.

fig 7.11 *a Zener diode in a typical circuit*



In this circuit the diode is reverse-biased (Zener diodes are always used this way) and the meter shows a drop of 2.8 V across it. The voltage drop across a reverse-biased Zener diode will be substantially constant for all values of current up to the diode's limit, and for all values of voltage higher than the Zener voltage. A resistor is necessary to limit the current through the circuit; in this case the current is calculated by Ohm's Law:

$$\frac{12 - 2.8}{470} = 19.6\,\text{mA}$$

If we were to double the battery voltage to 24 V, the current through the circuit would increase to

$$\frac{24 - 2.8}{470} = 45\,\text{mA}$$

but the voltage across the Zener diode would remain at 2.8 V. The Zener diode is therefore a valuable device for *voltage regulation*, giving a fixed reference voltage over a range of input voltages.

Zener diodes are named after C. M. Zener, who in 1934 described the breakdown mechanism involved. In fact, Zener's description applies only to diodes with a Zener voltage of less than about 3 V. Above this, a mechanism called *avalanche breakdown* begins to take over. However, both types are generally lumped together under the heading 'Zener diodes', and are available in a range of voltages from 2 to 70, and with power ratings from 500 mW to 5 W. The power dissipation of the Zener diode is calculated by means of the usual formula

$$P = V_Z I$$

where V_Z is the Zener voltage. By way of example, the Zener diode in

Figure 7.11 is dissipating

$$2.8 \times 19.6 = 54.9\,\text{mW}$$

The conduction characteristics of a Zener diode can be drawn in a diagram similar to Figure 7.10. Look at Figure 7.10 again—the only difference in the Zener diode's characteristic is in the *reverse breakdown* (lower left quadrant), which will occur at the Zener voltage. The important factor—not shown in the diagram of characteristics—is that the Zener diode is undamaged by reverse breakdown provided that the current through the diode, and thus its heat dissipation, is limited to a safe value.

7.7 THE VARICAP DIODE

When a diode is reverse-biased, the junction behaves like a capacitor. The *p*-type and *n*-type regions form the plates of the capacitor, and the depletion region acts as a dielectric. You will remember that the capacitance of a capacitor depends on the thickness of the dielectric, and it is this fact that makes the *pn* diode useful as a variable capacitor. If the voltage used to reverse-bias the diode is increased, the difference in the energy-levels of the *p* and *n* regions is made greater. As this happens, more of the junction thickness is depleted of charge carriers, increasing the width of the depletion regions. Effectively, the capacitor's dielectric has become thicker, reducing the capacitance.

A *varicap diode* is a diode in which the change of capacitance with applied reverse voltage is enhanced as far as possible by the diode's physical construction. Even so, both the overall capacitance and the amount of change are fairly small, though they are sufficient to enable the varicap diode to be used in the tuning circuits of televisions, for example. A typical varicap application circuit is shown in Figure 7.12.

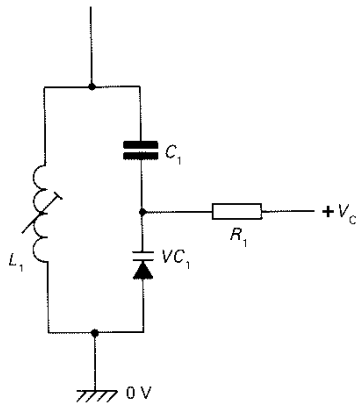
A tuned resonant circuit (see Chapter 15) is formed by L_1 and C_1, VC_1 (L_1 is made variable for initial adjustments) and a variable voltage applied via R_1. If this control voltage (V_C) is changed, the capacitance of the varicap diode is altered, varying the resonant frequency of the tuned circuit. The control voltage could come from a TV channel-changing system, probably based on digital electronic techniques, and the resonant circuit could be used in the TV tuner.

Typical varicap diodes have a capacitance swing from 6 pF to 20 pF, for an applied reverse voltage of 2 to 20 V.

7.8 THE LIGHT-EMITTING DIODE

The light-emitting diode (LED) is an extremely useful device, and replaces miniature incandescent lamps in a whole range of applications. Almost

fig 7.12 *a varactor tuning circuit, such as might be used in the tuner of a television; changing the voltage at $+V_c$ changes the capacitance of the varicap diode and alters the resonant frequency of the circuit*



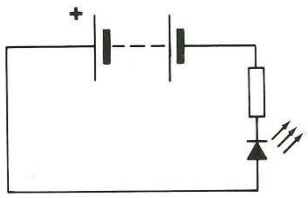
everybody is familiar with the LED display, used commonly for digital alarm clocks. LED displays are usually ruby red, but green and yellow is also available (although more expensive and therefore less common).

Like ordinary *pn* diodes, LEDs have no inherent current limiting characteristics, and must be used with a resistor to limit the current flowing—generally to around 20 mA for a small LED indicator. A suitable circuit is shown in Figure 7.13. It is clear from the circuit that LEDs are used in the forward-biased mode: indeed, they must be protected from reverse-bias as they usually have a reverse breakdown voltage of only a few volts. If breakdown occurs, the LED is destroyed. LED displays are described in rather more detail in Chapter 16.

Mechanism of light emission

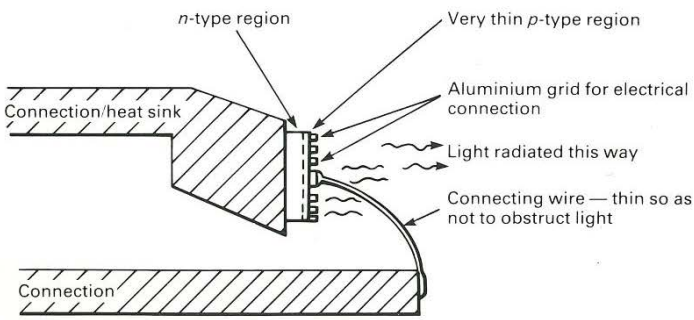
To move an electron from the valence band into the conduction band requires energy; and example of this is the creation of an electron-hole pair (see Figure 6.9). Similarly, if an electron falls down from the conduction band to annihilate a hole in the valence band, energy is given off. In normal *pn* diodes, this energy takes the form, ultimately, of heat. In the LED, part of the energy is given off as light, the wavelength (colour) of the light depending on the distance the electron falls, which in turn depends on the width of the forbidden band. LEDs are made with a variety of rather exotic materials: gallium arsenide phosphide for red or yellow, gallium phosphide for green.

fig 7.13 *a typical circuit for driving a LED; the resistor is necessary to limit the current through the LED, which is used in the forward-biased mode*



The physical design of a LED is important. Because the semiconductor material is rather opaque, most of the light produced never reaches the surface. To improve matters, the junction has to be very close to the surface, only about 1 μm or so. This means that the *p* or *n* region has to be very thin. Electrical connections to this layer have to be made in such a way that they do not obstruct the light. Figure 7.14 shows a typical LED. In practical applications LEDs are often pulsed. This is because the light output increases quite rapidly with increasing current, so a LED that is pulsed at 50 mA (and is on half the time) will look brighter than it would with a continuous 25 mA, while dissipating the same amount of heat. The pulse-repetition frequency has to be rapid enough for persistence of vision to make it look as if the LED is on continuously—more than about 50 Hz.

fig 7.14 *physical construction of a LED; light is emitted through the very thin p-type region and the device must be designed so that this light is obstructed as little as possible; a large connection area is made to the n-type region and this also serves to conduct heat away from the junction*



Like a *pn* diode, the LED exhibits a *forward voltage drop*. This varies according to the type and colour of the LED but is generally between 2 and 3 V.

QUESTIONS

1. What is the value for the forward voltage drop of (i) a silicon diode, (ii) a germanium diode?
2. If a silicon diode is passing 2 A, what is the amount of power that is likely to be dissipated by the diode?
3. How is the forward voltage drop of a diode affected by variations in temperature?
4. What is the main use of a Zener diode?
5. Which characteristic of semiconductor diodes is emphasised in varicap diodes?
6. A LED cannot be connected directly to a battery. What additional component is required, and why?

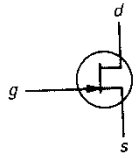
CHAPTER 9

FIELD-EFFECT TRANSISTORS

Field-effect transistors are a more recent development than bipolar transistors, and make use of a completely different mechanism to achieve amplification of a signal. Field-effect transistors (FETs) are *unipolar*, and involve only one type of charge carrier (electrons or holes) in their operation. There are also two major types of FET, junction-gate field-effect transistors (JUGFETs) and insulated-gate field-effect transistors (IGFETs). There are, as we shall see, subdivisions within these two classes.

Operationally, FETs are more similar to valves than are bipolar transistors. The main distinguishing characteristic compared with bipolar transistors is the fact that they are *voltage-controlled* rather than current-controlled. The circuit symbol is given in Figure 9.1; a voltage applied to the gate is varied to provide a corresponding charge in the resistance between the *source* and *drain*.

fig 9.1 *circuit symbol for an n-channel field-effect transistor*



Unlike the bipolar transistor's base connection, the gate of the FET has a very high input resistance, at least a few tens of megohms and in some cases, gigohms (see Figure 2.6). The amount of current drawn by the gate is therefore extremely small.

FETs can be used in amplifier circuits, just like bipolar transistors. Compare the circuit of Figure 9.2, which shows a typical JUGFET, with that in Figure 8.5. The difference is the lack of a gate (base) resistor;

because negligible current flows in the gate connection, such a resistor would make no difference to the operation of the circuit, adding a moderate amount of resistance to one that is extremely large in the first place. Just as there are *n*-type and *p*-type resistors, so there are *n*-channel and *p*-channel JUGFETs. Figure 9.3 shows a *p*-channel version, while those in Figures 9.1 and 9.2 are *n*-channel. As in the case of bipolar transistors, practical operation is the same but with all polarities reversed.

fig 9.2 *a simple amplifier using an n-channel FET*

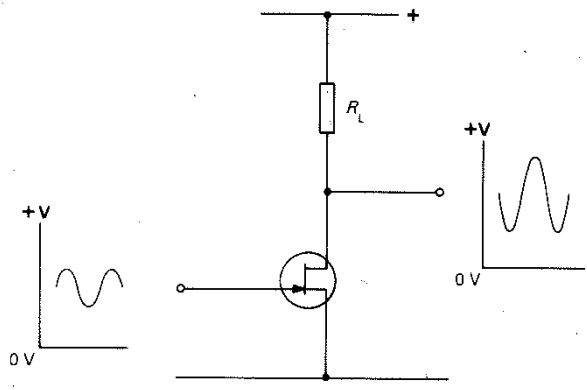
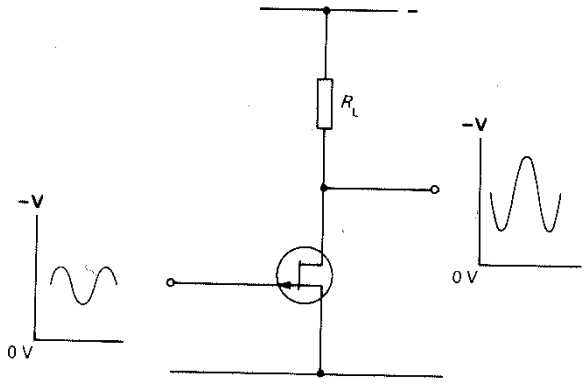


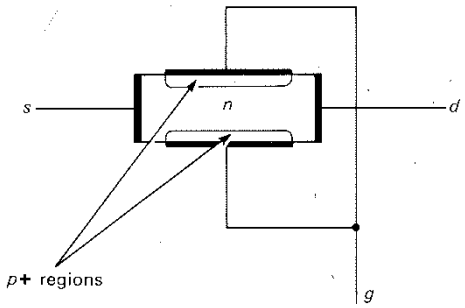
fig 9.3 *a simple amplifier using a p-channel FET (note the different polarity of the power supply, compared with Figure 9.2)*



9.1 FET PHYSICS

The JUGFET has a physical structure that can be represented by a diagram like the one in Figure 9.4, though in practice it is not easy to diffuse impurities with *both* sides of the wafer, and a rather different layout is used. Chapter 11 gives details.

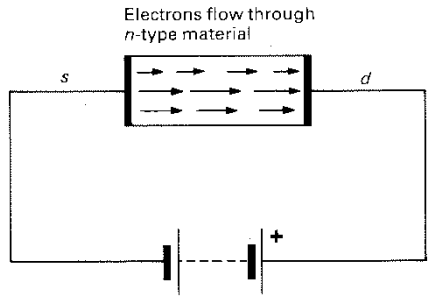
fig 9.4 *theoretical construction of an n-channel depletion-mode FET*



A bar of *n*-type semiconductor (almost invariably silicon) is made with shallow *p*-type regions in the upper and lower surfaces. These are connected to the gate terminal, and the two ends of the bar are connected to the source and drain.

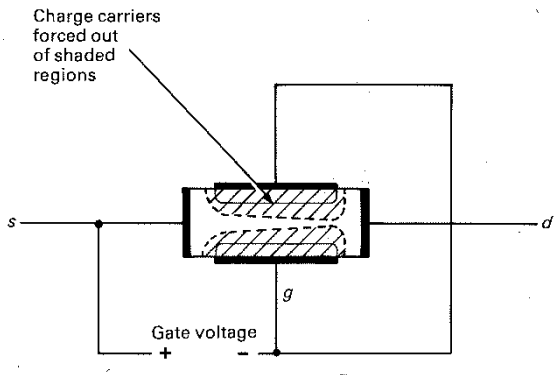
If the bar is connected to a voltage source, current will flow through it. Since the bar is symmetrical, it can flow either way, the source and drain being interchangeable. The current flow consists of electrons moving through the *n*-type semiconductor (Figure 9.5).

fig 9.5 *movement of electrons with no voltage applied to gate*



Now observe the effect of a negative potential applied to the gate regions. The junction between the *p* and *n* regions forms a reverse-biased diode (see Chapter 7) so no current flows, but an electric field extends into the *n*-type bar from the *p*-type regions. This charge forces current carriers (electrons) away from the region, reducing the amount of bar available for conducting the current between the source and drain (shown diagrammatically in Figure 9.6).

fig 9.6 *the electric field forces charge carriers away from the plates when a negative potential is applied to the gate*



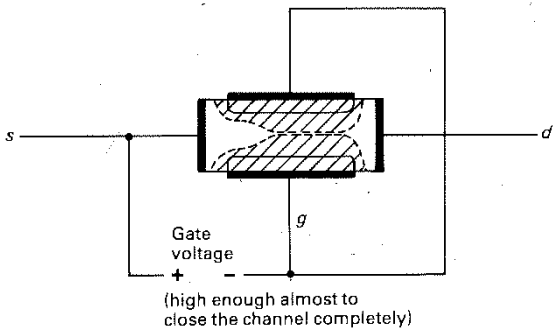
If the potential applied to the gate is made sufficiently negative, the electric field will extend across the whole thickness of the bar of *n*-type semiconductor, hardly any charge carriers will be available for current flow, and the current available from the drain will drop to a very low value (never to zero, for it is physically impossible for the channel to 'close' completely: see Figure 9.7).

Changes in the voltage applied to the gate will cause corresponding changes in the current flowing between the source and drain, which makes the operation of the FET very similar to that of a bipolar transistor.

9.2 THE INSULATED-GATE FET

Generally known as a MOSFET (metal-oxide semiconductor FET), the insulated-gate FET is one of the most important devices in the electronics industry. There are two basic categories of MOSFET, known as *depletion* MOSFETs and *enhancement* MOSFETs. They work on a different principle from the JUGFETs that we have been looking at so far in this chapter.

fig 9.7 *although the current flow can be greatly reduced, it is impossible to stop it completely—there will always be a small gap between the two areas without charge carriers*

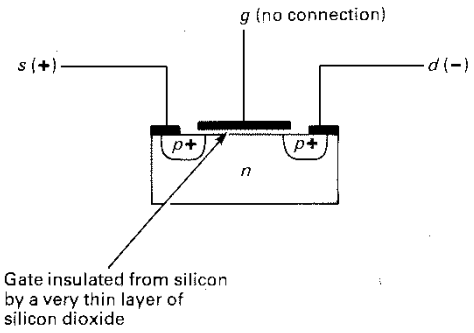


The structure of a *p*-channel enhancement MOSFET—again, a theoretical structure—is shown in Figure 9.8.

The most striking feature is the gate. It is insulated from the silicon by a thin layer of silicon dioxide. The layer is *very* thin, typically only about 0.1 μm thick. Although very thin, the silicon dioxide layer has an extremely high resistance, so the gate input resistance is very high indeed, at least 10 GΩ.

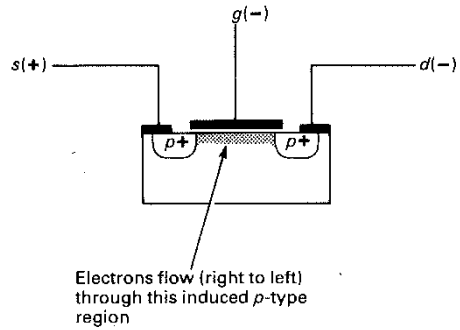
The *n*-type silicon has two regions of heavily doped *p*-type impurity, connected to the source and drain. With no applied gate voltage, one of the

fig 9.8 *theoretical construction of an n-channel enhancement mode FET (note that the gate is completely insulated from the rest of the structure by a very thin layer of silicon dioxide)*



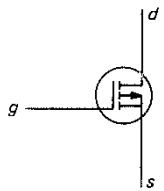
p–n junctions (depending on which way round the source and drain have been made) will act like a reverse-biased diode and block any flow of current. If a negative potential is applied to the gate electrode, holes from the *p*-type regions are attracted into the area immediately beneath the electrode. This effectively, if temporarily, makes a narrow *p*-type region just beneath the gate, illustrated in Figure 9.9.

fig 9.9 *a p-type region is induced in the n-type bar when the gate is made negative*



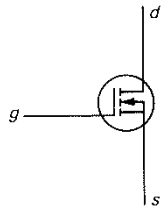
The blocking *p–n* junction is bypassed by this induced channel of *p*-type material, and electrons can flow through the device, between the source and drain. The circuit symbol for the *p*-channel enhancement MOSFET is given in Figure 9.10. The centre connection in the circuit symbol (with the arrowhead) is a connection to the silicon substrate (the chip itself). Often the connection is made internally to the source, but sometimes manufacturers fit a fourth lead so that the substrate can be used as a second 'gate', the conduction characteristics of the device then depending approximately on the difference in potential between the gate and substrate.

fig 9.10 *circuit symbol for a p-channel enhancement MOSFET*



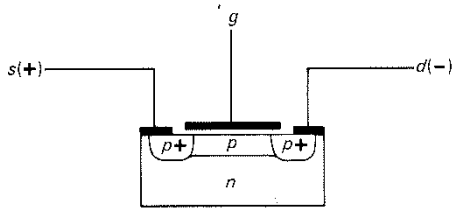
When the circuit symbol has an arrowhead pointing away from the gate, it symbolises a *p*-channel device. It is, however, equally possible to make an *n*-channel device, with all the polarities reversed, using *n*-regions diffused into a *p*-type bar. In this case the charge carriers are electrons attracted from the *n*-regions; otherwise operation is the same. The circuit symbol for an *n*-channel enhancement MOSFET is given in Figure 9.11.

fig 9.11 *circuit symbol for an n-channel enhancement MOSFET*



The second class of MOSFET is the depletion MOSFET. The structure is shown in Figure 9.12, and is similar to that of the enhancement MOSFET. Notice that there is an 'extra' region. A narrow strip of *p*-type impurity has been diffused into the space below the gate, so that the depletion MOSFET, with *no* signal applied to the gate, looks rather like the enhancement MOSFET when its gate is connected to make it conduct. Compare Figures 9.12 and 9.9.

fig 9.12 *theoretical structure of a depletion MOSFET*

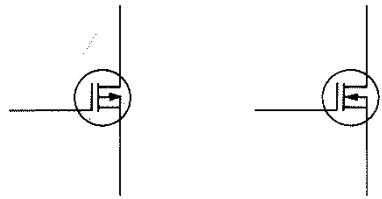


Applying a positive signal to the gate causes electrons from the *n*-type region to be attracted to the area under the gate electrode, neutralising some of the holes in the *p*-type channel and reducing the amount of current flowing between the source and drain. The higher the positive potential (for a *p*-channel device, of course), the more the source-drain current is cut off.

Actually the depletion MOSFET can also be used in the enhancement

mode as well. Applying a *negative* voltage to the gate of a *p*-channel device will increase the source–drain current by adding to the number of holes available as charge carriers. Figure 9.13 shows the circuit symbols for *p*-channel and *n*-channel depletion MOSFETs.

fig 9.13 *circuit symbols for (a) p-channel and (b) n-channel depletion MOSFETs*



9.3 USES OF MOSFETs

At first it seems paradoxical that, even though the MOSFET is a vitally important device, individual MOSFETs are rarely used. Only in the highest-quality communications receivers do we find MOSFETs used as a matter of common practice. MOSFET technology is used overwhelmingly in integrated circuits.

Before leaving FETs for the time being, let us just look at one or two interesting aspects of the devices:

(1) Because of the necessity for a very narrow channel, it is difficult to make FETs that will carry high currents. Most FETs are therefore low-power devices.

(2) Very high frequency operation of FETs is hampered by the internal capacitance, a byproduct of the very narrow regions. There is an effective capacitance of a few picofarads between the source (and drain) and the gate.

(3) It is possible to make FETs very small.

(4) Because of the very high input resistance and the thinness of the silicon dioxide insulating layer, MOSFETs are very liable to damage from high voltages accidentally applied between the gate and the other terminals. For this reason, most MOSFETs are made with a Zener diode connected between the gate and substrate. Normally, the Zener diode is non-conducting, but if the gate voltage rises too high, it conducts and discharges the gate. Even so, MOSFETs are easily damaged by electrostatic voltages. The

charge on a person wearing shoes with rubber or plastic soles can rise, on a dry day, to several kilovolts. This is often enough to destroy the MOSFET, Zener and all. Even *touching* the gate terminal can therefore destroy the device! The terminals of a MOSFET are generally shorted together with metal foil or conductive foam until it is installed in a circuit. Extra care is needed when servicing any equipment that might include MOSFETs.

QUESTIONS

1. What is the major difference in characteristics between IGFETS and JUGFETs?
2. Special precautions are needed when handling IGFETS—why?
3. Why are high-power FETs uncommon?
4. Draw the circuit symbols for (i) *p*-channel enhancement FET, (ii) *n*-channel enhancement FET, (iii) *p*-channel depletion FET, (iv) *n*-channel depletion FET.
5. What limits the higher frequency performance of FETs?

CHAPTER 11

FABRICATION TECHNIQUES AND AN INTRODUCTION TO MICROELECTRONICS

Chapter 7 covered the basic physics of semiconductor devices, and just touched on the methods used to manufacture the components themselves. Semiconductor manufacture, or 'fabrication' as it is more generally called in the industry, is a highly specialised and very difficult subject, but it is useful for the electronics engineer or technician to have some idea of the principles involved.

Various different semiconductor materials are used, but the most common (and cheapest) is silicon. In this chapter we shall look at the fabrication of semiconductor devices based on silicon, but you should bear in mind that roughly similar techniques—although with different materials—are used to deal with germanium and the other semiconductors.

11.1 PURE AND VERY PURE . . .

The first step is to take a single large crystal of pure silicon. This is not as easy as it seems, for silicon that only thirty years ago would have been called 'chemically pure' would be hopelessly contaminated for the purpose of semiconductor manufacture. In the early days of transistor manufacture (1960s) the purity of the raw material, usually germanium, was the major obstacle to reliable production. A process that had worked perfectly for weeks would suddenly start turning out 100 per cent rejects, and would have to be stopped. The batch of semiconductors would be thrown out, everything cleaned, and (with luck) the process might be restored to correct operation after a month. This sort of thing was one of the main reasons for the high cost of early transistors. Today, manufacturers of semiconductor-grade silicon aim for *no* impurity atoms at all. This level is never reached, but crystals having impurities less than 1 part in 1 000 000 000 are routinely made.

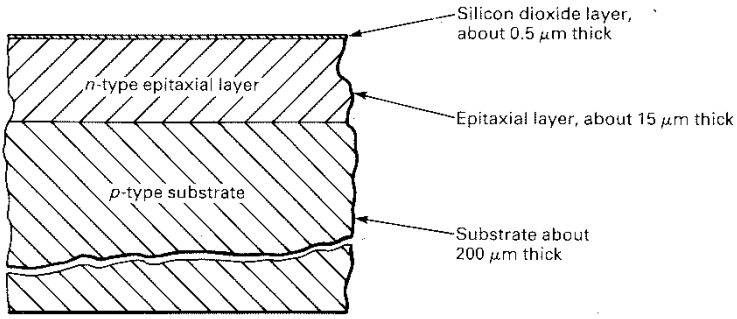
One way of making pure silicon crystals was described at the end of Chapter 6. We begin with the sausage-shaped single crystal of silicon that is drawn out of the pure silicon 'melt'.

The silicon crystal is sliced up (like salami) into circular *wafers*, typically 50 mm diameter and 0.5 mm thick. The surfaces are ground and polished perfectly flat, leaving the wafer about 0.2 mm thick. The wafer is finally cleaned with chemical cleaners.

The technique for making a single *pn* diode is as follows.

Beginning with a wafer of *p*-type silicon, made by adding a tiny amount of *p*-type impurity such as indium or boron to the pure silicon, an *n*-type *epitaxial layer* about 15 μm thick is 'grown' on the wafer. This is done by heating the wafer to about 1200°C in an atmosphere of silicon and hydrogen tetrachloride with a trace of antimony, phosphorus or arsenic. Next a thin (about 0.5 μm) layer of *silicon dioxide* is grown on top of the wafer by heating it to about 1000°C in an atmosphere of oxygen or steam. Silicon dioxide has three very useful properties. It is chemically rather inert and is not attacked by gases in the atmosphere, or indeed most other chemicals, even at high temperatures. Second, it is impervious and prevents the diffusion of impurities through it. Third, it is an excellent electrical insulator. The wafer at this stage is shown in cross-section in Figure 11.1.

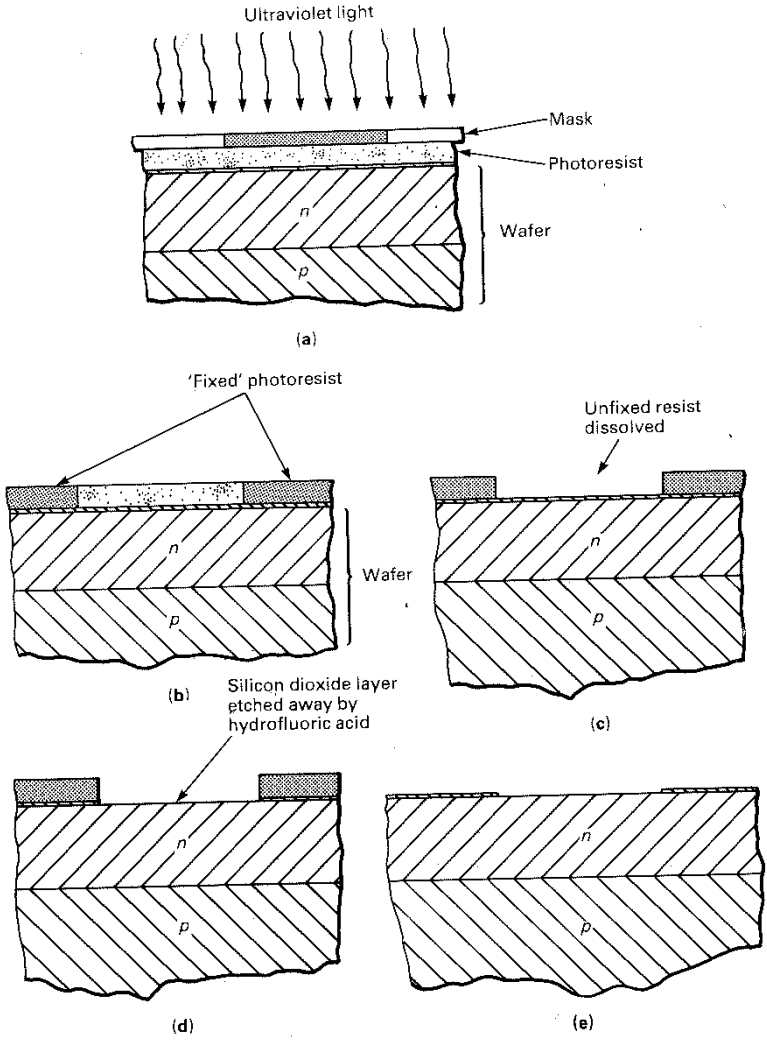
fig 11.1 *silicon wafer in cross-section, after formation of a silicon dioxide layer on its surface*



Next, openings have to be made in the silicon dioxide layer, in the right places. This is done photographically. The wafer is coated with a *photoresist*, a light-sensitive emulsion similar to the emulsion on a black-and-white film. A pattern, having the necessary cutouts, is placed over the photoresist, and the wafer is exposed to powerful ultraviolet light. The pattern, or *mask*, is removed, and the wafer washed with *trichlorethylene*, a chemical that dissolves the photoresist only in the places where it was *not* exposed to the ultraviolet light. This leaves the photoresist as a 'negative' of the mask.

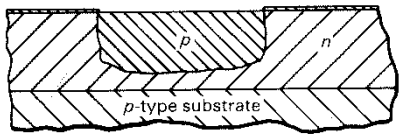
At the end of this stage, the wafer is washed in *hydrofluoric acid*, a very powerful acid that will actually dissolve glass—but not the special photoresist (nor, fortunately, the silicon wafer!). The wafer is washed again, and the resist removed with hot *sulphuric acid*. The various steps are shown in Figure 11.2.

fig 11.2 *the various stages in masking and etching the silicon dioxide layer*



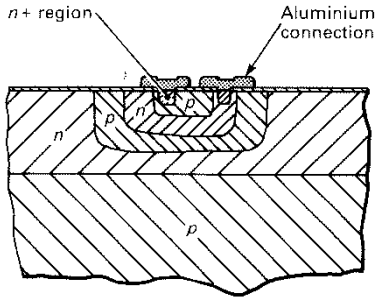
The end-result of this procedure is to produce the right pattern of holes in the silicon dioxide layer. Our wafer now goes back in the furnace, with an atmosphere of *p*-type impurity. The impurity diffuses into the wafer below the holes in the silicon dioxide, but nowhere else (see Figure 11.3).

fig 11.3 *p-type impurity is diffused into the upper n-type layer*



The whole process is repeated, with a different mask, to diffuse another *n*-region into the *p*-region just created. Then the whole process is repeated a *third* and then a *fourth* time, to diffuse more doped regions into the wafer. Figure 11.4 shows the final result.

fig 11.4 *a completed silicon planar epitaxial pn diode*



Connections have been made to the heavily doped *n*+regions, by evaporating an aluminium film on to the wafer, after suitable masking. If this seems a terribly complicated way of making a *pn* diode (compare Figure 7.7 with Figure 11.4, for example) then it is for a reason.

Those extra *p* and *n* layers below the diode itself—the *pn* junction connected to the outside world—serve to isolate the diode from the rest of the wafer. A reverse-biased *pn* junction surrounds all parts of the diode, preventing leakage to the rest of the wafer. Now this is not a very useful feature if we are going to cut the wafer up into individual diode *chips*. But suppose we cut the wafer into larger sections? The aluminium layer could carry current from one part of the wafer to another, and transistors

and diodes could all be combined on a single chip to build a complete circuit!

And this idea is the basis of microelectronics. The first integrated circuits used just this technique, though more recent devices are rather more subtle, and rely on electrical connections inside the silicon structure rather than on superimposed aluminium 'wiring'.

11.2 INTEGRATED CIRCUITS

Diodes and transistors—bipolar and FETs—can be produced on a silicon wafer. So, too, can resistors, either deposited on top of the wafer in the form of *tantalum* (a poor conductor), or built into the wafer as a 'pinch' resistor. The pinch effect is similar to that observed in a partially turned-off FET (see Figure 9.6) and relies on a very thin region for conduction. Resistance values of up to 100 kΩ can be produced in this way.

Capacitors are more problematic, and although it is possible to make capacitors on an integrated circuit, the values are generally limited to a few picofarads if the capacitor region is not to be excessively large. Capacitors are to be avoided in integrated circuits as far as possible.

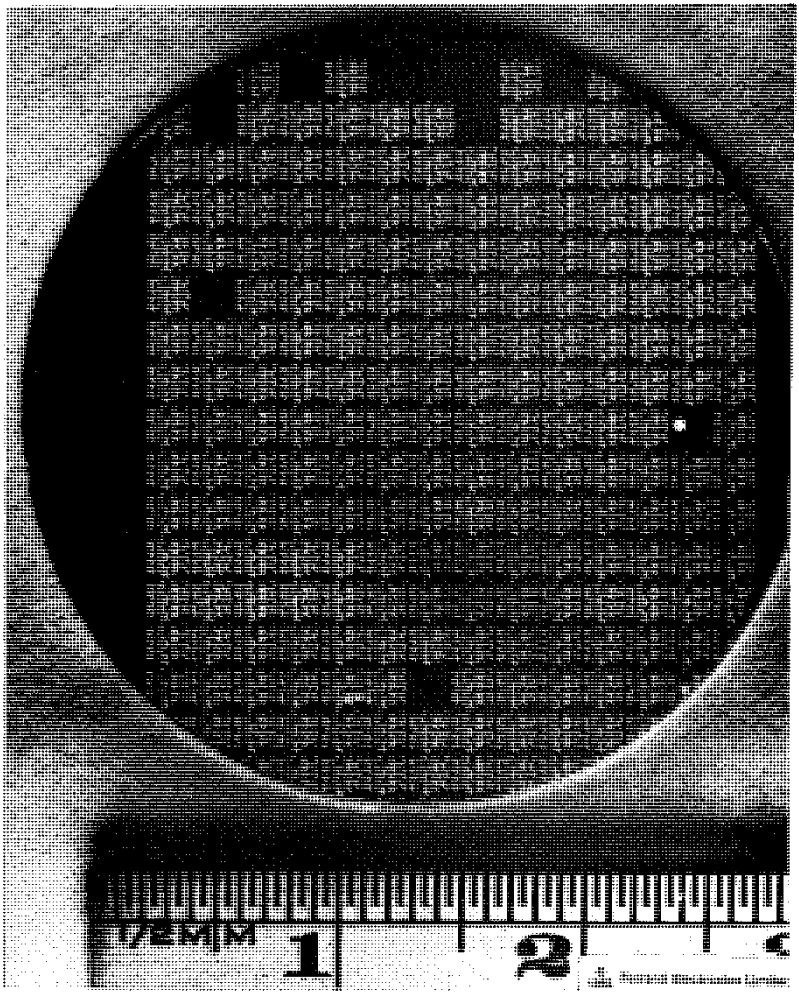
There is no equivalent of an inductor, but fortunately most circuits can be designed to avoid this requirement.

11.3 TESTING AND PACKAGING

The silicon wafer may eventually contain several hundred complete integrated circuits. They are, of course, all produced simultaneously, the masks used during processing consisting of hundreds of identical units. See Figure 11.5.

It is usual to test the individual circuits before the wafer is cut up. Testing is done automatically, and the machine marks any faulty circuits with a blob of ink. The *yield*, i.e. the percentage of 'good' circuits, varies considerably. Complex circuits might have a yield as low as only a few per cent, whereas simple circuits would be much better. The reason is clear—if a number of faults are scattered over a wafer, the larger the number of circuits on that wafer, the more fault-free circuits there will be. Once the wafer has been tested, it is cut up and the faulty circuits discarded, again automatically. The chip is mounted on a suitable frame, and connections are made between the aluminium 'pads' on the chip and the frame which forms the connections to the chip. The chip is still usually hand-wired to the frame, using ultrasonic welding to make the connections. The work is done under a binocular microscope, using special micro-manipulators. Fine gold wire is often used for the connections.

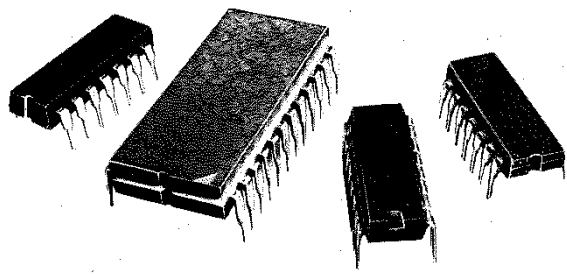
fig 11.5 *a silicon wafer—with a ruler to indicate the scale—showing a little under 200 individual integrated circuits, ready to be tested and cut up*



photograph by courtesy of Ferranti Electronics Ltd

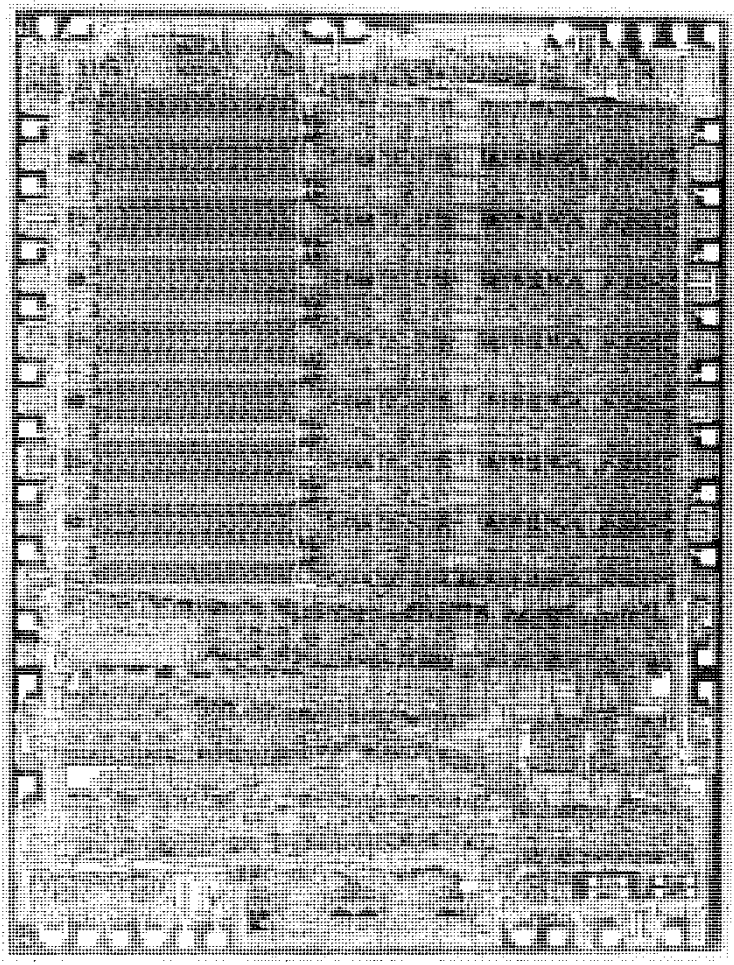
Various styles of encapsulation are used, but by far the most common is the DIL pack (Dual-In-Line). DIL packs are based on a standard $\frac{1}{10}$ inch matrix. The pins down each side are always 0.1 inch apart, and the distance between the two rows of pins is a multiple of 0.1 inch, generally 0.3 inches for packages of eighteen pins or less, and 0.6 inches for twenty pins or more. Figure 11.6 illustrates typical plastic DIL packs.

fig 11.6 *a selection of DIL-packs, the form in which integrated circuits are generally sold*



Ceramic DIL packs are used as well as the plastic ones, but only for the most expensive devices, or those requiring a very high degree of environmental protection. The number of individual semiconductor devices that can be packed on to a single chip is astonishing. Although a 14-pin DIL pack may contain a relatively simple circuit (see below) a 40-pin DIL pack can contain the electronics for a complete computer! Figure 11.7 shows a complex integrated circuit. It is in fact the central processor unit of a computer. Devices like this have a huge number of transistors and diodes, approaching 10 000 on a chip less than 10 mm square.

fig 11.7 *an example of a large-scale integrated circuit, a microprocessor; this IC contains all the important parts of a computer central processor unit*



photograph by courtesy of RCA

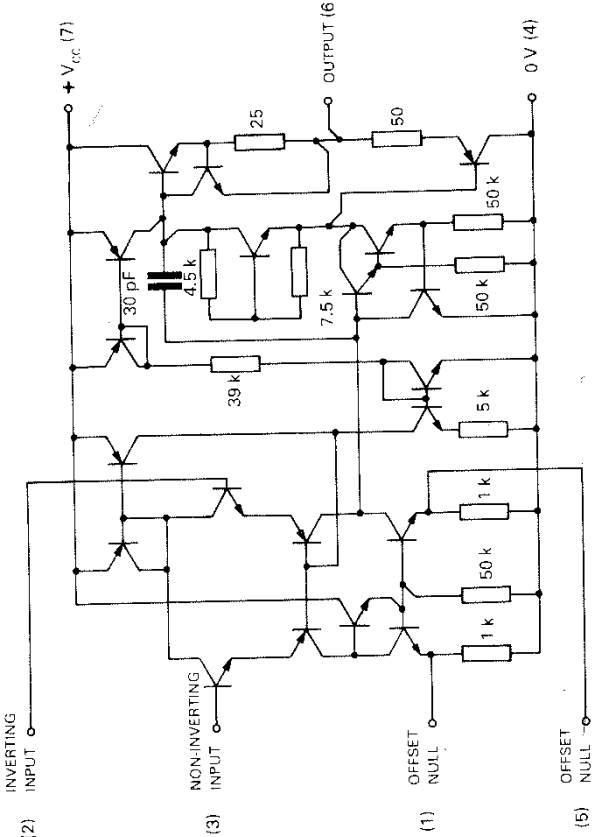


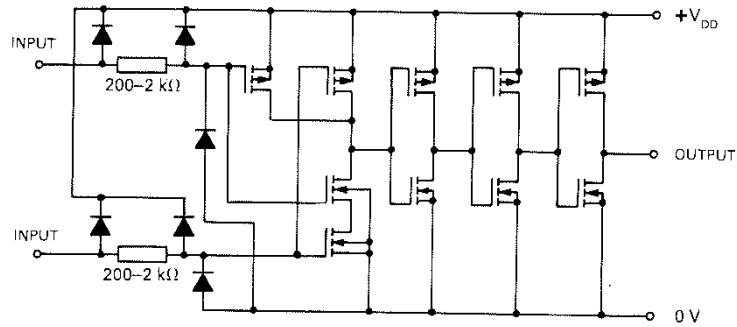
fig 11.8 *circuit diagram of a typical bipolar IC, operational amplifier type SN72741*

11.4 SOME EXAMPLES OF SIMPLE INTEGRATED CIRCUITS

In the next chaper we shall look at *operational amplifiers*, so it seems sensible to take one of the most popular operational amplifiers as an example of the kind of circuit design used (see Figure 11.8). This is the operational amplifier type SN72741. As you can see, there are many transistors, a few resistors (they take up more room on the chip than transistors or diodes), and as few capacitors as possible (one). But this is a relatively simple circuit; it is very cheap, currently costing less than a cup of coffee. Such is the scale of mass production in the microelectronics industry!

The second example–in Figure 11.9–relates to Chapter 19, and shows a 2-input AND gate. A 14-pin DIL pack contains a chip with four such gates.

fig 11.9 *a typical simple IC using MOSFET technology; components can be packed up to twenty times more densely than with bipolar technology (this diagram shows one of four identical circuits included in the CD4082 quad 2-input AND gate)*



The use of MOSFET technology means greater economy, as it can be packed on the chip up to twenty times more densely than bipolar transistors. Also, the power requirements are much lower. The CD4082 quad 2-input AND gate is as cheap as the SN72741 operational amplifier.

More complicated chips, such as the Z80 microprocessor, are more expensive, but still amazingly cheap. The Z80 currently costs about the same as a pound of coffee beans.

QUESTIONS

1. Why would a complex IC have a lower production 'yield' than a simple one?

2. What method is used to manufacture pure silicon such as might be used for semiconductor manufacture?
3. Silicon dioxide layers feature prominently in the manufacture of all silicon semiconductor devices. What are important properties of silicon dioxide?
4. What is the pin spacing of the standard DIL pack?
5. Name two 'impurity' materials that are added to pure silicon to produce (i) *n*-type, (ii) *p*-type silicon.

CHAPTER 15

RADIO, TELEVISION AND VIDEO

Radio transmission and reception was perhaps one of the earliest applications of electronics, and is—so far—the application that has made the greatest impact on society. Oddly, we can use radio, predict its properties and design circuits that work very efficiently, but we know little about the real nature of radio. Ask an electronics engineer what radio is, and the answer will be a confident, 'Electromagnetic waves.' Ask a physicist what electromagnetic waves are, and he will begin to hedge, or he will tell you that really we don't know. We do know that electromagnetic radiation is a form of energy, and that it behaves as if it is propagated as waves. The model becomes more of a model and less like reality when we discover that radio travels through a vacuum. How can there be waves in a vacuum? Perhaps in the future, theoretical physics will give us an answer. In the meantime, we use radio, describe it mathematically, and design and use electronic circuits that function happily despite our underlying ignorance.

Possibly the hardest concept to grasp in radio is the way in which circuits can be made that broadcast radio waves into their surroundings.

We begin with the *LC* tuned circuit, given in Figure 10.20. You will remember how the circuit, when oscillating, stored all the available energy, first in the capacitor and then in the magnetic field associated with the inductor, reversing the situation with each swing of oscillation. Consider the way in which energy is stored in the capacitor (Chapter 3)—it is in the form of an electric field, whatever that may be! All capacitors that we use in electronic circuits develop an electric field between two parallel and very closely spaced plates. Moving the plates apart has two effects. First, it decreases the capacitance (the principle of variable capacitors); and second it makes the electric field occupy a larger volume of space. We could make a capacitor with two plates a metre apart, which would give a tiny value of capacitance unless the plates were *very* large indeed.

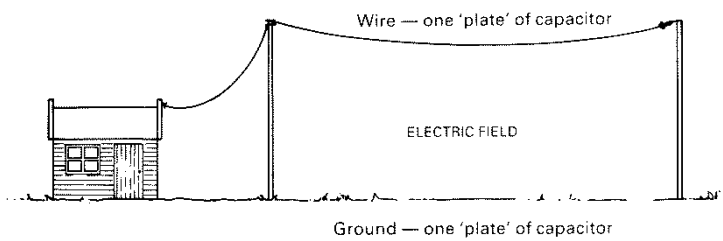
Operating this cumbersome arrangement in an *LC* oscillator circuit

would yield an unexpected result, for not all the energy put in to the circuit could be accounted for by waste heat emitted by the coil windings, etc. Some of the energy has escaped–leaked away, if you like–from the area between the plates (which are rather a long way apart). This escaped energy is, of course, electromagnetic or 'radio' energy, and has been broadcast away from the circuit. In order to make a radio transmitter, we need a circuit that will give the maximum possible 'leakage' of energy from the *LC* oscillator.

As you might expect from the hypothetical experiment outlined above, what is needed is a capacitor with an enormous gap between the plates, and designed in such a way that the largest possible amount of radio is 'leaked'.

We can, in fact, use the largest possible plate for one of the plates of the capacitor–we can use the earth! The other 'plate' has to be rather smaller, and it is convenient to use a single, long wire; this has the advantage that it can be made an excellent 'leaker' of radio energy. The earth and wire as a capacitor is shown in Figure 15.1. The picture is a familiar one, that of a radio aerial.

fig 15.1 *an aerial*



It also turns out that there is an optimum length for the aerial wire, if the 'leakage', which I shall now call 'broadcast energy', is to be maximised. Conveniently, it is when the aerial is half, or one-quarter, of the *wavelength*. The term 'wavelength' refers to the physical distance between complete cycles of the broadcast wave. Radio waves travel at the speed of light, which is about 300 000 km per second. If the radio waves were produced at the highly unlikely rate of one a second, the wavelength would be 300 000 km. Two a second, and the wavelength would be 150 000 km, still a little on the long side for a convenient aerial to be one quarter of the wavelength!

Radio is broadcast on *much* higher frequencies than this, but the formula for working out the wavelength is the same:

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

(where λ is the Greek letter 'lambda', which is conventionally used to represent wavelength).

At a radio frequency of 10 MHz, the wavelength would be

$$\lambda = \frac{300\,000\,000}{10\,000\,000} \text{ metres}$$

$$\lambda = 30 \text{ m}$$

A convenient quarter-wavelength aerial would be 7.5 m, not too bad, but unsuitable for portable equipment. For portable radio transmitters, we can exchange convenience for efficiency, and use a 1/8 wavelength or a 1/16 wavelength aerial.

The length of the aerial will affect the capacitance (refer to Figure 15.1 and you will see why this happens) and this will affect the frequency of oscillation. For a transmitter circuit to work properly, the *LC* circuit must be 'tuned' to the transmission frequency, and the aerial must be the correct length. Radio frequencies between 30 kHz and 20 GHz or so are used. Figure 15.2 illustrates the different frequency 'bands' and the names given them.

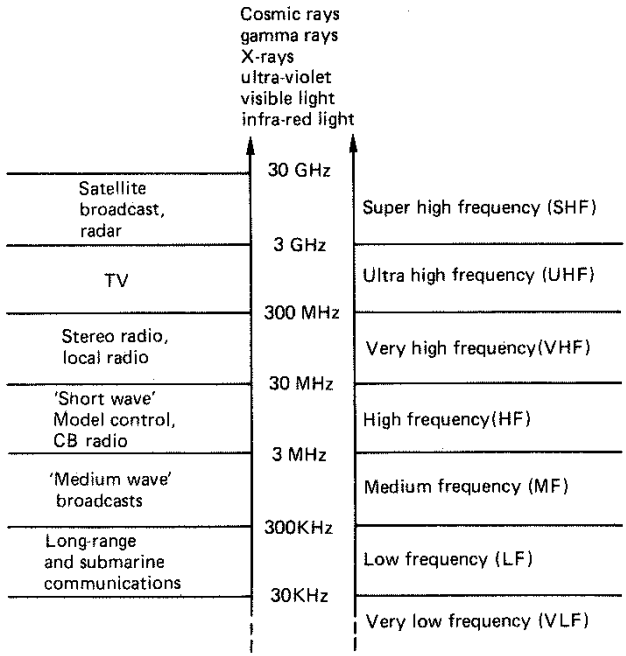
15.1 **RADIO TRANSMITTERS**

We can now look at a practical transmitter circuit, operating in the high frequency band at 27 MHz. The choice of frequency is not arbitrary, for this is the model radio control frequency, and in the UK and many other countries it is legal to operate a low power transmitter at this frequency without any form of license. Many readers can therefore build and operate this circuit—provided it is properly tuned—without breaking the law.

In order to ensure that they are operating on a legal frequency, model control transmitters have to be *crystal controlled*, i.e. the frequency of oscillation must be controlled by a quartz crystal oscillator, to ensure accuracy and stability. A frequency close to the middle of the model control band is 27.095 MHz, and crystals at this frequency are readily available from good electronic component or model shops.

The crystal-controlled oscillator circuit in Figure 10.24 is a good atarting-point. This oscillator provides the necessary radio frequency sinewave output, which is then amplified by a simple transistor amplifier (see Figure 15.3).

fig 15.2 *the radio spectrum*

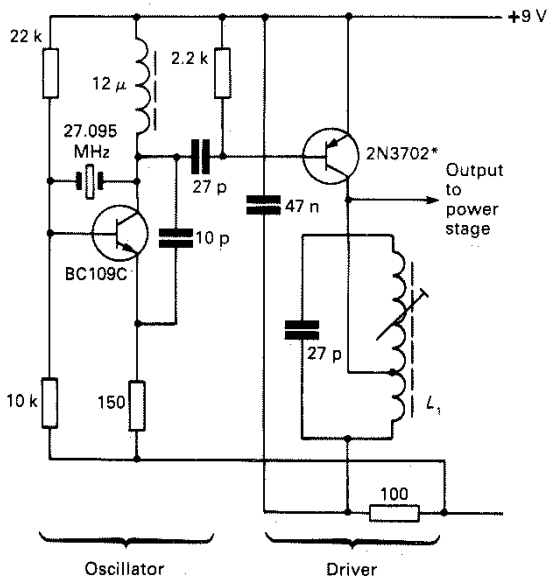


Note the use of a tuned circuit for the transistor amplifier, something which is fairly common in radio circuits. The use of a tuned collector load improves the efficiency, for the impedance of the load is highest at the 'wanted' frequency, giving the best output and optimum bias. In this case the frequency of the resonant circuit is about 27.1 MHz.

The specification of inductors is always problematic, because there are many different ways of making, say, a 10 mH inductance, and not all are suitable for all applications. In practical circuits it is usual to give constructional details of any 'purpose-made' inductors, to ensure that the right design is used for the circuit. Fortunately for us, the amounts of inductance required in high frequency radio circuits are generally quite small, and suitable inductors are easily made with a few turns of enamel-insulated wire over a small ferrite core (or no core, in which case the inductor is called 'air-cored').

To give a reasonable range with a typical model control receiver, our transmitter needs an output of 200–500 mW. A third, *power stage*, is used to give the high output current needed.

fig 15.3 *a crystal-controlled oscillator circuit, followed by a tuned amplifier stage; this practical circuit can form the basis of a transmitter*



L_1 details
6¼ plus 3¼ turns of 24 s.w.g. enamel-insulated wire, wound on a ¼ inch former with dust core (adjustable)

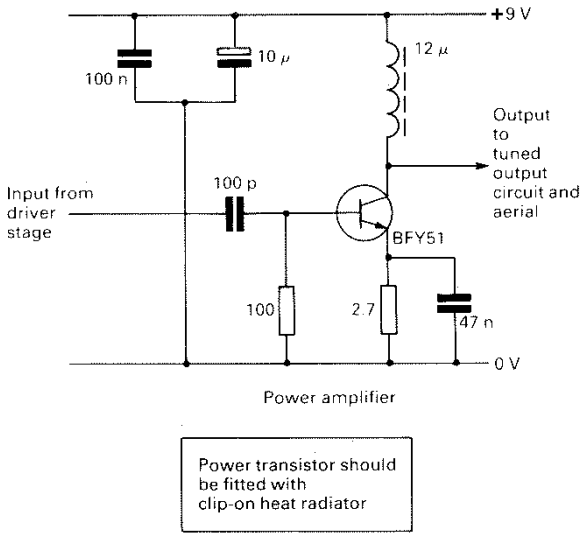
*For later projects in this book, the 2N3702 emitter is disconnected from the positive supply. It is worth making the emitter accessible!

The power stage is very similar to the preceding stage, using a very simple transistor amplifier. However, a larger, high-power transistor is needed, and the collector load (and thus the bias resistor) are much smaller in value to provide the required current. The circuit is given in Figure 15.4.

Notice the resistor values associated with the power stage, and notice also the small capacitor (100 pF) used for coupling the stage to the one preceding it: the high frequency means that a small capacitor is able to pass sufficient current.

The last part of the circuit, following the power stage, is the output tuning circuit. There are various forms, but it is usual to have two tuned

fig 15.4 *the output stage of a small transmitter; this can be driven directly by the circuit in Figure 15.3*



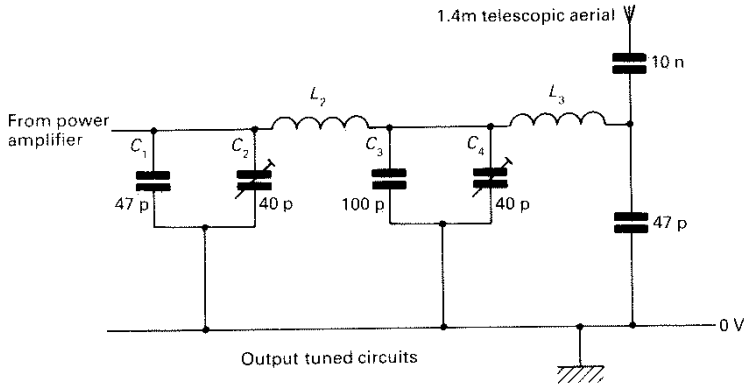
circuits—both *LC* circuits—to improve efficiency. The circuit shown in Figure 15.5 is typical.

There are a number of noteworthy points about this circuit. First, the inductors are generally air-cored (again, for efficiency) but may be wound on ferrite cores to increase the inductance if the transmitter has to be small. Second, the capacitors are variable, so that the circuit can be tuned accurately—the *LC* circuits have to be accurately 'peaked' (i.e. tuned to the exact frequency) for the transmitter to work properly. The same is true of any other tuned circuits, such as the collector load of the stage driving the power stage.

The capacitor that couples the aerial itself to the circuit is there for safety reasons, and does not play any part in the circuit's functioning. If the aerial is accidentally shorted to the transmitter casing (which is earthed to the 0 V power line) this would short-circuit the output transistor and could damage the driver transistor; but the 10 nF capacitor prevents this.

In practice, the circuit can be built on a printed board, and layout is uncritical. Matrix board is not suitable, since the strips of copper act as unwanted capacitors—at high frequencies a few picofarads can be important. The coils (in this design) are intended to be mounted facing the same way (i.e. with the same axis).

fig 15.5 *output tuned circuits for a 27 MHz transmitter; this circuit can be driven directly by the output amplifier in Figure 15.4*



L_2 L_3 details

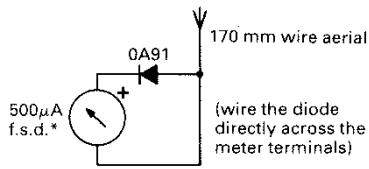
each 12 turns of 18 s.w.g. wire, air core about 8 mm diameter. Wind so that the turns are touching; the coils can conveniently be wound round a pencil.

The circuit shown in Figures 15.4 and 15.5 is entirely practical, and can be made as a construction project. The crystal *must* be suitable for operating on the model control band, between 26.995 and 27.245 MHz in the UK.

The transmitter can be built, but we need some kind of detector to check that it is working, and to allow us to tune the *LC* circuits. A suitable detector is shown in Figure 15.6. This is very simple, and will react to the 27 MHz radio frequency when held a few centimetres from the transmitter aerial.

The radio signal induces a small current in the wire, and this flows through the meter. Without the diode, an alternating current would flow, which would not perceptibly affect the meter—the sluggish needle is unable to swing to and fro 27 million times a second! The diode *rectifies* the signal, and makes the current flow through the meter unidirectional.

fig 15.6 *an RF detector suitable for use with a low-power 27 MHz transmitter*



*Full-scale deflection.

The meter can measure the current, and responds to the radio signal. Such a device is referred to as a *detector*, or *RF meter* (radio frequency meter). The diode specified is germanium, not silicon. Thermal effects and leakage are unimportant in this application, but forward voltage drop *is*–and the germanium diode is better than silicon in this respect (germanium semiconductors still have their place).

The RF meter can be used to tune the transmitter. Connect the power supply to the transmitter, and bring the RF meter close up to the fully extended aerial. Adjust the core of L_1 with a plastic screwdriver (not a metal one, which would affect the inductance). At one point, the meter will show a small deflection. Adjust the core for maximum deflection of the meter. Now adjust the output capacitors to improve the deflection. You may need to move the RF meter further away from the aerial, but keep on adjusting L_1, C_2 and C_4 until no further improvement can be obtained. This is the way to ensure that all the tuned circuits are set as accurately as possible to 27.095 MHz.

The final adjustment should be made with the transmitter in its metal case, held in the hand. The reason is that the connection to earth is made through the operator, who will be holding the transmitter and will have his feet touching the ground.

15.2 MODULATION AND DEMODULATION

Although this transmitter can send out quite a powerful signal, the signal itself carries no information. The RF meter can detect the presence or absence of the signal, but if the radio waves are to carry useful information, such as speech or control signals, the system needs additions.

There are two common methods of adding information to the radio signal, which is called the *carrier*. Each involves changing the carrier

slightly, and both systems are in common use. The first, and most obvious, is *amplitude modulation* (AM). Amplitude modulation involves nothing more complicated than changing the power, or amplitude, of the carrier in sympathy with the modulating signal. This is clearly illustrated in Figure 15.7, which shows the carrier being modulated with an audio-frequency sinewave.

fig 15.7 *amplitude modulation*

(a) Radio frequency carrier

(b) Modulating audio signal

(c) Carrier modulated with audio signal: AM

AM has the additional advantage that it is easy to recover, or *detect* at the receiver. Assuming that the signal received by the receiver is roughly the same as that shown in Figure 15.7*c*, we cannot simply feed the output into a speaker. The output is, at audio frequencies, symmetrical so that any increase in positive signal is exactly balanced by a similar increase in negative signal and the result is zero.

Fortunately, the audio signal can be extracted simply by rectifying the receiver's output, to make it asymmetrical. The carrier is then removed with a small capacitor. The circuit is shown in Figure 15.8, along with the waveforms associated with it. Comparison with Figure 15.7 shows how the modulating waveform is recovered more or less unchanged.

The second method is known as *frequency modulation* (FM). Instead of changing the amplitude of the carrier in sympathy with the modulating waveform, the frequency is shifted a little higher or a little lower. The amount of frequency shift is very small. Frequency modulation of a carrier is illustrated in Figure 15.9; compare it with the same diagram for amplitude modulation, Figure 15.7.

Detection of the FM signal is much more complicated than detection of AM signals. Various circuits have been developed, and up to the beginning of the 'integrated-circuit era' a circuit known as a *ratio detector* was most commonly used to recover the modulating signal. The circuit is given in Figure 15.10. The functioning of this circuit is quite complicated, and it is now little used.

Nowadays a circuit called *phase-locked loop* is often used. A system diagram is shown in Figure 15.11. The circuit operates as follows. The

fig 15.8 *detector circuit designed to recover an audio waveform from an AM radio signal*

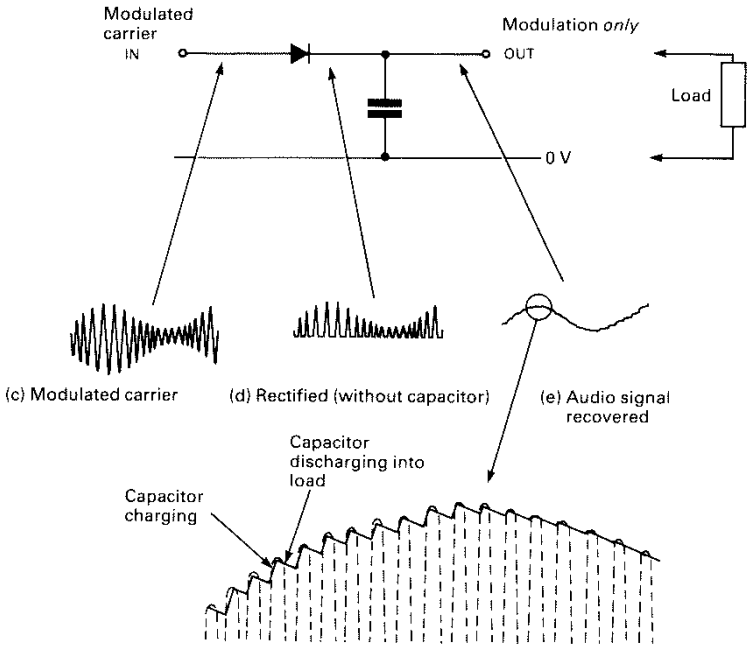


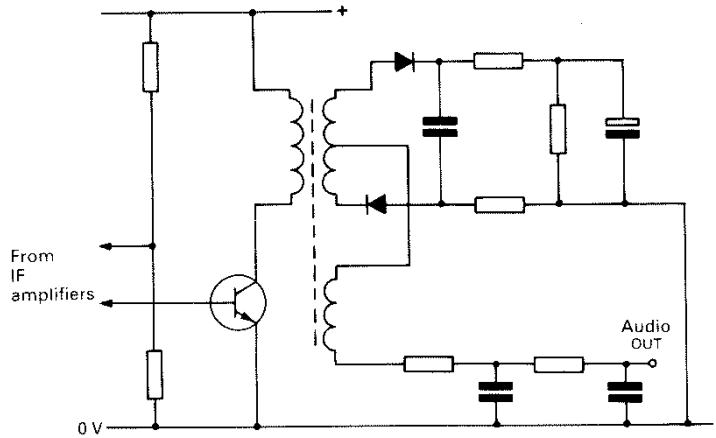
fig 15.9 *frequency modulation*

(a) Radio frequency

(b) Modulating audio signal

(c) Carrier modulated with audio signal: FM

fig 15.10 *a ratio detector circuit used for recovering the audio signal from an FM transmission*



frequency-modulated signal from the receiver is fed into one input of the *phase detector*. The other input of the phase detector is connected to the output of a *voltage-controlled oscillator* (VCO), working at the same frequency as the carrier. The phase detector compares the phases of the two signals –see Figure 15.12–and, if the two signals are out of phase, produces a positive or negative output according to the direction of phase error.

The output is fed to the VCO, which changes frequency in such a way as to move its output signal back into phase with the incoming frequency-modulated carrier. The VCO therefore 'tracks' the frequency changes of the carrier, continuously altering its own frequency to correspond with the incoming frequency modulation. The VCO 'locks' on to the carrier, through a feedback loop (hence 'phase-locked loop'). Now look at the control voltage applied to the VCO by the phase detector. A few moments' thought will show that this control voltage accurately reflects the changes in frequency of the incoming signal–it has to, in order to make the VCO change frequency to track the modulation. The control voltage therefore reproduces very accurately the modulating signal, and is simply amplified to provide an audio output.

Another, very important, advantage of the phase-locked loop is that it not only tracks the modulation but will, within a certain range, track overall changes in frequency of the carrier, or will compensate automatically for any 'drift' in its own operating parameters. This feature enables

fig 15.11 *system diagram for the phase-locked loop system of detecting a frequency-modulated signal*

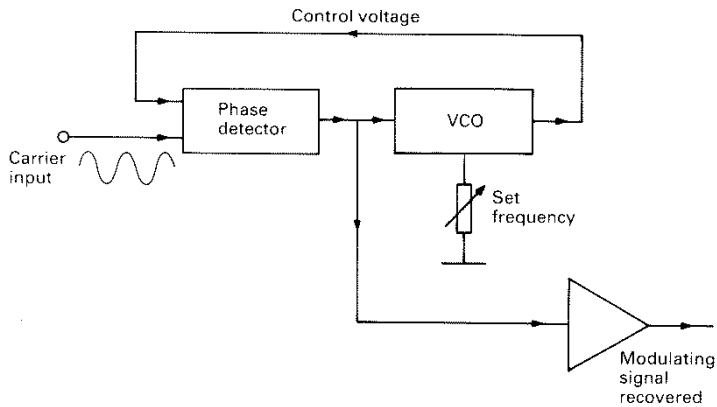
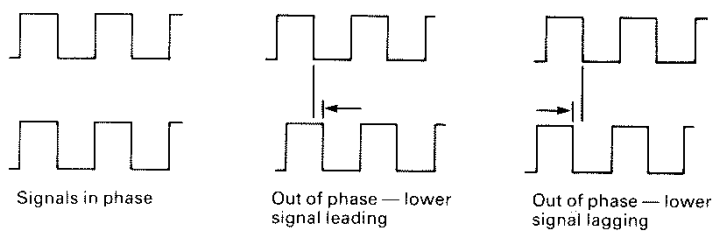


fig 15.12 *signals in and out of phase*



the detector to lock on to a signal when the receiver is tuned sufficiently nearly to the transmitter frequency. You will notice this feature working in almost all modern FM radio receivers.

Phase-locked loop integrated circuits specifically designed for radio receivers will usually incorporate other 'convenience' features. These include an output to operate an indicator (usually a LED) when the circuit is locked on to a signal, a means of turning off the feedback to unlock the circuit, and a means of changing the 'capture range' of the circuit (i.e. the amount of deviation from the selected frequency before the circuit unlocks).

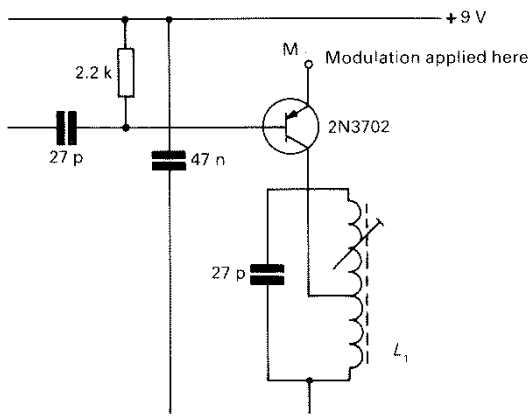
It is obvious that the phase-locked loop is actually a rather complicated circuit. Its internal design is certainly complex, but this is unimportant in the context of ICs. The IC is treated as a single component by manufac-

turers of radios, and as a 14-pin DIL pack (typical packaging for this component) it simplifies and cheapens assembly of an FM radio receiver.

To return to the practical transmitter circuit of Figures 15.3–15.5, the output can be amplitude-modulated by varying the gain of the driver stage with an applied signal. A simple way of doing this is to remove the emitter connection of the driver transistor from the positive supply rail and connect it to the output of an amplifier or transistor switch that can swing between 0 V and the positive supply voltage. The output of the transmitter will vary in more or less linear fashion with the voltage. There are, of course, many different circuits for modulating a transmitter, all achieving the same result.

If the demonstration transmitter has been constructed, a modulator designed for model control can be added. Details are given in Chapter 22, since the pulse-width modulation system used involves digital systems. In the UK at least, it is illegal to add an audio modulator to this transmitter. Figure 15.13 illustrates the connection for the modulator—refer also to Figure 15.3.

fig **15.13** *modification to the transmitter circuit to enable modulation to be applied*

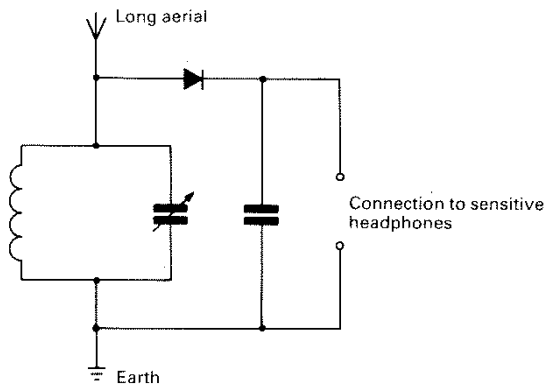


Frequency-modulation techniques are beyond the scope of this book, but a simple FM modulator uses a varicap diode, in parallel with the oscillator crystal, to obtain small changes in transmission frequency according to an applied modulating voltage.

15.3 RADIO RECEIVERS

The simplest radio receiver consists of just a tuned circuit, a diode, a capacitor, a pair of headphones, and as much aerial wire as possible! This is the 'crystal set' of the 1920s, and is shown in Figure 15.14.

fig 15.14 *a 'crystal set'*



The electric field from the transmitter induces a tiny current in the aerial wire—which needs to be a few tens of metres long—and the *LC* resonant circuit selects the required frequency. It does this because the *LC* circuit has low impedance at frequencies other than the resonant frequency, and this 'shorts out' unwanted radio transmissions. At the resonant frequency the *LC* circuit has a high impedance, so the transmission at the selected frequency appears across the circuit as a small radio frequency alternating voltage.

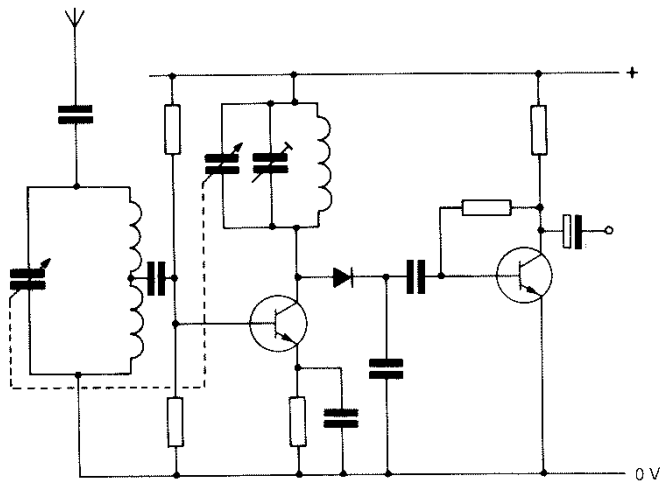
This is rectified by the diode to recover the audio signal, using a diode with a low forward voltage drop (germanium). The capacitor removes the RF component—Figure 15.8—and a pair of sensitive high-impedance headphones produce a (just) audible output.

The output can, of course, be amplified. However, results from this circuit are still unsatisfactory, chiefly because the tuning cannot separate one radio station from the next. Often, several stations will be received together.

There is a need to make the receiver more *selective*. As we saw earlier, when discussing transmitters, the use of a tuned *LC* circuit as a transistor amplifier's collector load—instead of a resistor—makes the circuit selective,

amplifying the wanted frequency much more than other frequencies. If, therefore, a tuned transistor amplifier of this type follows the tuned aerial circuit, selectivity is enormously improved. The *sensitivity* of the receiver is also better, as the radio frequency signal is amplified before it reaches the detector diode. A receiver incorporating a single tuned RF amplifier is shown in Figure 15.15.

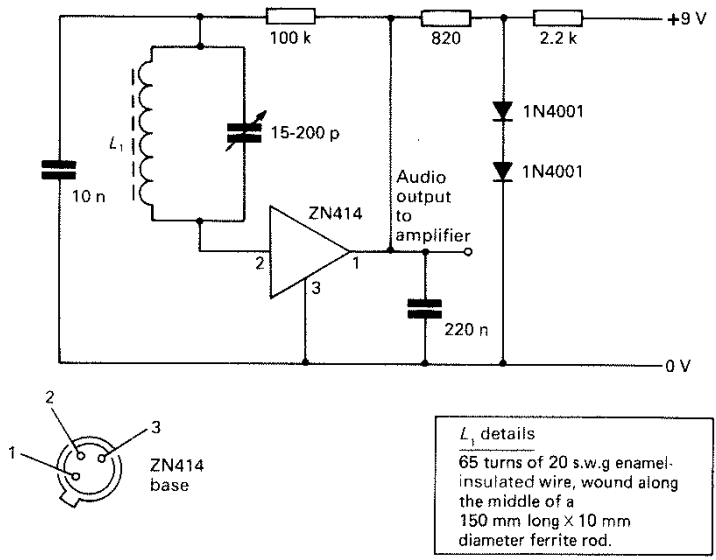
fig 15.15 *TRF receiver: only moderately efficient*



Receivers of this type can be made to work quite efficiently, and are very simple to construct. There is, in fact, an integrated circuit version of this *tuned radio frequency* (TRF) receiver, the ZN414. The ZN414 makes an ideal construction project as it is by far the simplest radio receiver to build. Unfortunately, it will not operate at frequencies higher than 3 MHz, so cannot be used with the model control transmitter. It can, however, be used very effectively with the amplifiers illustrated in Figure 13.9 and 13.3 and will connect directly to them. The circuit for the ZN414 radio is shown in Figure 15.16. The two forward-biased diodes appear to short-circuit the power supply to the ZN414, but this IC requires only a 1.5 V supply, and the combined forward voltage drop of the diodes (0.7 V + 0.7 V = 1.4 V) produces a reference voltage that is near enough right. The circuit takes only 300–500 μA from the supply!

There is no external aerial. With the sensitivity of the receiver vastly increased compared with the simple crystal set, the inductor itself can pick

fig 15.16 *circuit of a modern IC TRF receiver; this provides a low-level output suitable for feeding to an audio amplifier—this is a practical circuit and will work well with the amplifier designs of Figures 13.9 and 13.13*



up a sufficiently large signal to work the receiver. A large ferrite core a few centimetres long is used for the aerial, with the coils of the inductor of the tuned aerial circuit wound on it. The whole assembly is called a *ferrite rod aerial.* The aerial is quite directional and can be used to tune out unwanted signals.

Although capable of a respectable performance, the TRF has definite limitations and the ZN414 may represent a development approaching the ultimate for this type of design.

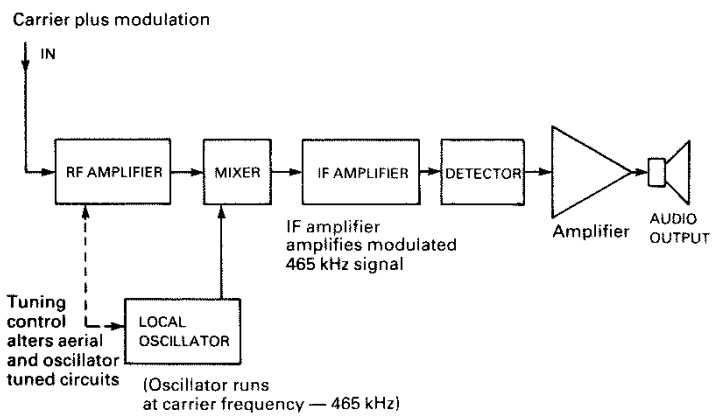
A much better bet from the point of view of selectivity in particular is a receiver consisting of a number of *tuned* amplifier stages. It turns out that this is difficult to build, and even if built gives a poor audio quality. If an audio frequency signal is used to modulate a radio frequency carrier, the result will not be a signal at one frequency alone, but a signal that has components at a whole range of frequencies. The lowest will be the carrier frequency *minus* the highest audio frequency and the highest will be the carrier frequency *plus* the highest audio frequency. A medium wave radio station transmitting at 500 kHz, with an audio signal between 40 and

9000 Hz, would therefore actually be broadcasting on frequencies between 500 – 9 = 491 kHz and 500 + 9 = 509 kHz. And the radio receiver would need to respond to all frequencies between these two figures equally, and ideally not respond at all to frequencies above 509 kHz or below 491 kHz. This is not possible to attain using a series of tuned circuits, for the combined effect is too selective at the chosen frequency.

Much more serious deficiencies come to light when we try to tune the receiver. With several stages, each with its own tuned circuit, we have to contrive a means of tuning *every LC* circuit simultaneously when changing frequency (station). Such a design would be very cumbersome, though a few were made in the 1930s.

There is a solution, one that is employed by all commercially made radio receivers. The design is called the *superheterodyne* (superhet) receiver, and a system diagram is given in Figure 15.17.

fig 15.17 *system diagram for a superhet radio receiver*



The incoming signal from the radio frequency amplifier (the RF amplifier may be missing in very simple receivers) is *mixed* with a signal from an oscillator, known as the *local oscillator*. The oscillator is tuned to a slightly lower frequency than the selected radio frequency. In just the same way that the transmitter produced sum and difference frequencies above and below the carrier frequency, so the mixer produces an output that is the difference between the frequencies of the received radio signal and the output of the local oscillator. The oscillator is designed, in most AM receivers, to run precisely 465 kHz below that of the received carrier. A ganged (double) variable capacitor is used to alter the tuning of the aerial

circuit and the oscillator simultaneously, always keeping them 465 kHz apart. The output of the mixer is therefore always a 'carrier' at 465 kHz, and this carrier is still modulated with the original audio signal.

Now it is possible to use two or three tuned amplifier stages if necessary. The lower frequency makes design much easier, and the tuning of the stages can be arranged to give a selectivity response close to the ideal. Once the signal has been amplified sufficiently, a detector diode (or, for FM, a suitable discriminator circuit) can follow, with audio amplication last in the line.

The frequency resulting from the carrier and local oscillator inputs to the mixer is called the *intermediate frequency* (IF), and the stages that follow are called *intermediate frequency amplifiers*. They invariably have tuned collector loads: tuned, of course, to the intermediate frequency. The intermediate frequency need not be 465 kHz. In model control receivers it is usually 455 kHz, and in VHF radio receivers it is 10.7 MHz. The range of intermediate frequencies used is limited, not for any technical reasons but for the purely practical reason that manufacturers produce IF tuned transformers, known as *intermediate frequency transformers* (IFTs) that are preset to the intermediate frequency, and it is convenient to make just a few types. IF amplifier stages invariably use transformer coupling, for the tuned collector load simply needs a secondary winding to make it into a coupling transformer as well, which saves components and makes for an efficient circuit. A transformer-coupled amplifier stage was shown in Figure 10.11.

Receivers built with discrete components generally combined the local oscillator and mixer into a single transistor stage by means of a circuit like the one illustrated in Figure 15.18. The inductor of the tuned circuit L_1 $C_1 + C_2$ is wound on the ferrite rod aerial, along with a coupling winding L_2 which works as a transformer and provides a signal current for the transistor. The transistor, as well as working as an RF amplifier, acts as an oscillator, L_3 and L_4 providing feedback from the collector to the emitter to sustain oscillation. The frequency of the oscillator is determined by the tuned circuit L_5 $C_3 + C_4 + C_5$ and the tuned IFT selects the intermediate frequency to pass on to the next stage (the IFT has a low impedance at radio frequencies). C_1, which tunes the aerial circuit, and C_3, which tunes the oscillator circuit, are mechanically connected (ganged) so that the two frequencies change together. C_2 and C_4 are preset variable capacitors used to balance the frequencies when the receiver is being adjusted, or *aligned*.

Figure 15.19 shows a typical IF amplifier stage that might follow the mixer/oscillator of Figure 15.18.

Note that in both these figures the IF transformer is shown surrounded by a dashed box; this indicates that it is constructed as a single component—IFTs are mass produced and are very cheap. The IFT is usually mounted

fig 15.18 *mixer/oscillator stage, with a tuned collector load*

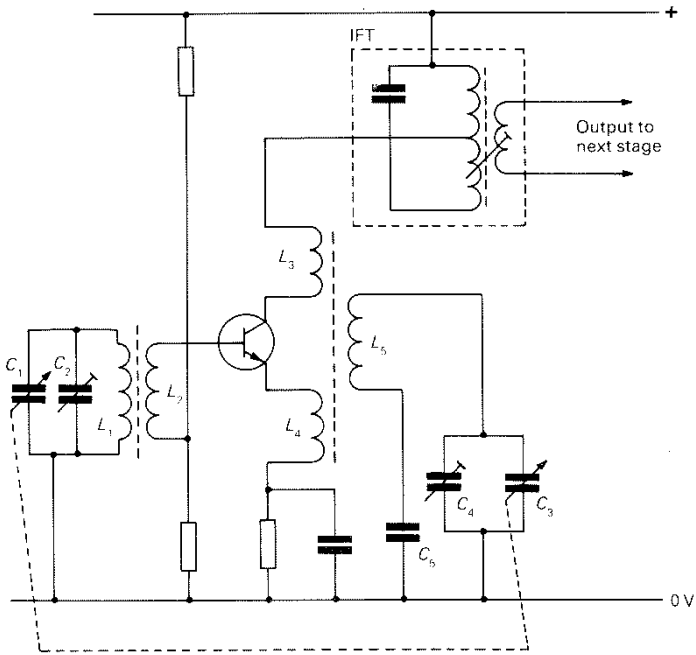
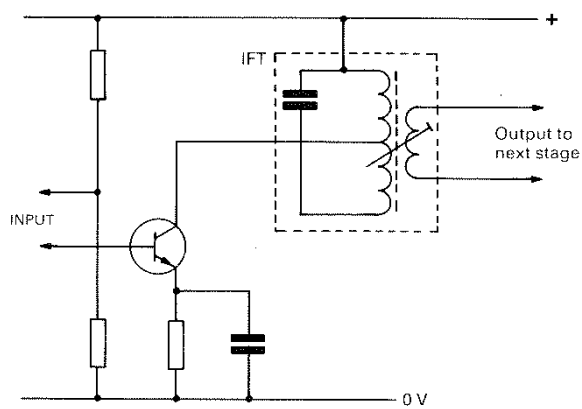


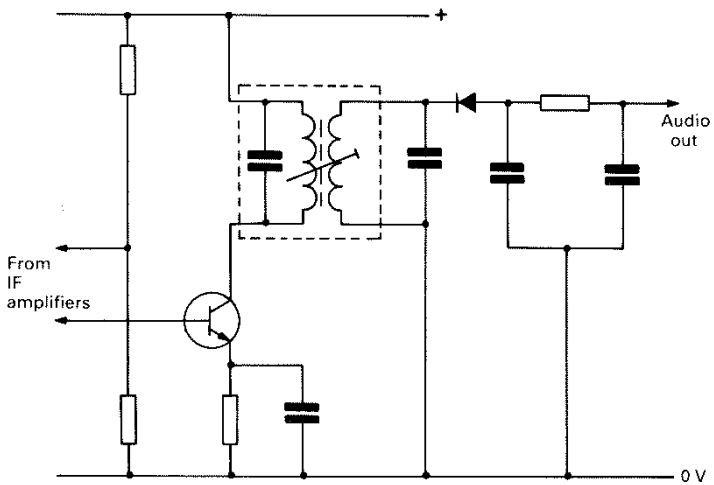
fig 15.19 *a typical IF amplifier stage*



inside an aluminium can which acts as mechanical protection and also as screening against stray electric fields. In practical circuits the aluminium can is connected to supply 0 V, to 'earth' stray signals. The IFT can has a small hole in the top through which the ferrite core can be adjusted to tune the *LC* circuit over a small range—this is used when aligning the receiver. The IFTs are clearly visible in the photograph of the model control receiver (see Figure 15.23).

Following two or even three IF amplifier stages is the detector. In AM receivers this is really just a diode, but usually with a slightly more sophisticated smoothing circuit than the single capacitor of the crystal set. A typical detector stage is shown in Figure 15.20. And last, there is an audio amplifier.

fig 15.20 *an IF stage followed by a detector*



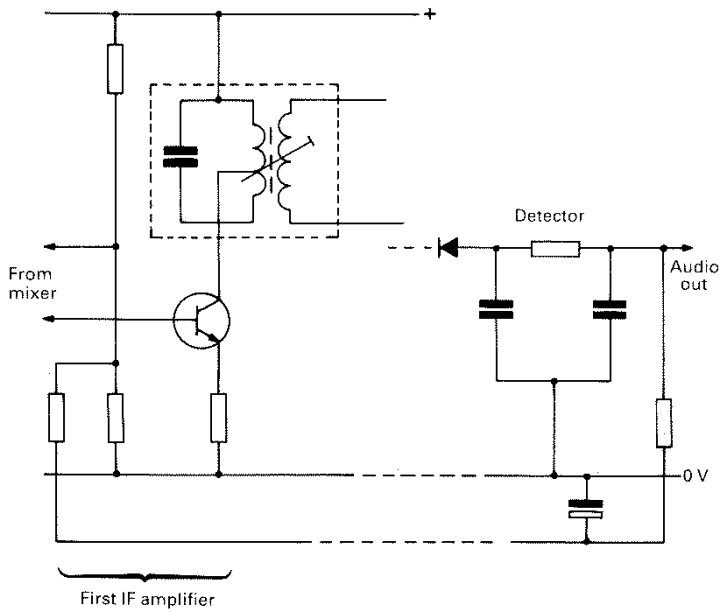
There is one more feature that appears in all but the most rudimentary radio receivers, and that is the provision of *automatic gain control* (AGC). If a sensitive receiver is tuned to a powerful nearby transmitter, the signal may be so strong that it overloads the later IF amplifier stages, causing severe distortion of the sound. To prevent this, AGC is added. The idea—and the circuit—is quite simple, and involves using a proportion of the rectified signal from the diode to control the biasing of the first IF amplifier transistor. A large signal at the detector diode lowers the bias potential at the base of the first IF transistor and reduces the gain—which, of course,

reduces the output signal. With the right circuit values the receiver balances out with the correct output and bias levels.

To prevent the AGC responding to transient changes in the audio content of the signal (and increasing the gain during the singer's pauses for breath!), a large-value capacitor is placed across the AGC feedback line to delay the operation of the AGC by acting as a reservoir—the AGC then responds to the average signal level, averaged over a period as long as a couple of seconds.

An implementation of an AGC circuit is shown in Figure 15.21 (refer also to Figures 15.19 and 15.20).

fig 15.21 *application of feedback to provide automatic gain control*



15.4 A PRACTICAL RADIO-CONTROL RECEIVER

It is quite possible to construct a working radio control receiver which, when used with the transmitter described at the beginning of this chapter, will have a range of up to half a mile. Commercially-designed radio control receivers generally use specially designed ICs that minimise the number of components required, but this one uses circuits based on discrete tran-

sistors. It means that there are a few more parts, but it does have the big advantage that you can see how the circuit works, and find your way through the functioning of every part of the system. The complete circuit of the receiver, from aerial to output, is illustrated in Figure 15.22.

This receiver is a little simpler than a broadcast receiver (and much easier to adjust) because it operates on a fixed frequency. The frequency is set by a crystal-controlled oscillator almost identical to the one used in the transmitter (and to the one in Figure 10.24), apart from a simplified bias circuit using just one resistor. In a broadcast receiver the local oscillator and the tuning circuit are adjusted in frequency together to preserve the correct intermediate frequency, which is then selectively amplified by the IF amplifiers. In the radio control receiver the requirement is for the tuning to be set to a single, very accurately determined frequency, which is why the more expensive crystal-controlled oscillator is used in preference to an *LC* oscillator.

The crystal is at a lower frequency than the crystal in the transmitter, differing from it by the selected value of the intermediate frequency, in this case 455 kHz. The transmitter crystal was 27.095 MHz, so the required crystal for the local oscillator in the receiver would be:

$$27.095 - 0.455 = 26.640 \text{ MHz}$$

This produces the correct intermediate frequency when mixed with the incoming RF carrier. A signal received from a transmitter working on a slightly different frequency would produce an intermediate frequency higher or lower than 455 kHz, which would not be amplified by the IF amplifiers and would therefore not produce an output.

The incoming radio signal is amplified by a JUGFET. This device is used in preference to a bipolar transistor because of its very high gate resistance, which maximises the efficiency of the aerial tuning circuit (L_2) that helps to reject powerful signals at unwanted frequencies. The output of the oscillator is transformer coupled to the source of the JUGFET (the signal does not need amplification), so that mixing of the radio and oscillator frequencies takes place. Following the mixer there are two IF amplifier stages. Bias for the transistors is obtained entirely from the AGC line - again, slightly simpler than the 'textbook' circuit in Figure 15.21.

The detector stage incorporates a diode between the base and emitter, to rectify the input signal for that stage. The AGC voltage is derived from this stage. Finally, there is a d.c. amplifier fitted with low-pass filters to reject unwanted noise from the earlier stages.

This circuit will work perfectly with the transmitter - but as it stands the transmitter will be sending only the carrier, and there won't be any output! To tune the receiver, a ***modulated*** 27.095 MHz transmission is needed. A borrowed AM model control transmitter could be used, fitted

fig 15.22 *the circuit of a complete superhet receiver; this is a practical design for use with the 27 MHz transmitter described at the beginning of this chapter. The receiver is crystal-controlled and is intended for model radio control work. A suitable decoder circuit is given in Figure 22.23. See also Appendix 1*

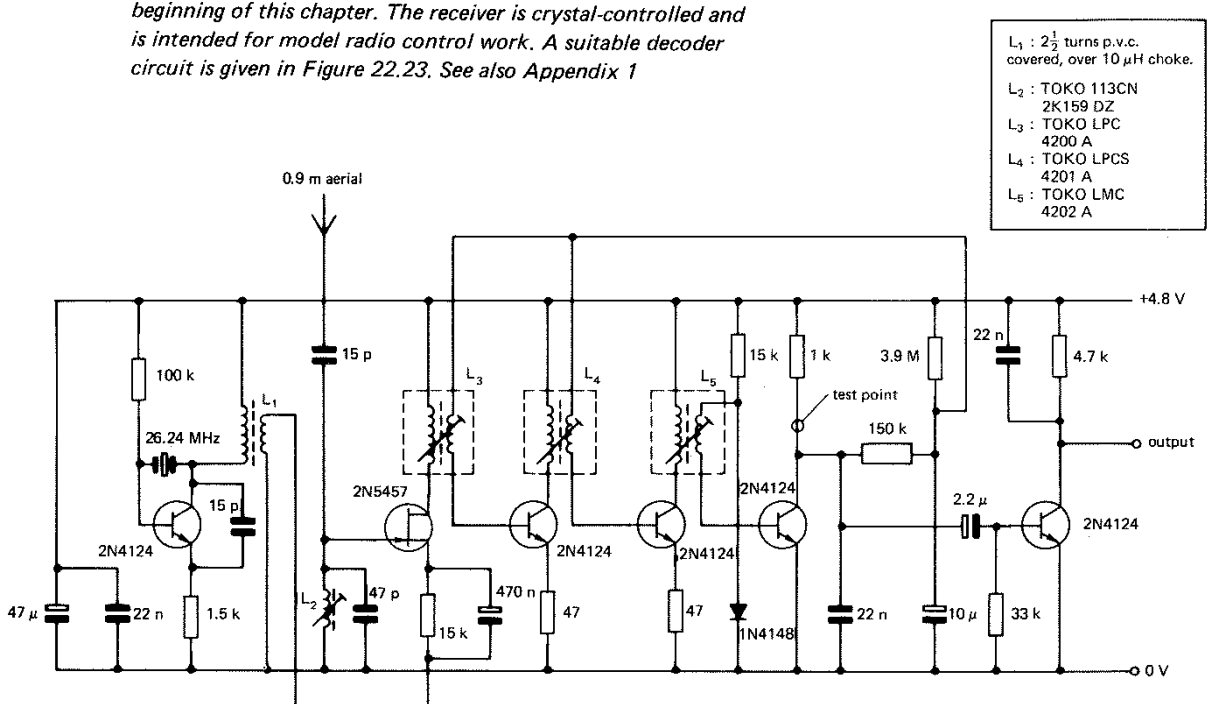
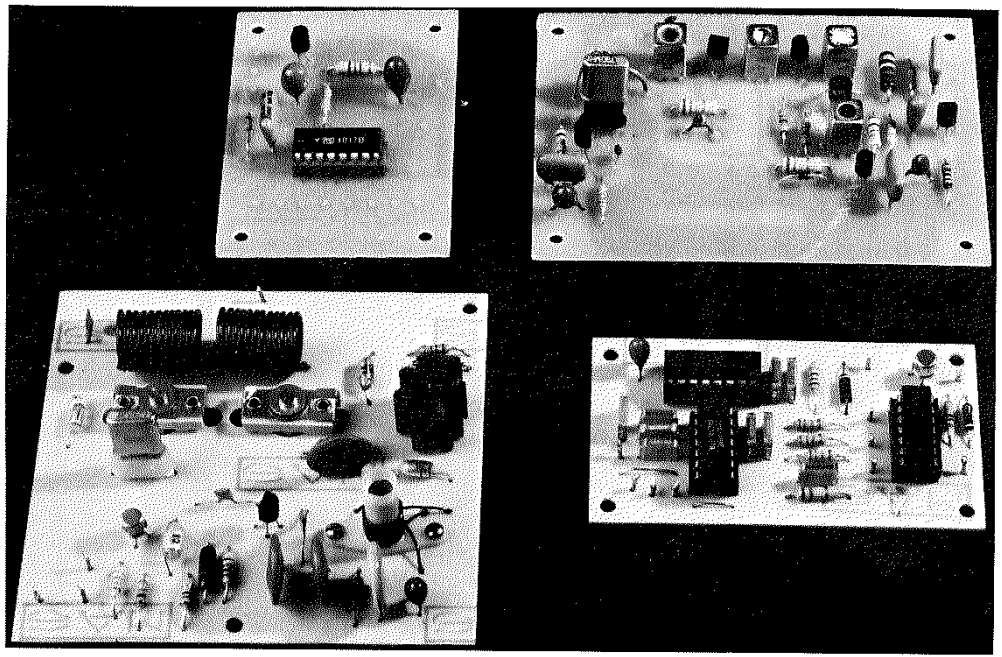


fig 15.23 *the complete radio control circuit boards - transmitter with its encoder, and receiver with decoder*



with a 27.095 MHz crystal, but the transmitter described in this chapter can be used with a modulation system described in Chapter 22.

To tune the receiver, the 0.9 m aerial must be connected and arranged roughly straight. Connect a voltmeter to the collector of the detector transistor (marked 'test point' in Figure 15.22) via a 4.7 kΩ resistor. Select a range reading around 5 V full scale. When the battery - ideally a 4.8 V nickel-cadmium accumulator - is connected, the meter should show little, in any, deflection. Switch on the transmitter with the aerial collapsed. When the transmitter is moved close to the receiver, the meter will begin to show a deflection. Using a plastic trim-tool, adjust L_2 for the maximum deflection, if necessary moving the transmitter away to keep the meter reading about 2.5 V.

Now repeatedly adjust L_3, L_4, and L_5 in turn to obtain the highest possible meter reading, if necessary moving the transmitter even further away; you may end up with the transmitter several metres from the receiver. When no further improvement can be obtained, the receiver is correctly aligned.

15.5 CONSTRUCTION PROJECT

Sufficient information has been given here (and will be in Chapter 22) to enable you to build the model control transmitter and receiver as a construction project. Although circuit layout is not particularly important, the printed circuit board designs given in Appendix 1 are recommended. The designs are fully compatible with modern radio control practice, and can be used with commercially available servos (see also Figure 15.23).

15.6 TELEVISION RECEIVERS

Having covered the principles of radio transmission and reception, it is now possible to consider the principles and practice of television. Even monochrome (black-and-white) televisions are extremely complicated in the details of their circuits, and television is a complete subject in its own right. In *Mastering Electronics* the systems involved will be explained, but without circuit details, except where this is essential to an understanding of the systems. To include a detailed study of television (and of various other topics, also) would be to make this book at least twice its present size and cost!

15.7 MONOCHROME TELEVISION RECEIVERS

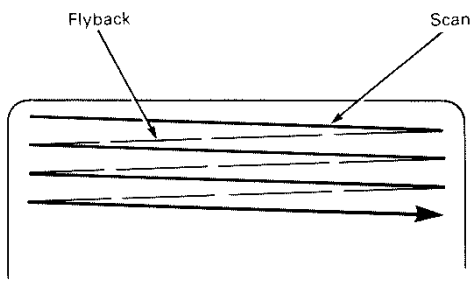
Most people now know that motion picture films produce the illusion of movement by presenting a rapid sequence of still pictures, each one slightly

different from the one before it. The eye 'joins up' the pictures and interprets them as a single image in smooth motion, a phenomenon called 'persistence of vision'. Home movies (silent) show 18 frames of film (i.e. 18 pictures) every second. Sound movies shown 24 frames per second.

The movement on the television screen is created in exactly the same way, and the television transmission is sent out as a series of 'still' pictures, at the rate of 25 per second in the UK and many European countries, and at the rate of 30 per second in the USA and elsewhere. (The rate is actually half the frequency of the mains supply in the country in question. In the UK the mains supply is at 50 Hz, and in the USA it is 60 Hz.)

Central to the television system is the TV tube, illustrated in section in Figure 5.8. The bright spot is made to scan the front of the tube continuously, in lines running across the screen, moving down from top to bottom in a pattern called the *raster* (see Figure 15.24).

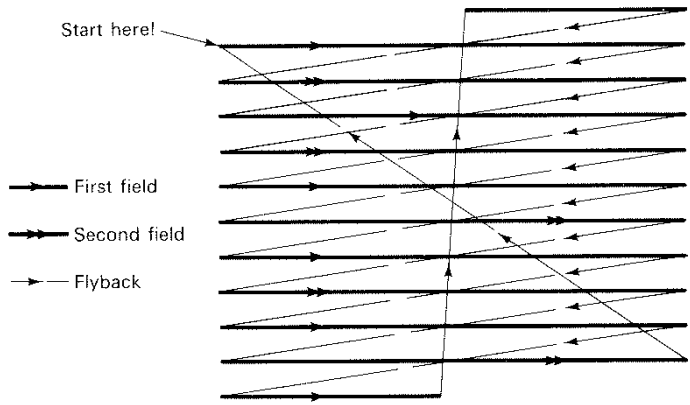
fig 15.24 *television screen showing the raster*



The brightness of the spot is reduced to nothing during the right-to-left 'return' stroke, the *flyback*. In the UK the screen has 625 horizontal scan lines in the raster; it is slightly different in some other countries. Unfortunately, a scan of this sort would produce flicker, even at 25 frames per second, whereas a motion picture film running at the same speed would not. The reason lies in the way the television picture is drawn from the top downwards, whereas the motion picture projector uncovers the whole of each frame virtually at once. To reduce the flicker, which has an odd appearance of running from the top down the screen, *interlaced scanning* is used, and the picture is produced in two halves, each taking 1/100 th of a second (UK). The electron beam scans the screen as above, but only with $312\frac{1}{2}$ lines. When it gets to the middle of the bottom of the screen (the $312\frac{1}{2}$th line!) it returns to the middle of the top and scans the picture again, but *between the first set of lines*, 'interlacing' the second scan with

the first. Using this technique makes the amount of flicker hardly noticeable. Interlaced scanning is illustrated in Figure 15.25, though with an 11-line screen instead of 625, to make the diagram clearer.

fig 15.25 *interlaced scanning*



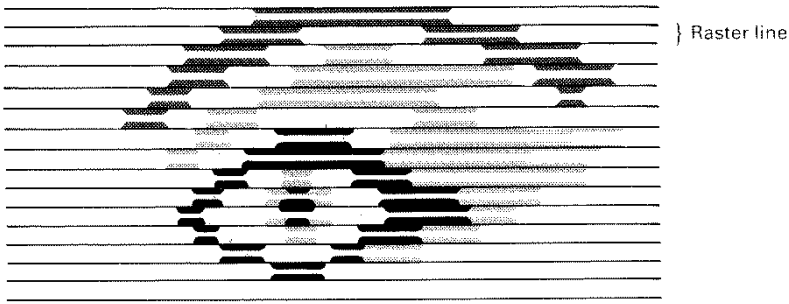
It is important to realise that in a monochrome television the phosphor on the screen is completely homogeneous; the lines are merely a product of the scanning that produces the raster.

Two time periods are critical to the television: the time the beam takes to scan the width of the picture—the line frequency; and the time taken for one frame—the frame frequency. Two oscillators within the receiver generate these frequencies, and they are called, respectively, the *line timebase* and the *field timebase*. The line timebase has to scan the screen $25 \times 625 = 15\,625$ times every second, so has an operating frequency of 15.625 kHz, and the field timebase scans the picture 50 times every second (twice for each frame; remember the interlaced scanning), so operates at 50 Hz (the mains frequency—you will see why this is a good idea later).

The timebases have not got to operate at about the right frequency, they have to work at *exactly* the right frequency, to scan the screen in exact synchronism with the cameras at the studio. For as the electron beam sweeps across the screen it is modulated so that the brightness of the screen is changed continuously to produce a picture. Figure 15.26 shows how a picture is built up.

Clearly the timebases have *got* to be exactly in synchronism with the timebases of the cameras at the broadcast studio, or the TV picture will be chaotic, with no recognisable images at all. The transmitted signal

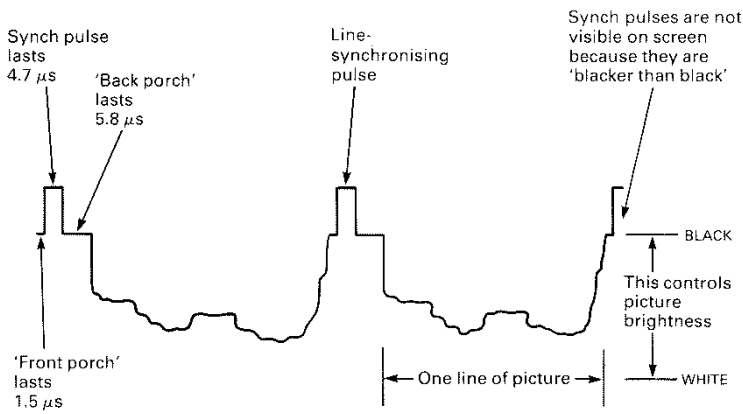
fig 15.26 *the brightness of the spot forming the raster is modulated to provide pictures on the television screen*



therefore includes information about the timing of the line and field scanning, and circuits in the receiver extract this information and use it to *synchronise* the two timebases.

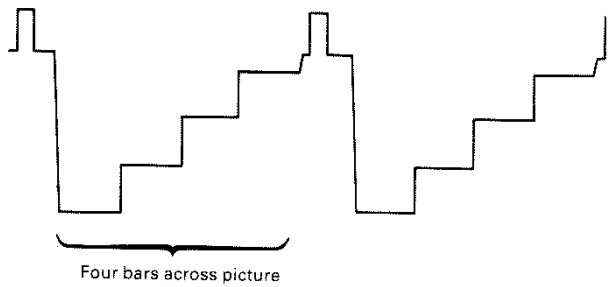
Every single sweep of the line timebase is *triggered* by a pulse in the signal received by the television. An RF carrier from the transmitter is modulated with a waveform corresponding to the brightness of the line of the TV picture. Figure 15.27 shows part of this waveform, corresponding to two lines of the picture.

fig 15.27 *the waveform of a television picture broadcast; this is the waveform that is recovered from the AM signal received by the television*



Compare the waveform in Figure 15.27 with the one in Figure 15.28, which shows two lines of a picture consisting of four vertical stripes, progressively darker from left to right.

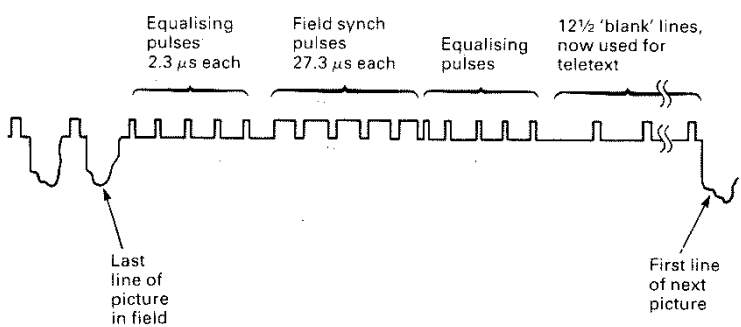
fig 15.28 *two lines of a television picture showing four vertical bars of increasing density*



The *line-synchronising pulses* each trigger the line oscillator into a single sweep across the picture, or more commonly are used to synchronise an oscillator to run at a precise speed. Although in theory it would work to have an oscillator that is triggered for each line, in practice this would mean that any tiny interruption in the received signal would result in the loss of a synch pulse, and the loss of overall synchronisation for the rest of that field.

At the end of each field there is a special series of line-synchronising pulses that also trigger the *field synchronisation*. The actual sequence is quite complicated and is shown in Figure 15.29.

fig 15.29 *the synchronising pulse sequence at the end of each field*

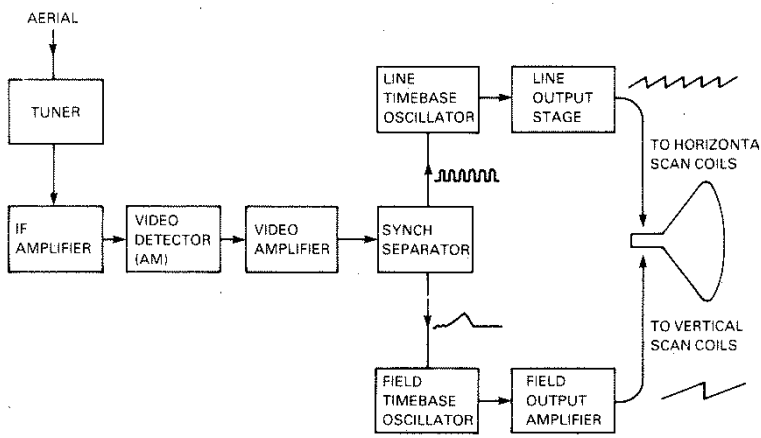


Following the last line of the picture in any particular field there are five *equalising pulses*, followed by five *field-synchronising pulses*, followed by another five equalising pulses. The exact function of the equalising pulses is subtle: suffice to say here that it makes the circuits in the receiver simpler. The five field-synchronising pulses are detected by the receiver circuits—they are five times longer than 'ordinary' synch pulses—and trigger the next sweep of the field timebase. Once again, a synchronised free-running timebase is used to prevent the picture collapsing entirely if the signal is lost momentarily.

Following the pulse sequence for field synchronising there are a further $12\frac{1}{2}$ 'blank' lines that are not used for the picture. These are put in to give the electron beam in the tube time to fly back to the beginning of the next field at the top of the picture. When all these various pulses and blank lines are taken into account, a total of twenty lines are 'lost' for each field, i.e. forty for each frame. For a television system like that used in the UK, having 625 lines per frame, only 585 lines actually appear on the screen.

A system diagram for the line and field scan of a television receiver is given in Figure 15.30.

fig 15.30 *system diagram of the synchronisation and scanning circuits of a monochrome television receiver*



15.8 TELEVISION SOUND

The signal for TV sound is transmitted on a completely separate carrier; in the UK it is 6 MHz higher in frequency than the carrier used for the vision.

The two carriers are amplified together in the receiver IF amplifier

circuits, and then combined in a mixer. Since the sound is frequency-modulated and the vision is amplitude-modulated, the result is a frequency-modulated difference frequency, centred at 6 MHz, and carrying the sound. This then goes its own way, being amplified separately, detected with an ordinary FM discriminator (phase-locked loop, etc.; see above), fed to an audio amplifier and finally to the speaker.

15.9 TELETEXT

The information used to update teletext systems in receivers equipped to show 'Ceefax' and 'Oracle' (in the UK) is transmitted in the form of binary code, a series of pulses, much more closely spaced than the line-synch pulses. Teletext arrived after the television standards were all established, and made use of a feature included for other reasons—the blank lines following the field pulse sequence. The 'blank' lines are no longer blank, but are filled with the coded binary information. The teletext system selects the binary pulses and feeds them into a digital system that processes and displays the teletext information on the screen when required.

Only part of the teletext information is transmitted at any time, for it would be impossible to fit in the huge amount of binary code, even over $12\frac{1}{2}$ lines. The teletext system therefore includes a computer memory circuit that selects and holds the required information. Pressing the keys for a new teletext page makes the system 'look at' the pulses, each group of which begins with a code number (in binary) corresponding to a page of text. When the teletext system detects the number you have selected, which might take several seconds, it feeds the binary information into its memory, changing the picture on the screen. The picture then remains in the memory—and on the screen, if you want it—but is updated by the teletext system every time the selected number is detected in the binary coded information.

Complex circuits are used to mix teletext and vision, display different things on different parts of the screen, and to provide a 'newsflash' facility that puts information on the screen only when a change in content for a particular page is detected.

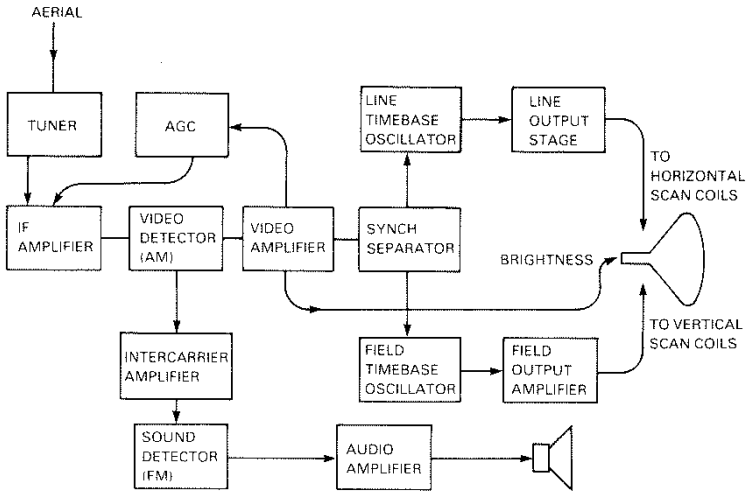
In theory, the amount of information that can be carried via teletext is unlimited, but the more different pages that are transmitted, the longer the delay in selecting a new page, as information for that page will come up less frequently.

15.10 THE COMPLETE TV RECEIVER

A system diagram for a complete TV receiver is shown in Figure 15.31.

The only circuit we shall consider, even briefly, is the *line output stage*.

fig 15.31 *system diagram of a monochrome television receiver*



This is the oscillator that provides the waveform required to make the electron beam in the tube scan the lines from one side of the tube to the other. As can be seen in Figure 15.30, the waveform is a 'sawtooth'; the waveform is applied to the horizontal scan coils, and relatively high voltages are needed. The flyback has to be *fast*, and the voltage changes very abruptly. It is possible to make good use of this. In order to produce the high voltage necessary to drive the horizontal scan coils, a transformer is used, driven by the line oscillator. An extra secondary winding is put on the transformer, called the *line output transformer* (LOPT), and with a large number of turns this winding can produce a very high voltage indeed, enough, in fact, to supply the final anode of the picture tube (see Figure 5.8). The line output transformer is therefore a vital–and dangerous– part of the power-supply circuits, as well as part of the line-scanning system.

15.11 COLOUR TELEVISION RECEIVERS

The problems facing the designers of the colour television system were formidable. The television system was well established with monochrome receivers, and it was important that the colour television system used would be cross-compatible. That is, a black-and-white receiver should be capable of receiving a colour transmission and would reproduce it correctly

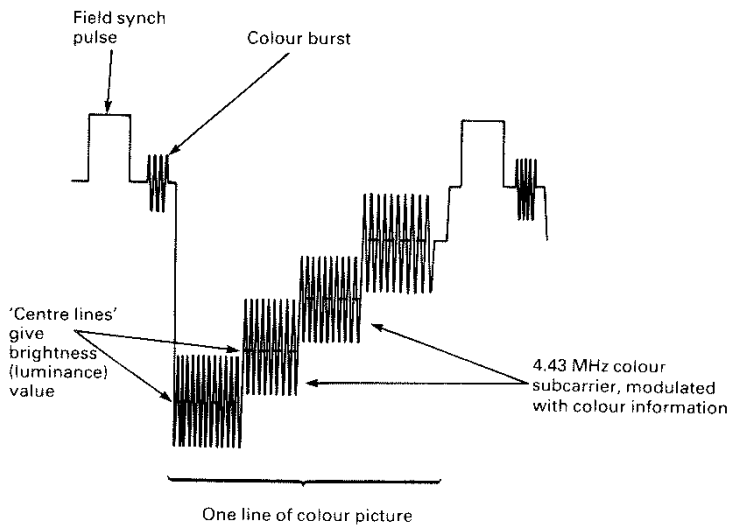
(in black and white); and at the same time, a colour receiver should reproduce a black-and-white transmission as well as a monochrome receiver.

The extra information required for the colour receiver therefore had to be 'fitted in' round the existing signal, and in such a way that it would not interfere with the operation of a monochrome receiver. The colour signal also had to be squeezed into the available bandwidth, which had been established as 8 MHz.

Since the eye is far less sensitive to colour than it is to brightness, it proved possible to transmit the colour information over a rather restricted bandwidth, so that the *colour* on a colour television is actually rather fuzzy. But if the brightness is controlled with a relatively wide bandwidth signal, giving a sharp picture in terms of brightness, the resulting picture is perfectly acceptable.

Figure 15.32 shows a waveform corresponding to one line of the same four vertical stripes illustrated in Figure 15.28 above, but with colour information added. The colour signal is actually coded during the vision information period, and at first sight seems to be likely to interfere with a monochrome receiver. But remember that the frequency of the colour signal–the *chrominance* waveform–is high, at 4.43 MHz, and would not

fig 15.32 *the waveform of one line of a colour television picture (the waveform is for a picture consisting of four vertical bars, like the one in Figure 15.28)*



be resolved by the circuits of the monochrome receiver, which would *average* the signal. The average is shown dotted in Figure 15.32, and is obviously the same as the brightness waveform—called the *luminance* waveform—of the bars in the illustration in Figure 15.28.

By means of a theoretically (and practically) complex method known as *quadrature modulation*, information about three colours can be obtained from the frequency-modulated chrominance *subcarrier*. (The waveform is called a 'subcarrier' because it is a carrier waveform, derived from modulation of another carrier of higher frequency.) The complete range of colours can be reproduced from these three colour signals, as we shall see. It is, however, necessary to have a 4.43 MHz reference oscillation, not only at precisely the same frequency as the reference oscillator back at the transmitter, but also exactly in phase with it (see Figure 15.12 for a reminder about phase relationship). This technical miracle is achieved by means of a crystal oscillator in the receiver, synchronised in frequency and phase with a short (ten cycles) burst of the 4.43 MHz transmitted just after each line-synchronisation pulse. The oscillator can be relied on to remain at the right frequency and phase for one line duration, 64 μs! The short burst of reference frequency is called the *colour burst*.

Even with a system like this, slight changes in the transmitted waveform, caused by weather conditions, reflections from aircraft, etc., can cause phase errors between the colour burst and the actual picture colours. The effect of this is a drift of the whole colour spectrum towards red or blue—not desperate, but enough to notice. Despite this problem, this system was adopted in the USA, the first country to have a broadcast colour TV network. The system is called the NTSC system, after the American National Television Standards Committee. (Opponents of the system claim that NTSC actually stands for 'Never Twice the Same Colour', but this is not true.)

The system adopted in the UK and many other countries, with the benefit of learning from the pioneers' mistakes, is known as PAL, which stands for *Phase Alternation by Line*. This overcomes the colour shift of NTSC by the ingenious method of turning round the modulation system on alternate lines. Colour shifts still occur, but alternate lines are changed towards blue and then red, and blue and then red, etc. The eye merges the colours and the overall impression is correct.

Very large colour shifts still look odd. The effect is a bit like a multi-coloured venetian blind, so a modification of the PAL system, known as *PAL-D* was introduced. PAL-D actually adds the colour signals of alternate lines electronically, by storing a whole line in a device called a *delay line* while waiting for the next line. The colour on the screen is always the product of two alternately coded signals, and the result is effectively perfect colour stability.

15.12 COMBINING COLOURS

The information recovered from the transmitted signal is about three colours only, but it is possible to use the three colours to make every colour in the visible range. In just the same way that the artist can make every colour (if he wants to) by mixing primary colours, so the three primary colours of the colour TV are mixed in different proportions to produce the whole range of colours. The three colours used are *red, green* and *blue.* Mixed together in equal proportions, these make white (white light is 'all the colours of the rainbow' mixed together), and other colours can be made by mixing two or three in different proportions. For example, green and red mixed together produce a bright yellow. This process is called *additive mixing*, and is not the same as the mixing of artist's colours, which is called *subtractive mixing.* The artist uses pigments that reflect only certain colours from the white light falling on them—they subtract colours. In the TV tube, colours are mixed—they are added together.

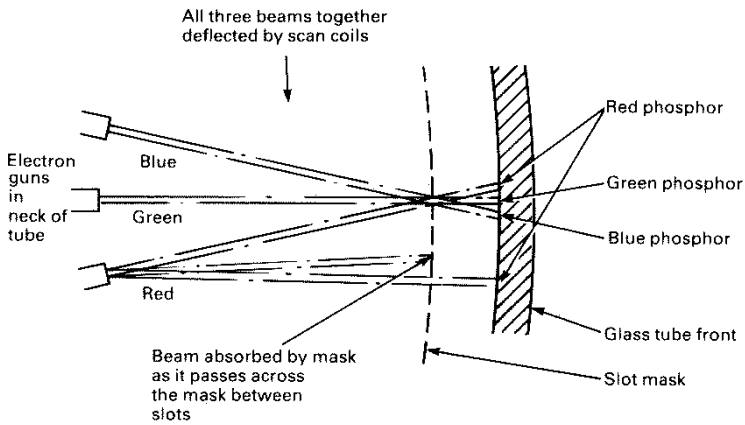
There are three factors that affect the picture on the screen. These are *brightness, hue* and *saturation*. Brightness is straightforward, and on the screen is controlled by the increase or decrease of all three colours simultaneously. Hue (colour) is controlled by the balance of the three colours, as described above. Saturation refers to the 'strength' of the colour, the amount of white light that is added to the basic colour. Red, for example, is a saturated colour. Pink is red mixed with white, and could be said to be a less saturated red. White light is not of course produced by the TV receiver, so saturation is in fact controlled by the difference between the brightness of the three primary colours; it is convenient to think of it as the colour mixed with white, however.

15.13 THE PICTURE TUBE

The colour TV picture tube is a development of the monochrome tube illustrated in Figure 5.8. The overall shape is the same, but there are three separate electron guns. In the most popular type of tube, the slot-mask tube, the three electron guns are arranged in a row, horizontally. The guns produce three electron beams, one for each colour, and although the beams are all deflected together by the vertical and horizontal scan coils, the brightness of the beams can be controlled separately.

All three beams therefore scan the front of the tube together, to produce a raster. Obviously, it is not possible to have coloured electron beams, so the colour is produced in phosphors on the front of the screen. Phosphors can be made almost any colour, and the right shades of red, blue and green can be produced easily, if expensively. The raw materials

fig 15.33 *principle of the slot-mask colour television tube*



(Not to scale!)

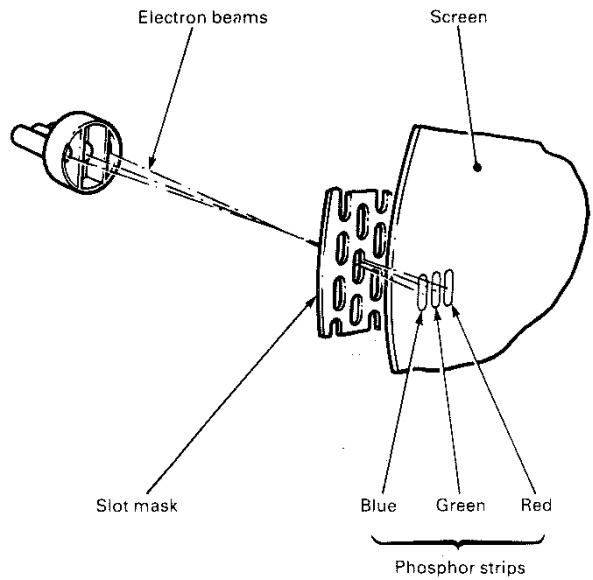
for the coloured phosphors actually contribute significantly to the price of a colour TV tube.

The trick is getting the 'red' beam to affect only the red phosphor, the 'blue' beam to affect only the blue phosphor, and the 'green' beam to affect only the green phosphor. This is done by simple geometry. Figure 15.33 shows the system diagrammatically, viewed from the top of the tube.

The slot-mask is fixed firmly in place behind the tube, and the relationship of the positions of the slots and of the phosphor strips on the back of the tube is such that the 'red' electron beam can fall only on the red phosphor, etc. Figure 15.34 is a perspective sketch of the tube, from which you can see that the slots are quite short, and are 'staggered' to produce an interlocking pattern. This arrangement is used because it is physically strong—great rigidity of the mask is clearly important—and because it gives a reasonably large ratio of 'slots' to 'mask'. The more transparent the mask is in this respect, the better, for electrons which hit the mask rather than go through the slots are just wasted power, and serve only to heat the mask. The larger the slots can be made in relation to the mask, the brighter the picture will be.

The slot-mask principle is used in the popular *precision in-line* (PIL) tube, which is manufactured complete with scan coils as a single unit. This, together with extremely sophisticated scan-coil design, produced in the first place with computer-aided design techniques, has led to a tube that is very simple to use, all the most critical alignments having been built in at the manufacturing stage.

fig 15.34 *perspective drawing of a slot-mask colour TV tube*



(Not to scale!)

A variation of the slot-mask tube is the Sony Trinitron® tube. This uses an *aperture-grill* instead of a slot-mask. The aperture-grill consists of vertical strips, like taut wires, with complete vertical phosphor stripes in different colours printed on the back of the tube. The Trinitron® tube has very good mask transparency and gives a good picture brightness. The Trinitron® tube also includes an improvement to the electron gun system, in which the three electron beams all pass through a single common point in the neck of the tube, so that they all appear to originate from the same point. This makes adjustments of the scanning system easy, and improves the focus (the sharpness of the picture on the screen).

Earlier colour TV receivers used 'delta-gun' tubes, which had the three electron guns arranged in a triangle, and circular holes in the shadow-mask tube. These tubes are not now used in new receivers, as they are rather difficult to adjust and give poorer picture brightness for a given beam power than either the PIL tube or the Trinitron®.

There is not space here to give more than the briefest glance at colour televison receivers. The circuitry is, in both theory and in practice, very com-

plicated, and extensive use is made of special-purpose integrated circuits to reduce the 'component count' and to make the receivers easy and cheap to manufacture. In real terms a colour television costs less that its monochrome counterpart of forty years ago, and gives incomparably better results.

Television servicing, especially colour television servicing, is highly specialised, and should never be attempted by the inexperienced or untrained. The voltage on the final anode of a colour receiver will be more than 25 kV, and is usually lethal.

Before leaving the subject of colour TV, get hold of a pocket magnifying glass, and take a close look at the surface of a colour TV tube, first with the receiver turned off, then with it turned on.

15.14 **SATELLITE TELEVISION**

If I were to stand on an enormously high mountain so that I was more or less outside the Earth's atmosphere, then fire a gun horizontally, the bullet would eventually curve down under the influence of gravity and hit the ground. All else being equal, the more powerful the gun, the further the bullet would go before hitting the ground. Figure 15.35*a* shows this.

Figure 15.35*b* scales things up a little, and shows that if I used a *really* big gun, the fact that the earth is a sphere and not flat means that the shell will travel further before it reaches the ground.

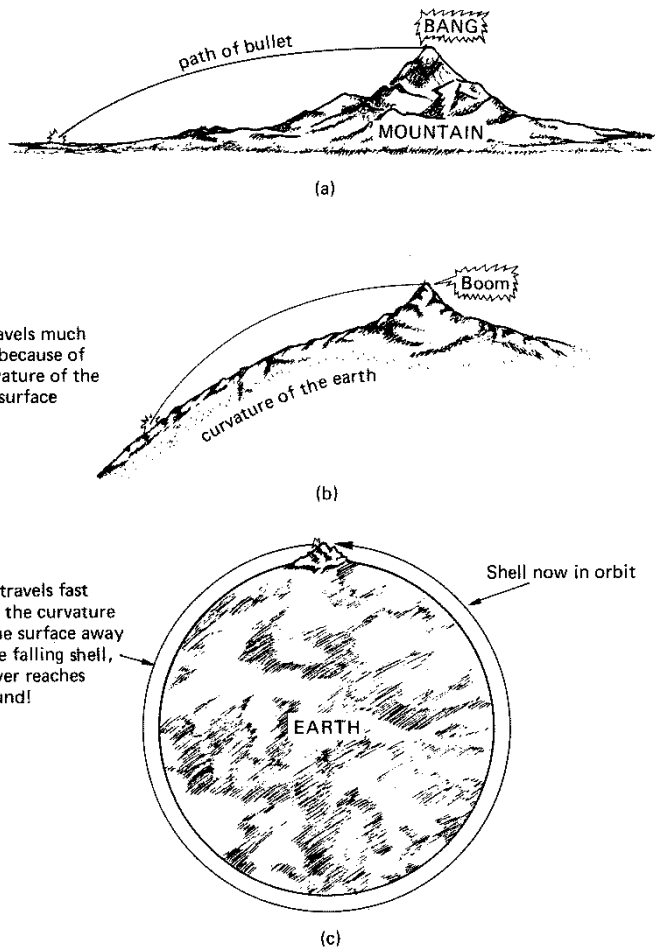
If the shell is travelling fast enough–Figure 15.35*c*–the Earth's surface will curve away underneath it just as fast as it falls . . . and unless something slows it (for example, air resistance) the shell would *never* land. It would be in *orbit*.

Now let us be more realistic and substitute a rocket-launched satellite for the shell. The faster it is travelling, the higher will be the orbit and the slower it will appear to go in relation to the earth (twice as far out it has to travel 4π as far). The moon, for example, is on average 384 400 km away and orbits the earth once every 28 days.

In between low orbits circling the Earth in a couple of hours, and very high orbits like that of the moon, there is a distance at which the orbit takes exactly one day for each revolution. That distance is 35 784 km above mean sea level. If the direction of the orbit coincides with the direction of rotation of the earth, then the satellite will appear to hover over the same point on the globe; it is in *geostationary* (or *geosynchronous*) *orbit*.

As early as 1945, the writer Arthur C. Clarke calculated the distance for a geostationary orbit and suggested that it would be a good place to put a radio transmitter. Today it is a reality. Television satellites broadcasting on

fig 15.35 *(a) the path of a bullet fired from a pistol on top of a high mountain; (b) the path of a shell fired from a cannon on top of an extremely high mountain; (c) a shell fired from a tremendously powerful cannon (on top of a ridiculously high mountain) goes into orbit*



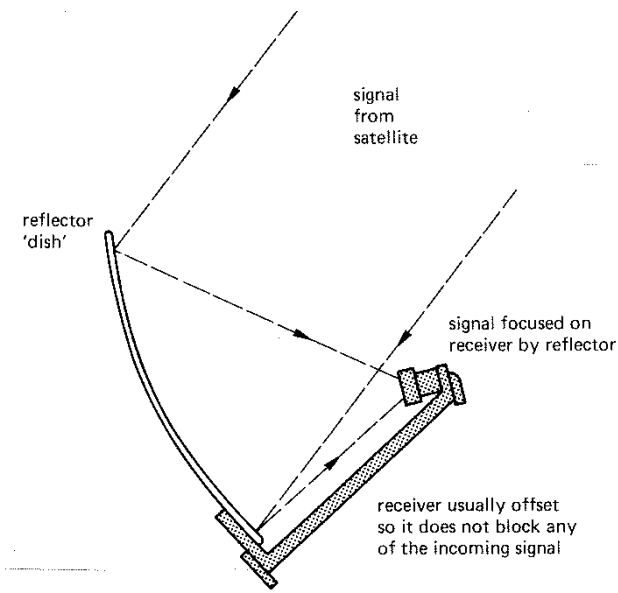
the Ku-band (11.710 GHz to 12.190 GHz) at a power of only a few tens of watts have the potential to reach tens of millions of viewers directly. Compare this to the ground-based Crystal Palace television transmitter in London, broadcasting at a power of 1 MW per channel and serving an area of only about 35 km radius!

A ground station transmits the broadcast television signal to the satellite, which automatically re-transmits it back to earth. The signal takes about a quarter of a second to reach the satellite and come back again. (Anyone who has telephoned between the UK and America will have experienced this annoying satellite link delay, which is just enough to make conversation awkward until you get the hang of it).

At 11 GHz to 12 GHz radio signals are very directional and can be reflected and focused just like light, so the familiar small 'dish' antenna is used to receive the signal directly from the satellite. A dish aerial works on the same principles as a reflecting telescope. The bowl gathers the signal from the satellite and focuses it onto a receiver (the LNB—see below). To prevent the receiver blocking some of the incoming signal from the transmitter, the receiver is often offset, as shown in Figure 15.36. This arrangement has the additional practical advantage that the dish can be more vertical and is less likely to fill up with snow. The bigger the dish, the more signal it receives, which is why larger aerials are needed in 'fringe' areas where reception is weak.

The dish focuses the signal onto the *low-noise block down-converter* (LNB) which uses the superheterodyne techniques referred to earlier to generate a 950 MHz to 1750 MHz intermediate frequency signal. This is

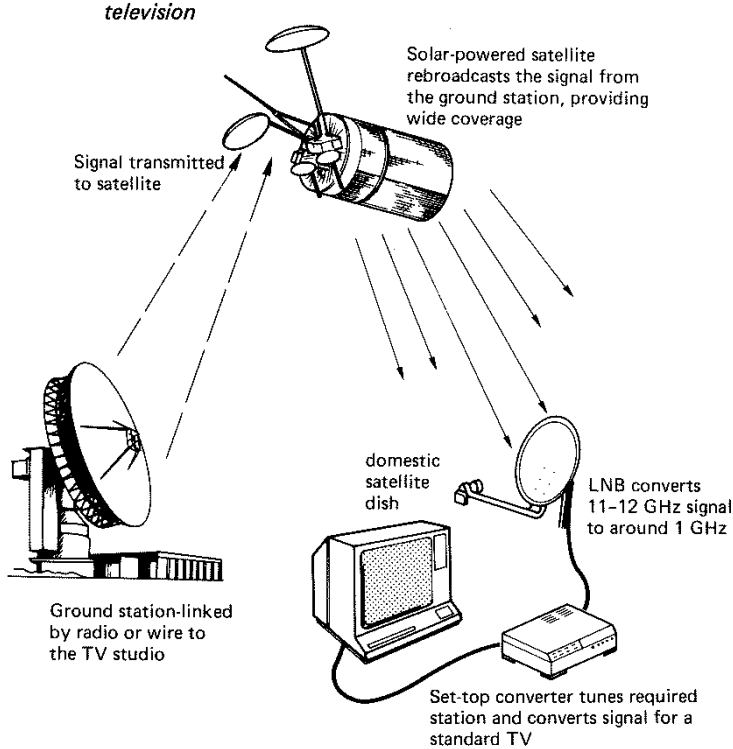
fig 15.36 *side view of a satellite TV dish aerial; the curved dish focuses the signal on the receiver unit which receiver is offset downwards*



then fed to the set-top unit, where the user can select the required transmission. The frequency is lowered again so the satellite receiver system can be plugged into a standard television. Figure 15.37 illustrates this system.

It is likely that in the near future the number of available television channels will far exceed the amount of (new) programme material there is to show.

fig 15.37 *satellite television; how the signal gets from the studio to your television*



15.15 VIDEO TAPE RECORDING

Video Tape Recorders

One of the most successful items of electronic equipment, in terms of sales, has been the *video tape recorder*, or *VTR*. The VTR can be used with a camera, or—more usually—you can buy a camera with a compact VTR built into it. The combined unit is called a *camcorder*. But by far the most common application is simply to record television programmes for subse-

quent viewing at a more convenient time. Additionally, a whole industry has built up around the hire and sale of video tapes of films made originally either for the 'big screen' of the cinema, or purpose-made for video.

Video tape recording was originally launched on several rival formats using different tape cassettes–*VHS* (video home system) and *Betamax* being the principal contenders in the UK. After a few years, Betamax has all but vanished, and VHS is the generally accepted standard. The VHS cassette uses ½ inch wide tape and provides a maximum of four hours' recording, although many VTRs provide a 'half speed' facility that gives eight hours' recording, albeit with reduced quality.

The signal that provides the picture in television can be recorded on tape in much the same way as an audio signal. It is, however, technically much more difficult to record because of the large *bandwidth*: the VTR must reproduce a much wider band of frequencies than an audio tape recorder. A magnetic recording head can record and reproduce a range of about 10 octaves (each octave doubles the frequency, so 10 octaves above 20 Hz is 20.48 kHz). VTRs use frequency modulation to provide a bandwidth of more than 18 octaves.

A second and more difficult problem is the necessity of recording high frequencies, up to 5 MHz. To record a single cycle of a waveform a length of tape equal to three times the head gap must pass by the head. Even with very small head gaps (a VHS VTR has head gaps of only 3 μm) high tape speeds are needed. Early VTRs used tape speeds of up to 9 metres per second, and it turns out that the optimum tape speed past the record head is actually quite close to this–about 10 metres per second.

The solution to the problem of obtaining a reasonable tape speed (and hence small cassettes) is to use rotating heads. Two record/playback heads are mounted in a rotating drum, and the tape wound round the drum at an angle, as in Figure 15.38. Tape is pulled steadily past the drum by the capstan–the mechanism is similar to that of an audio recorder–but at the same time the drum is spun rapidly in the other direction. Although the tape moves at only about 30 mm per second, the tilted rotating drum makes the heads whizz diagonally across the tape at the required high speed.

This system is known as *helical scanning*. Each complete scan across the tape is equal to one television field, which has the bonus that if the capstan drive is stopped, but not the drum, a still picture will be left on the TV screen. Figure 15.39 shows, although not to scale, how pictures are recorded on the tape.

It is also possible to show a recognisable picture at high speed either forwards or backwards, provided the tape speed is carefully controlled so that it is pulled round the drum at a speed that allows the drum to scan a whole number of fields. At four times the normal speed the drum will scan the first quarter of a field, the second quarter of the field after, the third

fig 15.38 *helical scanning of video tape provides a very high writing speed with a relatively slow transport speed – the sound track is recorded in a normal linear fashion along the edge of the tape*

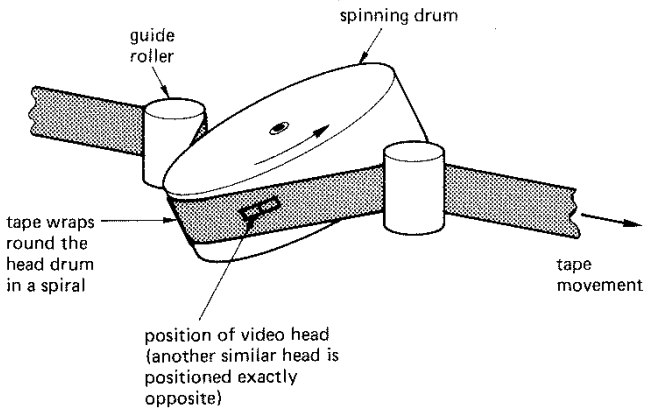
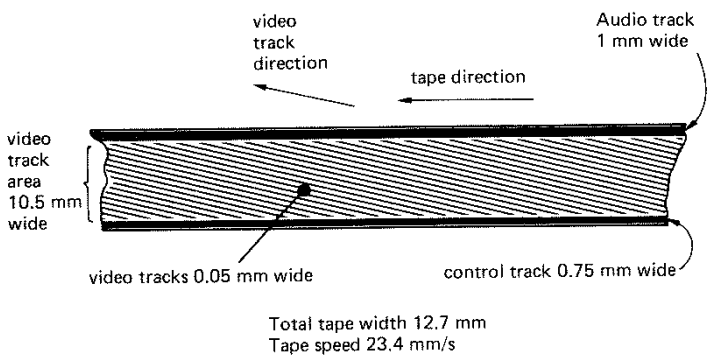


fig 15.39 *the recording format of VHS video tape (almost to scale) – the video tracks are in reality both smaller and closer together than can be shown in this illustration*



quarter of the third field, and the last quarter of the fourth field. The result will be familiar to those who have used the 'fast scan' feature on a VTR: a picture consisting of parts of successive fields with narrow bars separating them.

The recorded tracks are very narrow, about forty *thousandths* of a millimetre on a domestic VHS machine. In order to pack in as many fields as possible for a given length of tape, various technical 'tricks' are used, including the offsetting of the record/playback heads (set on opposite sides of the spinning drum) by 6 degrees in opposite directions, to reduce interference between adjacent tracks on the tape.

The VHS cassette is moderately compact, measuring roughly 19 x 10 x 2.5 centimetres, but is a bit on the large side for camcorders. The newer *video 8* and *VHS-C* ('C' standing for 'compact') formats use smaller cassettes; conveniently, VHS-C cassettes can be replayed in a standard VHS VTR with a special adapter. Most camcorders use the smaller cassettes.

Domestic VTRs include a *tuner* so that the broadcast television channels can be received; this allows you to record TV programmes off-air. They also include many 'convenience' features such as timers, remote control, etc., often under the control of a *microprocessor* (see Chapter 24).

Video Cameras

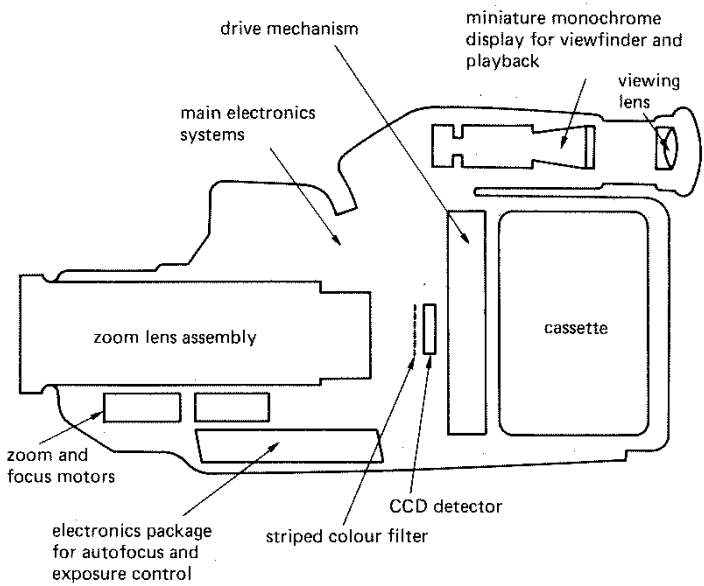
Video cameras are now about as small as the Super 8 movie cameras that were so popular in the '60s. Compared to a sound movie camera, they have the major advantage that the running cost of tape is a fraction of one percent of the cost of film, even assuming the tape is not recorded over and re-used. The only outstanding difficulty (compared with film) is in editing the recording. This is a (fairly) simple scissors job for film, but it is complex for video tape, which cannot be cut and joined.

A video camera works very much like a movie camera, except that instead of the film, film transport, and shutter mechanisms there is a *CCD (charge-coupled device) detector*. Details of the operation of a video camera are outside the scope of this book, but Figure 15.40 shows the general idea.

The striped filter in front of the CCD detector is used to provide colour information. Perhaps contrary to what you might expect (if you have read Section 15.11, dealing with colour television), it has two colours, generally green and magenta. The third colour (cyan) is calculated by the system. The resolution and sensitivity of the CCD detector is by any standards extraordinary, far superior to the older 'video tubes' that were used in earlier cameras.

There are many other aspects of the video camera that would take a whole book to explain properly. Features like automatic brightness control; automatic white balance; sound (with automatic volume control, of course); automatic focus (an electro-optical-mechanical system); zoom (optical-mechanical); playback through the viewfinder; pause; editing; and dubbing. Behind even the simplest features there are often very subtle electronic systems.

fig 15.40 *the main components of a VHS-C camcorder – most camcorders follow more or less this layout*



For example: it is almost impossible to make, for a reasonable price, a CCD detector that is completely free of defects. They all have the occasional 'bad pixel' resulting from the manufacturing process. To minimise the effect of such defects, each CCD detector is accompanied by a *ROM* (*read-only memory*) chip (see Chapter 23) that records the positions of the defects. When the camera is in use, the system refers to the ROM and 'fills in' the missing elements so that the defect is invisible on the finished recording.

I have made a more detailed exploration into a complex system in Chapter 25, looking at the basics of a compact-disc player; but it should be clear that modern consumer electronics are often *very* complicated.

QUESTIONS

1. What is meant by wavelength?
2. What kind of oscillator is used for radio control applications? Why?

3. Why does a video tape recorder produce a still picture when the tape is stopped?
4. Describe amplitude modulation, and compare it with frequency modulation.
5. Draw a circuit for a 'crystal set' radio receiver, and describe how it works.
6. What are advantages of the superheterodyne receiver?
7. Why do satellite TV receivers use a 'dish' aerial instead of the kind of aerial used to receive ground-based transmissions?
8. Why do television systems use interlaced scanning?
9. Describe the functions of the television line and field synchronising circuits.
10. Which part of the television's circuits generates the EHT supply for the final anode of the cathode ray tube?
11. Which colours are used to make the whole range of visible colours in a colour television tube?
12. Sketch the internal layout of a colour TV tube, showing how the electron gun assemblies produce the relevant colour on the front of the tube.

CHAPTER 18

INTRODUCTION TO DIGITAL ELECTRONICS

Digital electronics deals with the electronic manipulation of numbers, or with the manipulation of varying quantities by means of numbers. Because it is convenient to do so, today's digital systems deal only with the numbers 'zero' and 'one', because they can be represented easily by 'off' and 'on' within a circuit. This is not the limitation it might seem, for the *binary* system of counting can be used to represent any number that we can represent with the usual denary (0 to 9) system that we use in everyday life.

First, compare the way a digital system might be used to multiply two numbers together compared with a linear, or 'analogue', system. Let us imagine that we want to multiply 7 by 6 and arrive at the answer: 42. In an analogue system we could start with a voltage of exactly 7 V. If this were fed into an op-amp with a gain of exactly 6, the output voltage could be measured accurately and it would be found to be 42 V. This kind of 'computer' is quite feasible, but has limitations. Now suppose we want to multiply 7234 by 27 300. This is a different problem. Clearly it would now be impractical to use 1 V to represent 1, so we must use, say, 1 mV to represent 1. The input to the amplifier is thus 7.234 V. There now comes the problem of the amplifier to give a gain of 27 300—a massive design problem in itself. If it could be made, and made to work, the output would be 197 488.2 V, or nearly 200 kV! Even representing the input as 1 μV per 1, the output level would be almost 200 V—not really a practical solution, even supposing we could measure 200 V to the nearest microvolt.

A digital system, on the other hand, would treat the problem in a different way. Back to 7 times 6. Seven, in binary notation, is 111, and six is 110—see Figure 18.1 for a table giving comparable denary and binary numbers. We can represent a 0 by 0 V, and a 1 by, say, +5 V. Both six and seven could be represented using only three wires at one of two different potentials. The answer, 42, is represented in binary as 101010, which

needs six wires. To increase the capacity of the system to calculate 7234 x 27 300 simply needs more wires; 7234 in binary is 1110001000010 and 27 300 in binary is 110101010100100. Hardly a readily recognisable or convenient number for us, but it is still a way of representing a big number without having to use electronic systems needing more than a few volts. Also, the accuracy of the system in terms of reading voltage levels is not required to be high; all that is needed is the ability to differentiate in the circuits between 0 V and +5 V. The solution to the problem, 197 488 200, is actually a rather long binary number, 101111000101011011001001110, 28 binary digits, or 'bits', as the computer people call them. Although binary numbers are long, the difficulty of manipulating them is nothing compared with the problems associated with the equivalent analogue systems. Rather than use 28 wires, a computer or calculator would use a *register* holding the 28 'ons' and 'offs' in bistables, and the digits would be processed in sequence rather than altogether, but such details are part of computer science, not mastering electronics! Digital systems clearly have big advantages.

fig 18.1 *some denary numbers with their binary equivalents*

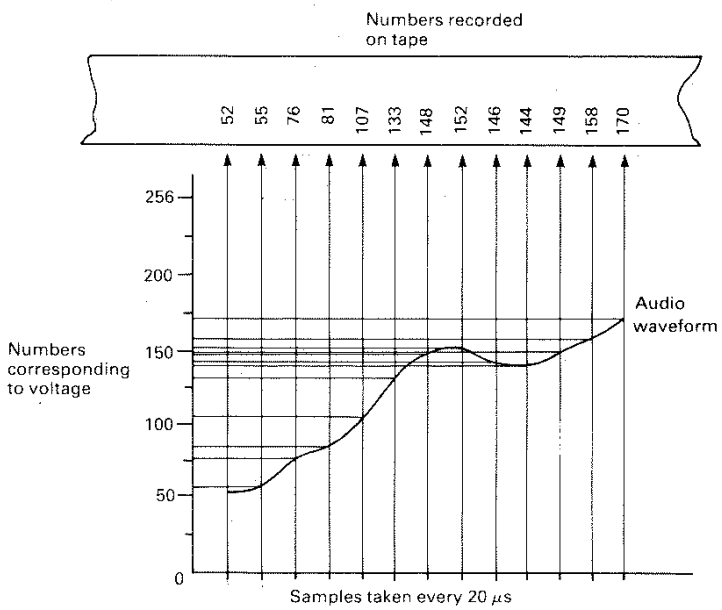
Denary	Binary
0	0
1	1
2	10
3	11
4	100
5	101
6	110
7	111
8	1000
9	1001
10	1010
11	1011
42	101010
100	1100100

A circuit involved in storing and manipulating binary numbers is rather complex, but in much the same way as a knitted sweater is complex, containing as it does a very large number of similar units. Digital systems remained the province of the specialist until the advent of microelectronics.

Digital systems are perfect for making as integrated circuits. They have large numbers of regular circuit elements, they can be made without capacitors or inductors, and they can be made to use any convenient voltage. With mass-produced ICs containing complicated circuits available for the same price as a single transistor, it has now become possible to use digital techniques in all sorts of applications that were previously the province of analogue systems.

Take digital sound recording, for example. Hi-fi sound has long been associated with the very best of linear electronics. Tape-recorders in particular have undergone spectacular developments to bring them to the present quality and cheapness—but it is digital systems that are now used to make the 'masters' for records and tapes, in the best studios. Figure 18.2 shows a very small part of an audio waveform. This could be reproduced with very little distortion on the best tape-recorders available. But using digital recording techniques the waveform is *chopped* into brief time intervals (more than twice the highest audio frequency), and the voltage level measured at that point in time. The voltage level is assigned a numerical value—8 bits would give 256 different numbers—from zero

fig 18.2 *a digital recording technique*



to maximum level. The numbers are recorded in binary form on a tape, at a rate (in this example) of somewhere over 300 000 bits per second.

To replay the tape, the numbers are turned back into voltage levels, and the original waveform reconstructed from the numbers. This may seem a complicated way of recording a concert, but the beauty of it is, if it works at all, it's perfect—or as near perfect as it needs to be.

A digital recording can be re-recorded over and over again, copied, re-copied, mixed, re-mixed, and if the numbers are readable (and they are read only as 'on' and 'off'), the music will be as perfect as the first recording, with no degradation whatsoever, and no background noise caused by successive re-recordings.

Digital recording is, it must be admitted, a rather 'high-technology' application of digital electronics. But digital systems are used in much simpler contexts, and there are surprisingly few different circuit elements in even the most elaborate system.

It is inevitable that digital electronic systems should lead to a study of computers and their organisation, for the digital computer is perhaps the greatest achievement of digital electronics technology. The following chapters, and the final section of this book, deal with digital circuit elements and complete systems, and, in the last two chapters, with microprocessors and digital computers. But you should remember that digital electronics covers not *only* computers but also a whole vast area of designs and ideas that make use of the ready availability of cheap digital integrated circuits.

One of the more impressive advances of digital electronics into the 'consumer' market has been the advent of the CD, or compact disc player. This is an entirely digital system that has provided a major advance in hi-fi audio reproduction. The CD system uses a 16-bit system, providing 65 536 different voltage levels, and a sampling rate of 44.1 kHz. Apart from the fact that nothing is lost between the original master recording and the millionth copy, the performance figures are, to the hi-fi enthusiast, quite compelling. Total harmonic distortion is 0.005%, compared with the best record at 0.15%. Signal-to-noise ratio is 90 dB, compared with the best record and player combination of 60 dB. Channel separation (that is, the amount of unwanted mixing that takes place between left and right stereo channels) is almost perfect at 90 dB, compared with a good record at 30 dB. Wow and flutter (variations in speed that affect the pitch or quality of the sound) are completely absent in the CD system.

CD systems are dealt with in Chapter 25.

CHAPTER 19

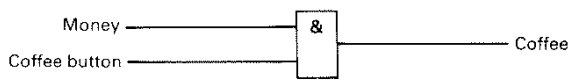
LOGIC GATES

Take a simple, everyday thing like a coffee and soup vending-machine. Before it delivers a cup of coffee, two things must happen: (i) some money must have been put in; and (ii) the button for 'coffee' must have been pressed. We can write down the requirement as

money AND coffee button = coffee

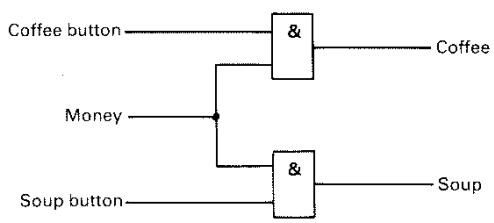
and represent this diagrammatically—see Figure 19.1.

fig 19.1 *an AND logic system*



Both inputs (money and coffee button) must be present before there is an output. If we want to design a system that gives soup as an alternative, then we need a schematic like the one shown in Figure 19.2.

fig 19.2 *a logic system using two AND gates*



Notice that one of the inputs (money) is common to both gates. This, plus the other input, gives the relevant output. As far as it goes, the system is workable, but it has a design fault. Sooner or later someone will discover

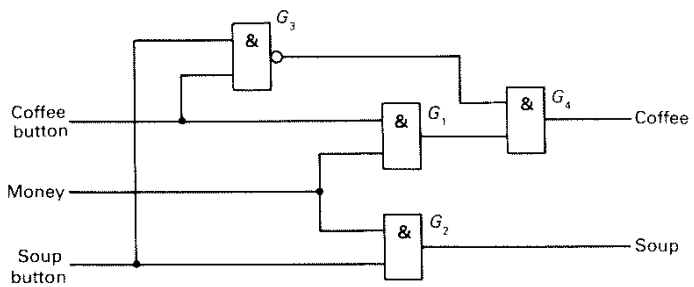
that if you press *both* buttons, you get coffee *and* soup. We could have predicted this by writing down a table of all possible combinations of logic inputs and their resulting outputs. Such a table is called a *truth table*, and a truth table for Figure 19.2 is given in Figure 19.3.

fig 19.3 *a truth table for a coffee/soup vending machine*

INPUTS			OUTPUTS	
Coffee	Money	Soup	Coffee	Soup
0	0	0	0	0
0	0	1	0	0
0	1	0	0	0
0	1	1	0	1
1	0	0	0	0
1	0	1	0	0
1	1	0	1	0
1	1	1	1	1

The last line is the one that shows up the problem. We can overcome our design problem by adding another stage to the logic, represented in diagram form by Figure 19.4.

fig 19.4 *a logic system for the coffee/soup vending machine*



We should examine the workings of this in more detail. G_1 and G_2 are our original system; G_3 and G_4 have been added. The symbol used for G_3 is slightly different, the ○ on the output representing a negation. All this means is that the sense of the output is reversed—the output is always 'on' unless both inputs are also 'on' (in which case it is 'off'). Thus the upper input of G_4 is usually 'on', and when the lower input is also 'on'

(coffee button AND money) the 'coffee' output is also on. If both buttons are pressed, the output of G_3 goes off, suppressing the 'coffee' output. All you get is soup, and it serves you right for trying to cheat! A truth table for Figure 19.4 is given in Figure 19.5.

fig 19.5 *a truth table for the system shown in Figure 19.4*

INPUTS			OUTPUTS	
Coffee	Money	Soup	Coffee	Soup
0	0	0	0	0
0	0	1	0	0
0	1	0	0	0
0	1	1	0	1
1	0	0	0	0
1	0	1	0	0
1	1	0	1	0
1	1	1	0	1

You have probably recognised the diagrams as *logic diagrams*, and the systems shown can be implemented, exactly as they are, using digital integrated circuits.

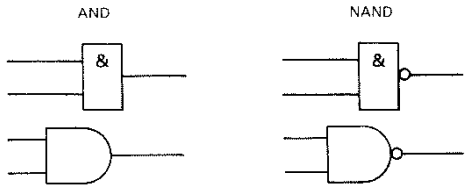
19.1 AND/NAND GATES

Before going any further with the system, it is worth pausing to look at the logic gates that are commonly used in simple logic systems. Figure 19.6*a* shows AND and NAND gates. We met these in the coffee-machine logic, and their function should be obvious by now—but the truth table in Figure 19.6*b* formalises the functions. The output of the AND gate is 1 only when *both* inputs are also logic 1; the output of the NAND gate is 0 only when both inputs are logic 1.

19.2 OR/NOR GATES

Figure 19.7 shows the symbols and truth tables for OR and NOR gates. The output of the OR gate is 1 if *either* of the inputs is 1, or if *both* of them are 1. In the same way, the output of the NOR gate is 0 if either or both of the inputs are 1.

fig 19.6



(a) *logic symbols for AND and NAND gates; the upper symbols are the recommended IEC/British Standard symbols, while those below are the American standard; both types of symbol are used, the American version being the more common*

AND

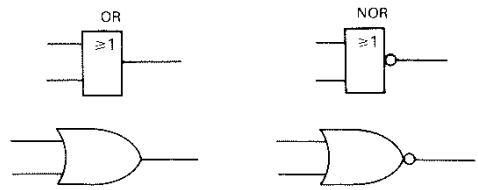
Inputs		Output
0	0	0
0	1	0
1	0	0
1	1	1

NAND

Inputs		Output
0	0	1
0	1	1
1	0	1
1	1	0

(b) *truth tables for the AND and NAND gates*

fig 19.7 *OR and NOR gates, along with their truth tables; once again, the upper symbols are IEC/BS, the lower ones the US standard*



(a)

OR

Inputs		Output
0	0	0
0	1	1
1	0	1
1	1	1

NOR

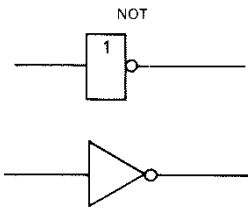
Inputs		Output
0	0	1
0	1	0
1	0	0
1	1	0

(b)

19.3 NOT GATE, OR INVERTER

Figure 19.8 shows the simplest logic element (no truth table necessary!), the NOT gate, more usually called an *inverter*. It simply changes the sense of the input, 0 output for 1 input and vice versa.

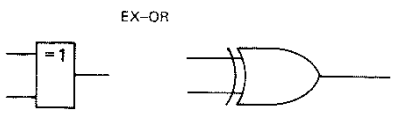
fig 19.8 *a NOT gate (or inverter); IEC/BS symbol on top, US symbol below*



19.4 EX-OR GATES

Figure 19.9 shows an Exclusive-OR gate, along with its truth table. The Exclusive-OR gate is just like an ordinary OR gate, *except* that the output is only 1 when just one of the inputs is 1. If more than one input is at logic 1, or if all inputs are at logic 0, the output is 0. EX-OR gates are less commonly found than the AND/NAND, OR/NOR varieties.

fig 19.9 *an exclusive-OR gate, with its truth table; IEC/BS symbol on the left.*

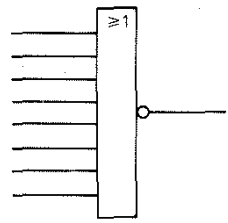


Inputs		Output
0	0	0
0	1	1
1	0	1
1	1	0

19.5 MORE COMPLEX GATES

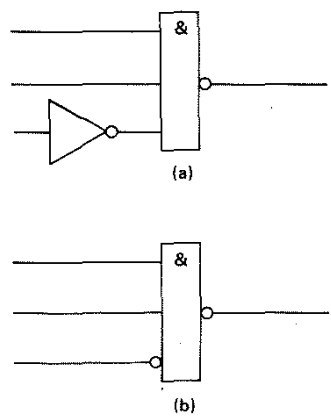
It is possible to ring the changes on the basic gates. Gates of all types often have more than two inputs–Figure 19.10 shows a CD4078 eight-input NOR gate. If one or more inputs are at logic 1, the output is 0. If all the inputs are 0, the output is 1.

fig 19.10 *an 8-input NOR gate, IEC/BS symbol: from now on, throughout this book, IEC/BS symbols are used; the equivalent US symbol is easily derived by comparison with Figures 19.6 to 19.9*



Gate *inputs* can also be negated, reversing their sense. This could be done using an inverter in the input line (Figure 19.11*a*); or if the inverter is part of the same IC as the gate, it can be drawn like the one in Figure 19.11*b*.

fig 19.11 *a NAND gate with one input inverted–two alternative representations*

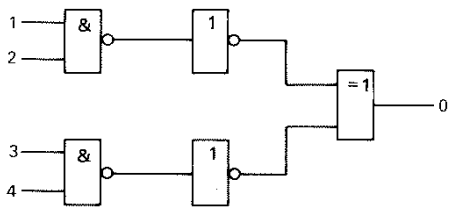


19.6 MINIMISATION OF GATES

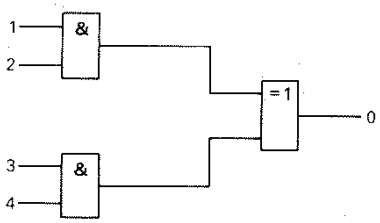
It is only sensible to design any logic system with the smallest possible number of gates, all else (the price of different types of gate) being equal. For simple systems it is usually possible to minimise the number of gates by inspection; for more complicated systems boolean algebra can be used to formalise the problem and arrive at a solution (see N. M. Morris, *Digital Electronic Circuits and Systems*, Macmillan, 1974).

Clearly, it takes only a short time to realise that the system in Figure 19.12*a* can be reduced to the system in Figure 19.12*b* without any change in its functioning. It is less obvious that Figures 19.13*a* and 19.13*b* are functionally the same—but it is obvious if you draw up the truth table for them.

fig 19.12



(a) *a logic system*

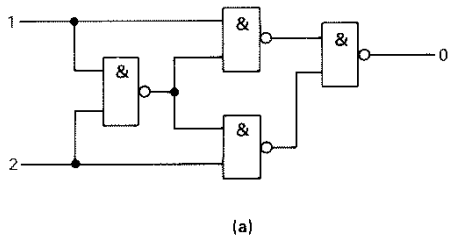


(b) *a simpler version, performing the same function*

Figure 19.14 shows a simple logic system that was actually used in a cassette tape-recorder intended for use with a microcomputer. F_1 and F_2 are inputs controlled by the computer, to stop and start the recorder motor. The two gates on the right form a *bistable* (see Chapter 21) which 'latches' the motor on or off. Switch '*M*' is for manual operation, and turns the motor on by applying a logic 1 pulse to the gate.

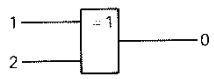
At first sight the circuit can be simplified by avoiding the double negation in the NOR gate and inverter, replacing them both with a single OR

fig 19.13



(a)

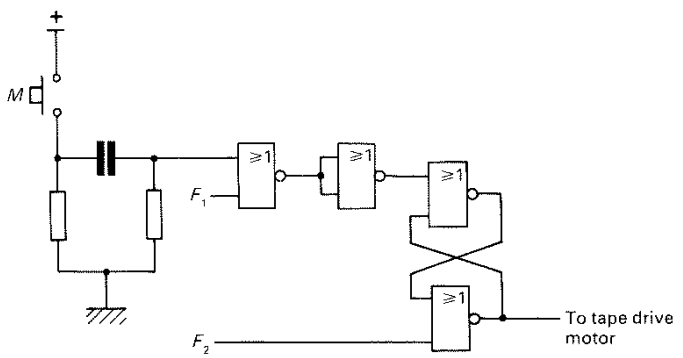
(a) *a logic system*



(b)

(b) *an exclusive-OR gate performing the same function*

fig 19.14 *a bistable control for a tape drive motor*



gate, as in Figure 19.15. So why would a sensible (and cost-conscious) manufacturer prefer the more complicated version in Figure 19.14? The answer lies in Figure 19.16, which shows the pin-out of the SN74LS02 integrated circuit.

This integrated circuit includes four NOR gates (it is known as 'quad 2-input NOR gate'), which is just right for the system in Figure 19.14. You cannot obtain a single integrated circuit that incorporates an OR gate plus NOR gates, so the apparently simpler version in Figure 19.15 actually requires two integrated circuits, which in turn means more circuit-board space, more soldered connections, and higher assembly costs. In almost all

fig 19.15 *a simpler but more expensive version of the circuit given in Figure 19.14*

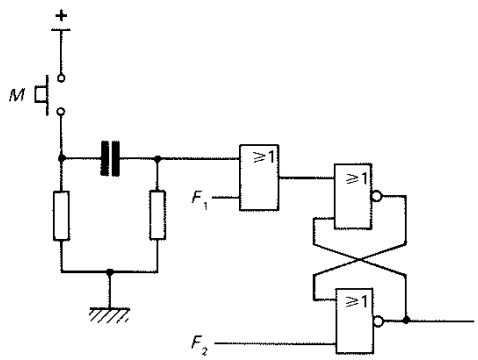
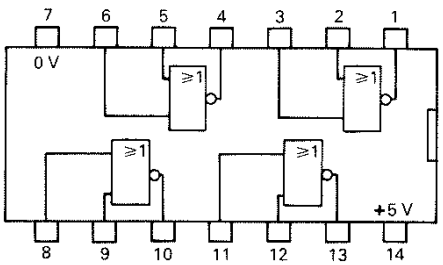


fig 19.16 *the pin connections of the SN74LS02 quad NOR gate*



SN74LS02 package

practical systems, minimisation of *components* and *cost* is the requirement. Reducing the number of gates in a given system may not always further that end.

Where large systems are involved (e.g. counters, clocks and computing circuits) it is almost always economic to use a large-scale integrated circuit (LSI) or even a microprocessor, despite the fact that only a fraction of the device's total capacity may be used. Very complicated LSI circuits are relatively cheap—a Z80 microprocessor chip costs less than twenty times as much as the simplest gate.

Having covered the basic types of gate, in the next chapter we shall consider the two major kinds of technology that are used in the implementation of integrated circuit logic elements.

QUESTIONS: CHAPTERS 18 AND 19

1. Why is the binary system of numbers used in computers?
2. Sketch a logic system in which the following truth table is implemented:

INPUTS			OUTPUTS	
1	2	3	1	2
0	0	0	0	1
0	0	1	0	0
0	1	0	0	0
0	1	1	1	0
1	0	0	0	0
1	0	1	0	0
1	1	1	1	1

3. Draw circuit symbols for (i) a 3-input NAND gate, (ii) an inverter, (iii) a 4-input OR gate with one negated input.
4. What is the largest (denary) number that can be represented by eight binary digits?
5. Draw a logic circuit including one EX-OR gate and one NAND gate, any connections. Now construct a truth table for the circuit.

CHAPTER 20

LOGIC FAMILIES

Although there are many ways of making integrated circuits, and several technologies are available to the manufacturer, two major 'families' of general-purpose logic ICs remain the most important. The first, and oldest, family is TTL. TTL stands for 'transistor-transistor logic', this logic family having replaced the older (now obsolete) DTL—diode-transistor logic. The second logic family is CMOS (pronounced 'sea-moss') which stands for 'complementary metal oxide semiconductor'. Both families include a very large number of different devices: gates, flip-flops, counters, registers and many other 'building-block' elements. In many cases there are direct substitutes, but, as we shall see, it is not usually possible to mix devices from the two families.

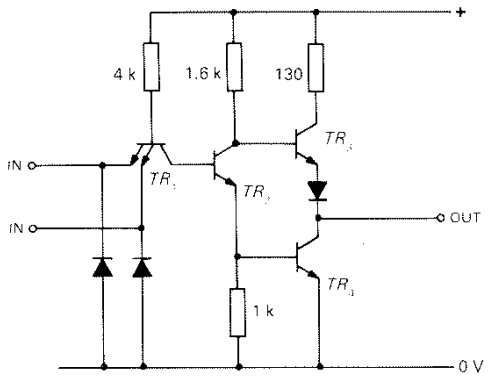
Both groups have their own special advantages and disadvantages, and when a designer is working on a digital system the choice of logic family will probably be one of his first decisions.

20.1 TTL

This family is based on the bipolar junction transistor, and was perhaps the first commonly available series of logic elements. Although the logic designer rarely needs to know how the inside of an IC works, it is instructive, when comparing logic families, to look at the workings of a simple gate. We can take as an example a 2-input NAND gate, numbered, appropriately enough, SN7400 in the TTL series. Figure 20.1 shows the circuit of this gate.

Notice that the circuit has a familiar 'transistor amplifier' look about it, except that the circles round the transistors have been left out—they represent the encapsulations. The input transistor is also unusual, by discrete component standards, in that it has two emitters. The circuit operates as follows. TR_2 is normally off (both inputs logic 0), which means that TR_3 is on, the 1.6 kΩ resistor providing the base current. TR_4 is off, its base

fig 20.1 *a typical bipolar 2-input NAND gate*



being clamped to 0 V by the 1 kΩ resistor. The output is therefore at a potential just below that of the positive supply, or logic 1. If *both* inputs are taken to logic 1, then a base current is provided for TR_2, and TR_3 switches off--its base is taken to 0 V via TR_2 and the base–emitter junction of TR_4. TR_4 switches on. The output is now at a potential near that of the supply 0 V, or logic 0.

It is clear that moderate amounts of power are required, since either TR_3 or TR_4 has to have a base current flowing the whole time. The SN7400 requires about 2 mA per gate at its operating voltage of 5 V—not a lot, but enough to make power supplies a headache in complex systems and enough to rule out battery operation for all but the simplest systems. It is equally clear that there is a definite limit on the amount of current each gate can provide; it varies from device to device, but the SN7400, which can provide at its output around 400 μA at logic 1 and 16 mA at logic 0, is typical. There is a big difference between the current that a TTL IC can *source* (i.e. when it is at logic 1) and *sink* (i.e. when it is at logic 0). Practically, this is of no importance when interconnecting gates since TTL inputs present a very high resistance when looking at a logic 1. It is of importance when *interfacing* TTL with other systems and devices, as we shall see later.

Fan-out

Manufacturers quote the capability of the output of a TTL circuit in terms of the number of standard TTL inputs it can successfully drive. This figure is termed the *fan-out*. For SN74-series TTL ICs, the fan-out is usually 10: each output can drive up to ten inputs.

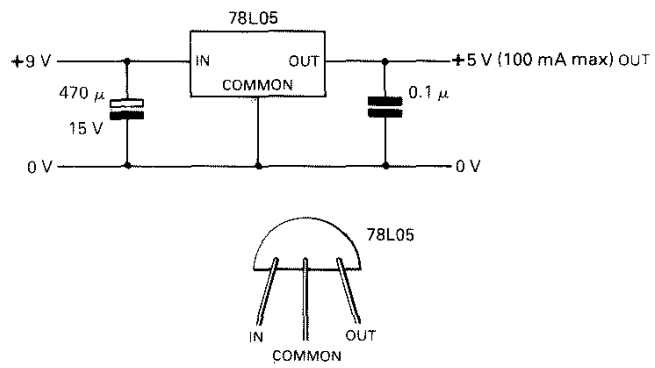
Operating speed

TTL gates work very quickly. The SN7400, for example, takes just fifteen nanoseconds to change state. TTL counters–which we shall be meeting in the next chapter–can count at frequencies up to 50 MHz, and the faster LS series (see below) at up to 100 MHz.

Power supplies

TTL logic circuits require a standard power supply of 5 V. The amount of tolerance is not large, and the limits of reliable operation are between 4.5 V and 5.5 V. It is therefore imperative that some form of power-supply regulation be used. Power-supply circuits form an interesting and quite important branch of electronics in themselves, and the design of a good power supply used to be very difficult. Fortunately, we now have cheap and extremely reliable power-supply *regulator ICs* which greatly simplify the job. Figure 20.2 shows a supply suitable for use with a 9 V battery (the ubiquitous PP9), capable of providing up to 100 mA.

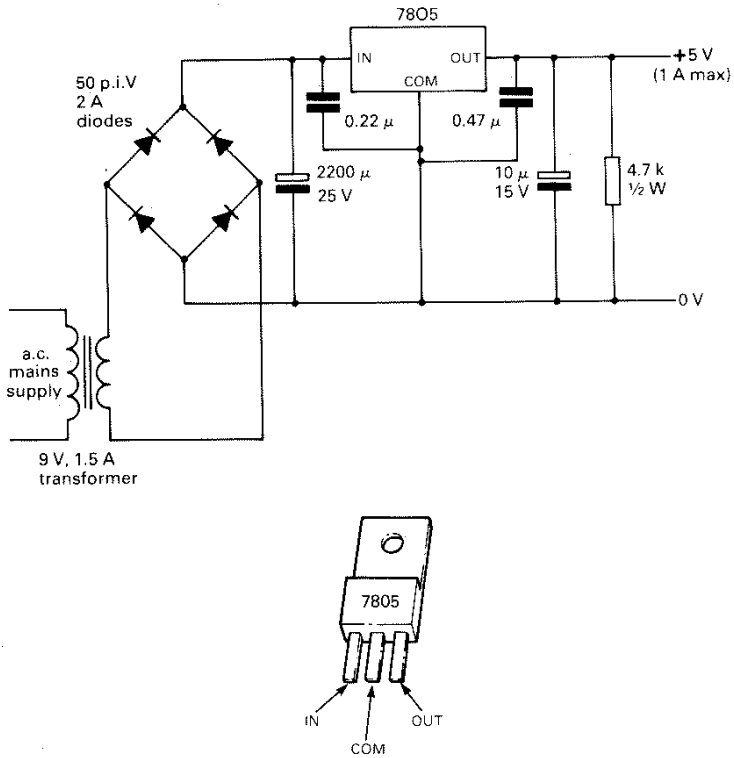
fig 20.2 *a power supply for operating small TTL logic systems*



The regulator IC, the 78L05, has a bigger brother, the 7805, which can supply up to 1 A. Regulators to provide up to 5 A at 5 V are commonly available. Since this represents 25 W (5 x 5), and the power is all dissipated by the ICs, it is clear that there is a cooling problem with large TTL systems! Fan cooling is quite commonly used.

A complete mains power supply for TTL circuits, capable of delivering up to 1 A at 5 V, is illustrated in Figure 20.3. The 7805 is protected against overheating and being shortcircuited, and even when used with experimental systems is quite hardy. Needless to say, the mains part of this unit must be effectively insulated.

fig 20.3 *a power supply for operating more complex TTL logic systems*



A TTL system can impose very heavy instantaneous loads on a power supply. This can in turn result in interference 'spikes' in the power lines, upsetting the normal working of parts of the system. A cure can usually be effected by connecting small capacitors directly across the power lines, as near as possible to the ICs themselves. Usually 0.1 μF ceramic capacitors are recommended, but 10 μF, 15 V tantalum 'bead' capacitors can sometimes prove even better. Whichever type is used, about one capacitor per five ICs seems to work well.

Unused inputs

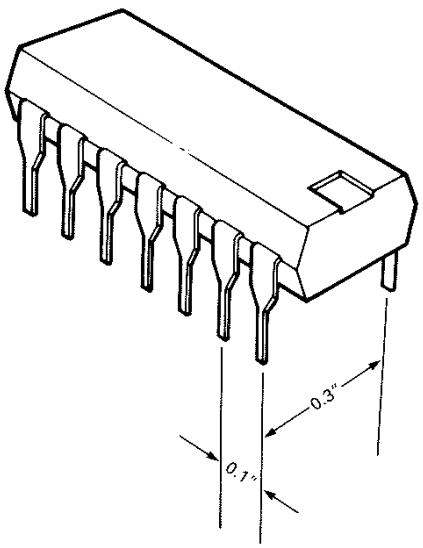
It is important to ensure that unused inputs of TTL gates are connected, either to *used* inputs (but remember the fan-out limit of the preceding device), to the positive supply, or to supply 0 V. Connections to 5 V

should be made through a 1 k resistor; you can connect several inputs to the same resistor. Failure to terminate unused inputs may lead to the gate oscillating, harmful because it may cause the circuit to overheat.

Encapsulations

Although many encapsulations are listed by the makers, the DIL package (dual-in-line) is practically universal. A 14-pin DIL package is illustrated in Figure 20.4. The pins are always numbered from pin 1, which is marked with a notch, or is the pin to the left of a notch in the end of the package, with the notch to the top. The pin numbers run in an anticlockwise direction, looking at the top of the package (which is held 'legs down'). The DIL package is usually made of plastic, though some high-quality devices may have ceramic packages, which are much more expensive.

fig 20.4 *a 14-pin DIL package*



Schottky TTL

TTL circuits are being improved continually, and a major advance has been in the introduction of a range of *low-power Schottky TTL* circuits. They use the same code numbers as the standard series, but with 'LS' before the type code—e.g. SN74LS00. Operating speeds are about twice as high, and

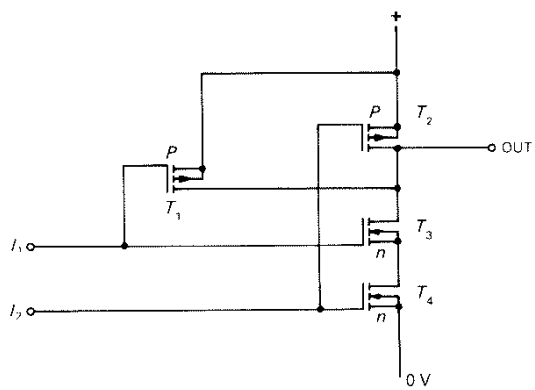
power consumption is as low as one-fifth of that of standard TTL. The only 'minus' is the cost, about 20-30 per cent more than the standard range.

Mainly because of the much reduced power consumption, the Schottky TTL is now more widely used than the standard TTL. In any digital system, and particularly in a large one, power supplies represent an appreciable fraction of the total cost, and the savings that can be made here will more than pay for the higher cost of the 'LS' series circuits. Indeed, it is the problem of power supplies that has been an important factor in the rapid growth and acceptance of the other major logic family, CMOS.

20.2 **CMOS**

Whereas TTL ICs are based on the bipolar junction transistor, CMOS ICs are based on the field-effect transistor. Figure 20.5 illustrates a CMOS 2-input NAND gate. Before looking at the way this circuit operates, bear one fact in mind: the MOSFET is a voltage-operated device, and the gate insulation means that, for all practical purposes, there is no gate current. It is this that enables CMOS ICs to achieve astonishing power economies.

fig 20.5 *a 2-input NAND gate using CMOS technology*



Clearly, therefore, the inputs for the NAND gate in Figure 20.5 will take no current. If the output is feeding another CMOS gate, then the output will not be required to deliver any current!

The operation of the gate shown in Figure 20.5 is as follows. Consider the circuit with both inputs taken low (logic 0). With I_1 low, T_3 will be off (high resistance) and T_1 will be on. With I_2 low, T_4 will be off, and

T_2 will be on. This much should be quite clear—if it is less than clear, you should re-read Chapter 9. The output terminal is held at a high potential (logic 1) via T_1 and T_2.

Now let us take I_1 to logic 1. This will switch on T_3 and switch off T_1, but the output will remain at logic 1 because T_2 is still on and T_4 is still off. The same thing happens if I_2 and not I_1 is taken to logic 1. If, however, both inputs are at logic 1, then T_1 and T_2 are both off, isolating the output from the positive supply, whereas T_3 and T_4 are both on, making the output logic 0. Obviously, the simple circuit of Figure 20.5 functions effectively as a NAND gate. It is a characteristic of CMOS that most power is used during the transition from one state to another, so the average power required is a function of the switching frequency. Typically, a NAND gate like the one in Figure 20.5 would use less tha 0.1 μA with a 5 V supply when working!

CMOS has some other advantages. It can be used with a wide range of power-supply voltages—the RCA 'B' range, for example, works from 3 to 18 V. Because of the lack of any appreciable voltage drop across the IGFETs when they are turned on, the logic 1 and logic 0 outputs are very close to the power-supply voltages—only about 10 mV below the + supply or above 0 V. CMOS logic also switches neatly at half the supply voltage. For a 5 V supply, the output switches as the input crosses 2.5 V; for a 12 V supply this happens at 6 V; and so on.

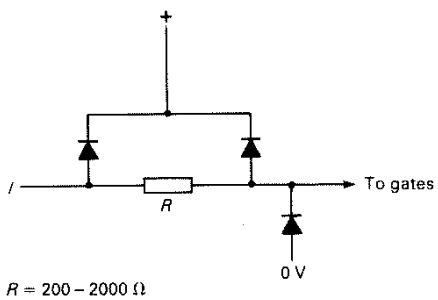
Finally, CMOS has far greater immunity to power-supply 'noise' than TTL and will tolerate noise of at least 20 per cent of the supply voltage without problems. Clearly, CMOS is the first choice for battery-operated equipment. A complex CMOS system can be run from a small 9 V battery, and the battery will have a very long life; the voltage can drop to below 6 V before the circuits are affected adversely. No power supply or smoothing circuits are generally required, though a small (10 μf or so) capacitor is often connected across the battery. In general, power-supply decoupling capacitors such as tend to be sprinkled liberally around TTL circuits are not needed for CMOS.

But there are disadvantages, too. CMOS is slower than TTLs. A CMOS counter might typically work at up to 5 MH—quite quick, but only a tenth as fast as a corresponding TTL circuit. The available output current is also low. This is not, of course, a problem within a wholly CMOS system (fan-out of a CMOS gate is more than 50) but means that buffering circuits are generally needed to drive indicators or control relays. The output is equivalent to a series resistor of 400 Ω–1 k, so by Ohm's Law we can see that with a 10 V supply only a few milliamps can be taken from the output. In practice, the output current is further limited by the safe dissipation of the device, around 200 mW per *package* (i.e. 50 mW per gate for a quad-NAND gate like the CD4011).

Handling CMOS

As we saw in Chapter 9, the MOSFET is susceptible to damage from electrostatic voltages. The same is true of the CMOS IC, though the manufacturers build in various safety features. A CMOS input actually looks like Figure 20.6. This protection system works for voltages up to 800 V–4 kV depending on the device. But since your body can easily pick up a charge of 10 kV when you walk across a nylon carpet, precautions are still essential. A suitable bench top for working with CMOS is made of a conductive material and is earthed.

fig 20.6 *typical protection circuit for CMOS inputs*



High electrostatic potentials can also appear on the bit of a soldering iron. A rather inconvenient cure is to earth the bit (but not through the earth pin of the power socket). A better cure is to use a soldering iron with a ceramic shaft—the ceramic conducts heat from the element to the bit of the iron, but with very high electrical resistance and low capacitance.

The use of a ceramic-shafted soldering iron and earthed conductive bench top should provide sufficient electrostatic protection for all modern CMOS devices.

Types of CMOS integrated circuit

CMOS technology has evolved to a point where various 'sub-families' are available. The first CMOS ICs were coded with type numbers in the range '4000–4999' (like the CD4011 mentioned above) and are known as *4000 Series* CMOS.

More recent types have even lower power consumption and much better high frequency performance, comparable with Schottky TTL. These CMOS devices have been given codes beginning with '74HC. . .', but are pin-for-pin compatible with the corresponding 4000 Series devices. For

example, the counter used in Figure 22.23 (marked '4017' in the diagram) could be an 'old' CD4017, or the newer 74HC4017.

Another range of CMOS devices, prefixed '74HCT. . .' are designed to be used in place of TTL (standard or LS) ICs, and feature inputs that have the same characteristics as TTL. They give a similar performance to TTL but use much less power.

For high-speed circuits, the '74AC. . .' series CMOS provides substantially faster switching than Schottky TTL with much lower power consumption.

20.3 COMPARISON OF TTL AND CMOS

It is useful to compare the main characteristics of some of the various 'branches' of the two major logic families. The circuit designer working with digital logic will generally choose the logic family according to his design requirements—power supply (mains or battery?), operating speed, outputs required, and of course price.

Figure 20.7 compares some of the major characteristics of the different types of integrated circuits for digital logic that are currently available. The older ones—like the 4000 Series—have few advantages over the newer types, and will gradually be phased out. Competition between major manufacturers makes it certain that new ranges will appear in the future.

fig 20.7 *comparing the characteristics of various types of digital logic integrated circuits*

type	supply voltage	typical supply current per gate	typical power per gate	output current		maximum operating frequency	fan-out
				logic 1	logic 0		
TTL							
74 . . .	5 V	40 μA	10 mW	500 μA	16 mA	35 MHz	10
74LS . . .	5 V	20 μA	2 mW	300 μA	8 mA	45 MHz	20
CMOS							
'4000BE'	3–18 V	10 pA	0.6 μW	1 mA	2.5 mA	5 MHz	50
74HC . . .	2–6 V	10 pA	1 μW	4 mA	4 mA	40 MHz	> 500
74HCT . . .	5 V	10 pA	1 μW*	4 mA	4 mA	45 MHz	> 500
74AC . . .	2–6 V	10 pA	1 μW	24 mA	24 mA	100 MHz	> 500

*power dissipated by 74HCT types increases with switching speed and is more than 1 mW at 4 MHz

QUESTIONS

1. The output of a TTL gate is connected, via a meter that reads current, to the 0 V supply line. Suggest typical current readings when the output is at (i) logic 1, (ii) logic 0.
2. What are typical power-supply requirements for (i) TTL, (ii) CMOS?
3. Compare the TTL and CMOS logic families for the following characteristics: (i) current consumption from the supply, (ii) speed of operation, (iii) immunity from interference, (iv) tolerance of variations in the supply voltage.
4. What precautions need to be taken when handling CMOS circuits?
5. Which logic family would you use in designing a system for the following uses: (i) a stop-watch, (ii) high-speed data transfer system for a computer, (iii) control logic for an electronic ignition system in a boiler?