*~~$5.95~~

Electronics — It's Easy

Vol.2

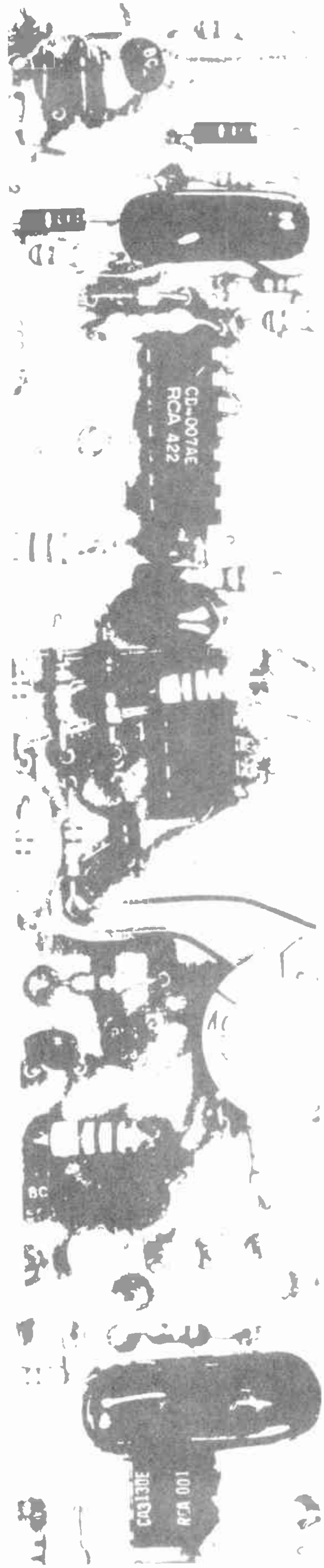
by Peter H. Sydenham, M.E., Ph.D., F.Inst.M.C., F.I.I.C.A., M.I.E.Aust, Professor and Head of School of Electronic Engineering, South Australian Institute of Technology.

eti

AN ELECTRONICS TODAY INTERNATIONAL PUBLICATION

Electronics —It's Easy Vol. 2

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



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

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

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

Acknowledgement

Electronics it's Easy was first published as a monthly series in Electronics Today International. It was republished as three separate volumes during the period 1975-1976.

A second edition was published in 1977-1978 and a revised version of Volume 1 was produced especially for the TAFE School of Electrical Engineering in Sydney in 1980.

This, the third edition of the complete series, has been totally revised and new material added where relevant. At the same time the material has been re-arranged so that the complete work now appears in two volumes rather than the original three. Apart from a card-cover edition, the work is also available in hard-back form — details at the end of both volumes.

The author would like to thank Brian Chapman (technical editor of Electronics Today International from 1973-1976) for editing the original edition. Thanks also to Ron Farley and Maree Breen for turning hundreds of my rough pencil sketches into clear and concise engineering drawings.

The very considerable work of turning my original individual articles into book form, and of editing and producing this latest revised edition, has been done by Janet Vernon B.A., special publications editor of Electronics Today International.

I wish to thank the following for use of illustrations:—
Philips; BWD; Science Museum (London); Advance Instruments; Electronics Australia; Hewlett-Packard; Teledyne Philbrick; Siemens; Tektronics; Non-Linear Systems; John Fluke; Heidenhein; Lloyds Bank; AEI Associates; Kirklands; Novus; Gerber Scientific; Honeywell; Motorola; National Semiconductor; Levers Rich; Andrew Antennas; Vu-Data; Data-Lab; Marconi; Australian Post Office; Kent Instruments' Record Electrical Company; Goerz, Hathaway Instruments; Devices; and Kepco.

P Sydenham

Peter H. Sydenham. Adelaide, April 1981.

21 Introducing digital systems

WE BEGAN this course with a discussion of electronic systems in general; what they do, how they do it and how we can progressively break down a complex system into fundamental building blocks. The example chosen then, a TV system, uses, in the main analogue signals. These we know from other parts of the series are those signals that contain information in the form of many continuously changing levels of an electrical voltage (or current).

Although we have already introduced the concept of the on-off, or digital kind of signal, the course so far has concentrated almost entirely on the linear, analogue circuits used in electronic systems. The time has now come to study an alternative philosophy and practice, by which tasks can be accomplished in another manner — the digital electronic approach.

DIGITAL OR ANALOGUE SIGNAL APPROACH?

By itself a purely electronic system has no real value until it is applied to the real world we exist in. At the input of a system physical variables are measured by sensors that convert the information, from the original form of energy, into an electrical signal. This electrical-input signal is then conveyed, through the system being modified, and converted in different ways as required. The output signal from the system is fed to actuators which convert this signal back to real-world variables at the output. It is the differences between two basic means of transmitting and converting information that we are concerned with now.

We have seen in the earlier part dealing with information that both analogue and digital signals can convey the same information between two points. It is a matter of how the information is coded on the signal. It is not possible to state categorically that one signal form is better than the other. Each has its advantages depending upon the application. Analogue systems can process the same information using far less components, than their digital counterparts, but they are unable to provide anywhere near the same ultimate accuracy, precision and long-term stability. In some uses, such as precise mathematical computation, digital techniques are a must. The same holds true for measuring equipment needing better than around 1 percent, or perhaps 0.1 percent, accuracy.

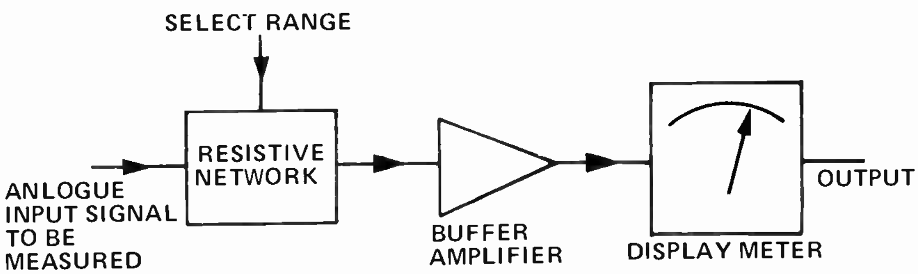


Fig. 1. Schematic representation of analogue multimeter system.

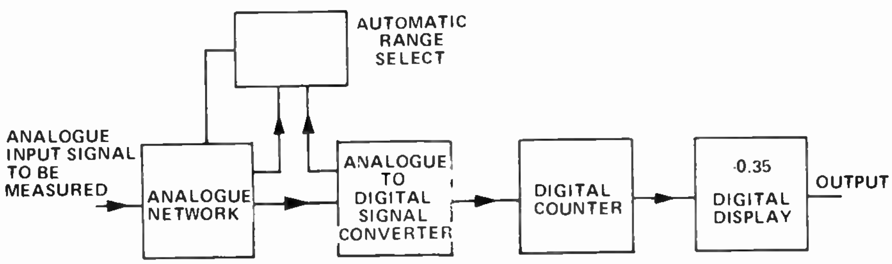


Fig. 2. Digital multimeters contain black boxes that operate with digital rather than analogue signals.

Other factors that decide the choice of signal form are the cost of components needed, the size of equipment and power supply demands. Today, the enormously large-volume production of digital circuits, especially when marketed as large-scale integrated systems, coupled with the tremendous effort that has been expended on digital techniques for computing markets, has now tipped the balance heavily in favour of using digital methods. This is now true even for what have traditionally been analogue applications. It may well now be cheaper to use a mass-produced digital assembly for a more unusual analogue requirement, even when analogue circuits could easily supply the need.

Take, for example, the choice confronted when purchasing a good quality multimeter. The traditional multimeter can be represented as a resistive network driving a display meter — see Fig. 1. The signal level can be ascertained by the degree of pointer deflection seen on the meter. High input impedance units incorporate a linear amplifier to buffer the signal source against a relatively low-impedance meter movement. Apart from the selector switch which has discrete settings, all components work with analogue signals and this means they must be linear in operation and adequately stable with time. Some components — the ballast and shunt resistors, for instance — must be made to tolerances that require expensive hand-made manufacture. We can summarize the situation as one where only a few components are needed but they are inherently expensive.

The alternative is to use a special circuit that we will discuss in detail in a later part. This is called an analogue-to-digital converter (or just A-to-D converter). As represented in the schematic of a digital multimeter given in Fig. 2, it converts the analogue input level into a digital signal form that is then used to drive a digital readout display. These units display the output value as a decimal number rather than as the position of a pointer as is used in totally analogue systems. We will see, as we delve more deeply into how such a system works, that the digital alternative uses literally dozens of active elements and many

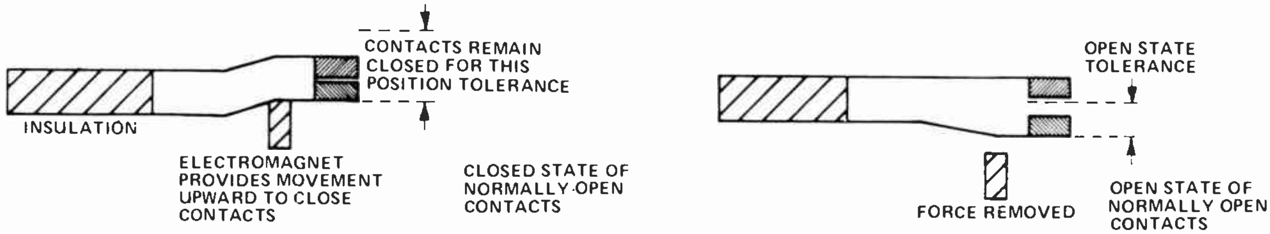
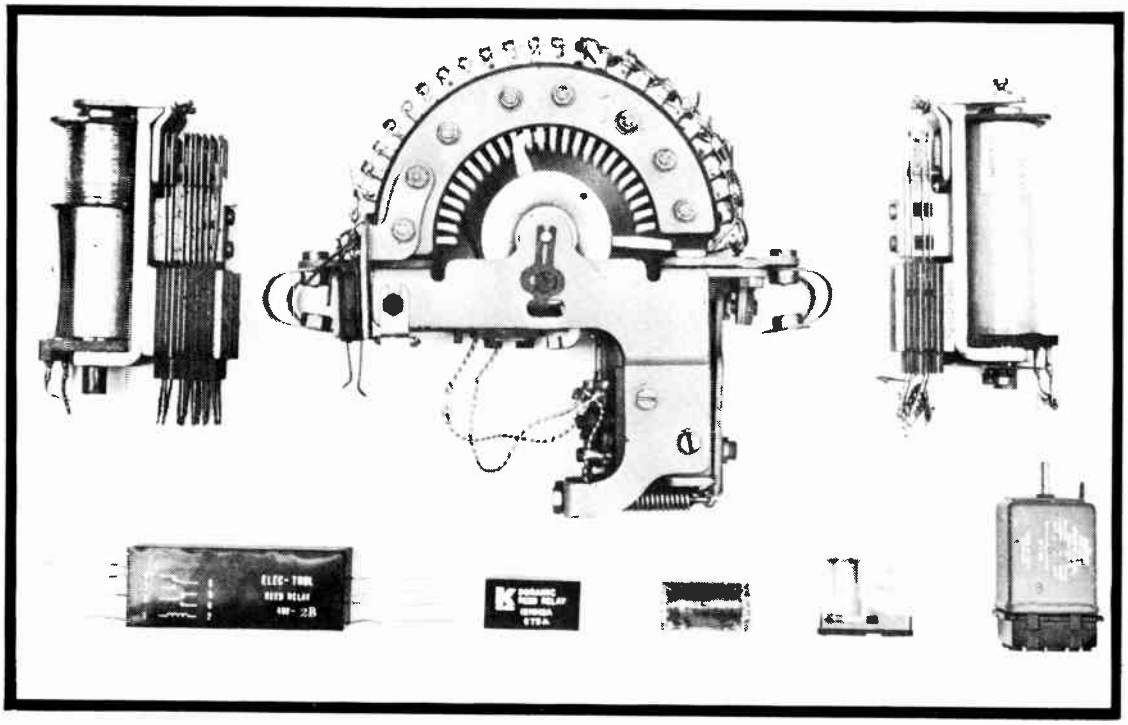


Fig. 3. Mechanical switches are designed to allow considerable lattitude in the open and closed contact positions, thus ensuring reliability. Electromechanical switches come in many shapes and sizes.



many more passive components than an analogue type of multimeter. Yet, today, there is little difference in the cost of either alternative for the same accuracy. The digital scheme, however, can be made considerably smaller, may be made more accurate, uses no more power and may even have circuitry that automatically selects the most appropriate range for itself.

Another example is found in computing. We have seen how operational amplifiers — those that perform linear arithmetic inherently — can be used to solve equations and do complicated arithmetical operations in what are called analogue computers. These can provide extremely powerful solutions of mathematical problems for quite small outlays. But only if the problem does not require high-accuracy — then digital computation is needed. Another instance where digital method is a must is when the problem involves logical type operations where yes-no decisions are needed. Digital computers can sort information into groups and decide which way to proceed at a decision junction. This will become clearer when we discuss the mathematics of logic which is quite unlike normal algebra.

As with the multimeter example, digital computers also involve many more components than the analogue units that would perform similar tasks. Yet, somewhat strangely, they can be far less expensive, much more accurate and more reliable. Undoubtedly the trend in electronic systems is toward more use of digital solutions — but this does not mean that analogue systems have no place in electronics.

One dominant reason why digital systems can be so reliable and positive to design is that the signal operations involve switching rather than continuous-mode action. We, therefore, begin our study of digital systems by looking at the design merits of various switching devices, starting with the mechanical kind.

MECHANICAL SWITCHES

The ON-OFF switch has only to define two states of circuit operation and hence the tolerances associated with each state can be very wide. Consider the basic mechanical switch having two contacts as shown in Fig. 3. When the contacts are disengaged it matters little how much further the designer separates them; the further they separate the less the chance of a spurious make-condition occurring. Conversely, when closed the spring action will ensure contact over a wide range of relative positions. The harder the two contacts are pressed together the better the reliability, but there will be negligible electrical change in the circuit-made state.

Continuing with the mechanical switch example we can also easily see that a switch with heavily over-travelled contact pressure or excessive opened distances will be slow

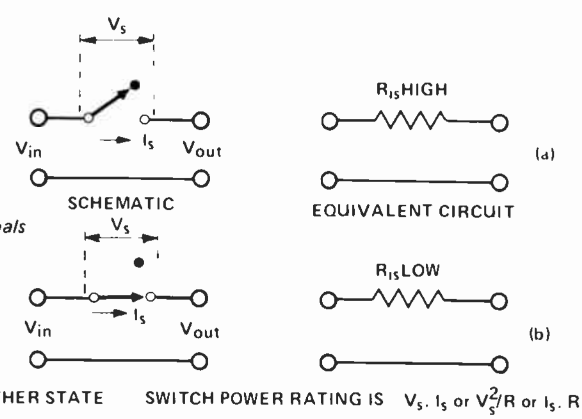


Fig. 4. Switches can handle large power signals with little loss (a) open (b) closed.

ELECTRONICS - it's easy!

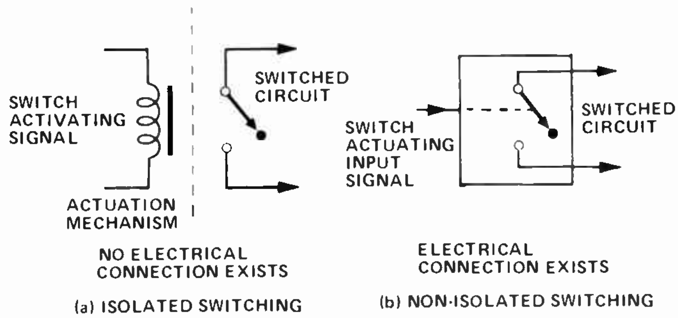


Fig. 5. Unlike electro-mechanical switches, most solid-state electronic switches do not provide ideal isolation between the actuating signal source and the controlled source.

to change to the opposite mode because greater force, or more travel, is needed to effect the change.

Another feature of the two-state switching circuit is that the switch's power rating can be very small compared with the load power being controlled. This arises because in each of the two states the switch has only to dissipate very small power losses. When open, see Fig. 4(a), the voltage across the switch is maximum but the current minimum. The power rating needed of the switch (neglecting arcing effects in this case) whilst open is, therefore, the product Vs.Is, and this is always very small, for only leakage currents flow when the switch is open. When closed, the situation is reversed; the current is now of the maximum value but the voltage drop is merely that due to resistive losses in the made contact (which can be very small). In practice the change of state from one condition to the other is so rapid that we can consider the switch as only ever being in the fully-off or fully-on case. This low-loss feature is used to effect in power-supply switching regulators where the "made" to "not-made" times of a vibrating contact are varied to pass the required amount of average power.

SOLID-STATE SWITCHES

Originally digital circuits did indeed use electro-mechanical switches; the relay as we know it. These are still used in some circumstances today but their size, cost, slowness of switching and possible unreliability now make them a poor choice, for logic applications, compared with solid-state switching alternatives.

A switch by definition, is a device that provides either a satisfactorily high or low resistance between two points, with the state being rapidly reversed by an external control input. It can be used in series or shunt to effect control. The degree of isolation provided is decided by the open-state resistance; the power rating is decided by the made-state switch resistance. What is high or low is purely relative, depending upon the impedances of the circuit elements connected to the switch. A perfect switch provides infinite open-circuit resistance and zero closed-circuit resistance. Typical resistances encountered in a small relay are from many megohms (contacts open) down to mere milliohms (contacts closed) thus giving excellent switching characteristics. Solid-state switches normally do not provide such large resistance ratios (some special devices come close) giving around a megohm to a hundred ohms change which is adequate for most logical tasks performed by digital systems.

Another disadvantage of most solid-state switches is that, as we will see below, the circuit connected to the switching part of the solid-state switch is not completely isolated from the circuit actuating the switch mechanism. This concept is shown in Fig. 5. At times this is most inconvenient and the development of workable solid-state switching systems has been influenced by the need to make-do with this shortcoming. A relatively recent newcomer to the solid-state switch, which overcomes the last disadvantage, is the solid-state opto-electronic isolator, shown in Fig. 6. This uses the actuating input to energise a solid-state light-emitting diode (LED); this, in turn, reduces the resistance of a light sensitive detector that acts as the 'contact'. This device is used in a minority of switching operations involved in digital circuitry where extremely high isolation is required between the switching and the switched circuits.

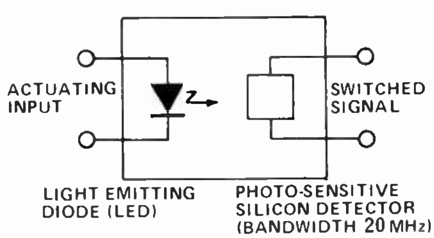


Fig. 6. Opto-electronic switches such as HP5082 series can provide a very close approximation to the low-power mechanical switch and are much faster in operation.

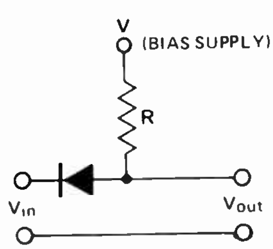


Fig. 7. Diode wired to provide switching action between input and output.

The two most commonly used solid-state switching techniques are those using two-terminal diode designs and three-terminal active element designs based on devices such as the transistor and other solid-state amplifying devices.

Let us first look at a diode wired to provide a switching function. In Fig. 7 a diode is connected to a bias supply V and to the input as shown. When the input voltage Vin is more negative that the bias voltage V the diode is forward-biased providing a quite low resistance path between the input and the output terminals. In this state Vout will be closely equal to Vin. If the bias voltage (or the input voltage) are changed to make Vin more positive than V the diode becomes reverse-biased placing a high-resistance between input and output. Thus, by changing V from positive to negative we have produced a switching action between input and output terminals.

A similar action is provided if the diode is wired in shunt across the line rather than in series as shown in Fig. 8. The state of V decides whether the diode shunts the line (when forward biased) or not (when reverse-biased).

In either design it is important that the diode resistances in the two states, the output impedance (R_S) of the preceding stage connected to the input, the load impedance (R_L) connected to the output and the

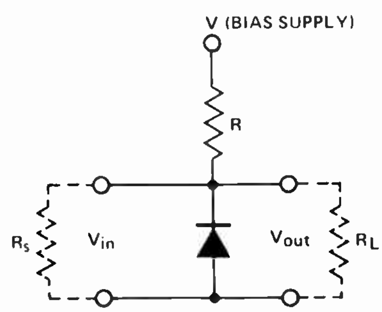


Fig. 8. Diode wired to provide shunt switching action.

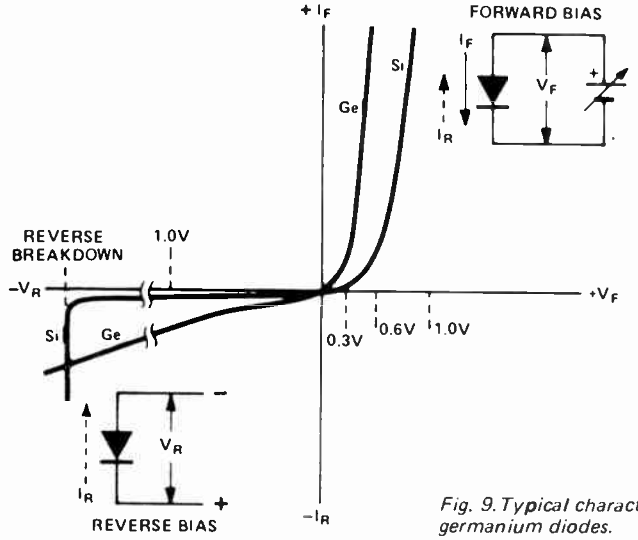


Fig. 9. Typical characteristics of silicon and germanium diodes.

bias-resistor value (R) are all chosen to have the right ratios in the two switching states. Adequate switching action will not result if the bias input is allowed to dominate the values being switched.

It is important that the bias voltage actually developed across the diode is sufficient to produce a diode forward-current greater than a value around the knee of the I_F against V_F characteristic — see Fig. 9. For a silicon diode this requires at least 700mV, a germanium diode at least 350mV: these values vary little with make or shape of particular device, being a parameter of the semiconductor material itself. Note how a quite large change in forward current hardly changes the dynamic resistance once the knee is passed. (Dynamic resistance is the slope of the characteristic which is reasonably constant beyond the knee). This reliable and constant loss switching (but not zero-loss) results over a very wide range of bias current conditions.

In its reverse-biased state the diode provides a larger resistance. Fig. 9 shows that germanium diodes do not provide as high an 'open' resistance as do silicon diodes — this is because the slope of the germanium characteristic is not as horizontal as that of silicon. Nevertheless both slopes represent higher resistance than in the forward-biased case, proving that resistance of the diode changes markedly. Again, we see that both reverse-biased curves are closely linear meaning constant resistance or, in other words, constant "open circuit" switch resistance.

When selecting the value of switching bias to apply it can be seen from Fig. 9 that too high a value for silicon devices will cause breakdown at the zener point, providing instead, a made-state that could cause total failure of the device.

The speed at which diodes can switch is a function of circuit values and the characteristics of the diode. It is routine practice with diodes to switch at tens of megahertz rates or higher. Mechanical switches are limited to less than 1 kHz at the very best.

Later in the course we will see how these basic diode switches can be used to perform logical operations by connecting more than one diode to the same bias source. Such connections are called gates.

Now to the use of three-terminal devices, transistors for instance, as switches rather than as linear amplifiers. This can be explained using the I_C versus E_C characteristics of a typical transistor, as is given in Fig. 10. The two switching states occur when I_b is either large or small. A chosen collector resistance value (in common emitter configuration) establishes the load-line on the characteristic. In a switching-mode the transistor operates around points A or B. At A, I_b is large; the transistor is, therefore, switched on with V_{CE} being very close to zero volts. At B I_b is small (practical circuits may apply a reverse polarity to ensure this); the transistor is switched off with V_{CE} being virtually at the supply voltage. In the on-stage the transistor provides a low-resistance path between its collector and emitter: when off, a high-resistance path.

The transistor switch, unlike non-amplifying diodes and mechanical contacts, does not directly pass the input signal but instead replicates a signal current in its base by providing an equivalent change in collector current or voltage. In reality a large proportion of digital circuits regenerate in this way with the output signal change closely following that of the input.

At either of the circuit operating points A or B the transistor is operated well within its allowable power dissipation. As we should expect, a given transistor used in a switching mode can handle a greater power than if operated as a linear amplifier. A little thought will also show that the load line can, in switching use, intercept the maximum dissipation curve, the reason being that the transistor does not dwell long enough in states other than A or B to produce deleterious heating. It is vital, however, in such designs to ensure that the switching action is rapid between states, and that the device never dwells on the way through. A ramp input signal may well destroy a stage designed to switch!

The above explanation is most basic — reality requires other criteria to be recognised to obtain more ideal switching. Like the over-travelled mechanical switch, a transistor switch with too much reverse-bias base current (off-state) or too much on state base current will be slower to operate than one not driven so hard. This is because the charge associated with the base current must be removed to alter the state and the more the charge there is to move, the slower will

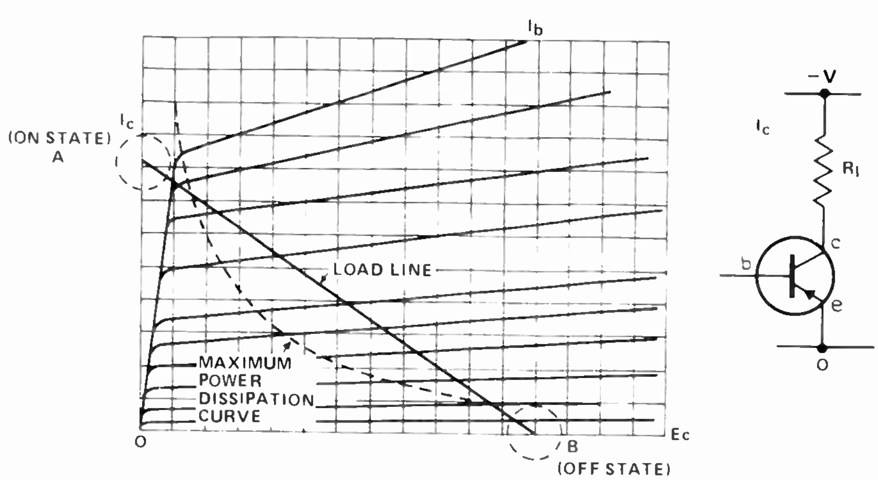


Fig. 10. In the switching mode the transistor is operated at either end of the chosen load line.

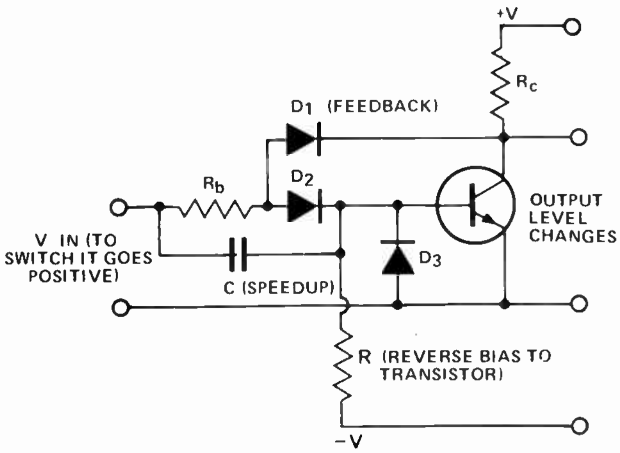


Fig. 11. High-speed switches are designed to remain non-saturated. This circuit employs feedback D1 with D2 providing a voltage supply needed. The speed-up capacitor is C. Diode D3 assists reduce the delay time.

be the switching time. Solid-state switches operated very positively by use of large drive currents are said to be working in a saturated state.

Certain circuit devices can be added to the basic solid-state switch to speed up the response. The first is to supply a much larger input signal than would be needed to just turn it on. This speeds up the charge movement but would take the device into deep saturation unless clamps are added that hold the circuit nodes at given values. Diodes acting as switches are often used to hold a point at a given voltage. A second circuit addition is the speed-up capacitor. This is a small value capacitor placed across the resistor feeding the input of the switching stage. When fast switching signals occur the capacitor provides a low impedance path around the resistor which must be of a reasonably high value to supply correct dc signal level requirements. Yet another technique is to use feedback between the collector and the base to speed up the switching transition yet hold the stage in a non-saturated state once switched. Fig. 11 is a non-saturating switch circuit — one of many possibilities. It shows how the basic transistor needs the addition of more components to realise fast switching in discrete designs.

The integrated-circuit revolution has provided us with inexpensive, ready-made digital circuits of great sophistication. These are extremely basic yet super fast — see Fig. 12. Rarely does one now have to consider the in-depth design of switching circuits. The task is usually one of devising a system using a few basic, digital system building blocks which have been so developed as to facilitate their ease of connection into systems.

The reliability of the switching state of an electronic circuit is one reason for the widespread use of digital techniques. There is another equally important reason for the use of digital signals and that is that philosophers and mathematicians of the past have developed powerful ways to process logical information by way of special algebra and techniques. This is employed to design complicated switching circuits and other digital systems with the simplest possible circuitry. In the next part we look at these philosophical concepts in readiness to return to a discussion of the basic, digital-system building blocks.

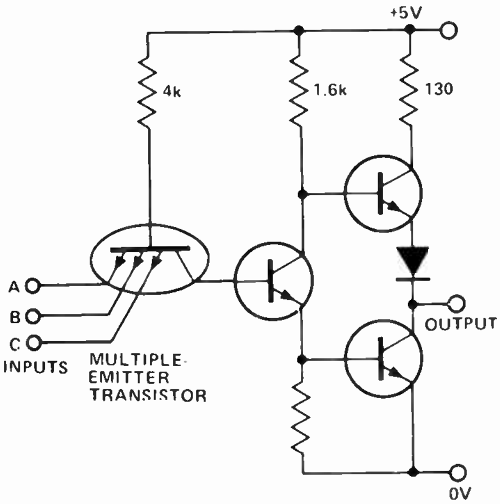


Fig. 12. Transistor-transistor logic, TTL is a commonly accepted integrated circuit manufacturing method. This gate combines up to three inputs switching with a delay of only 10 ns. The whole circuit is integrated on a common chip.

Notes

22 The algebra of logic

MATHEMATICS is a kind of shorthand language which enables us to present a physical process, on paper, with symbols which may be manipulated in order to gain a better understanding of the process. It is thus a tool which aids understanding.

The familiar kind of algebra which relates two variables, x and y, in combinations such as x+y, x−y, x.y, x/y, x^y and others is a linear process because the two variables can hold any value. It is this kind of algebra that is performed by analogue operational amplifiers.

However, if x and y can only have two possible states, such as a voltage which is there or not there, we can ignore the actual value of the voltage (or whatever) and regard the variables as behaving according to a two-state or binary number system. Just what the two states are is of no importance whatsoever — they can be high or low, positive or negative, there or not there and even true or false.

A mathematical algebra has been developed to cope with such binary systems. It is known as Boolean algebra — the algebra of logic, and it's rules are somewhat different to those of linear algebra. Before delving into the operation of Boolean algebra, it is worth tracing it's historical development.

HISTORY OF SWITCHING MATHEMATICS

Philosophers, those people who apply special skills to resolving paradoxes by the use of logic, have existed since the earliest civilisations. The Ancient Greeks were so impressed with logic that they wrote plays around Aristotle's formally arranged rules of logical deduction. The rules for this process of reasoning were handed down, largely by word of mouth, through the Dark Ages, with little, if any, recognition of their value for logic in computation. It was not until the early 19th century that the use of logical rules in calculation was established. This work was very much the result of George Boole's 1854 work (see Fig. 1) entitled "An Investigation of the Laws of Thought on which are Founded the Mathematical Theories of Logic and Probabilities", Augustus de Morgan, a contemporary, also contributed to the first systematic arrangement of Aristotle's logic.

Boole took the concepts further than the Ancients by substituting mathematical symbols in place of the basic logical situations. This symbolic logic became known as Boolean algebra.

Little was achieved with Boole's work for the next few decades. The first machine to utilize his algebra to solve logic problems, faster than by hand, was William Jevons' logical piano of 1869. Boole's contribution, however, had to wait until the early 20th Century to find extensive application. One by one, logicians advanced the techniques of logical algebra: Pierce, Venn, Dodgson, Marquand, Pastore, Bollee. The "Principia Mathematica" of Whitehead and Russell (1910-1913) and the Hilbert and Ackermann work "Mathematical Logic" (1928) were further milestones in digital computer realisation.

Fig. 1. In 1854 George Boole, an English logician, showed how ordinary algebra could be applied to logic situations.

Shannon's 1938 paper "A Symbolic Analysis of Relay and Switching Circuits" was a paper of very practical relevance for it described how to put Boole's rather abstract logical albegra to work in engineering and computer design. But this was not the first recorded use of electrical logic circuits. In a letter Charles S. Peirce wrote to his former student, Marquand, around 1890 he expressed, in the words and circuit diagrams shown in Fig. 2, that logical algebra could be performed with three switches in parallel or in series, also stating that he felt electricity to be one of the best ways to implement logical equipment.

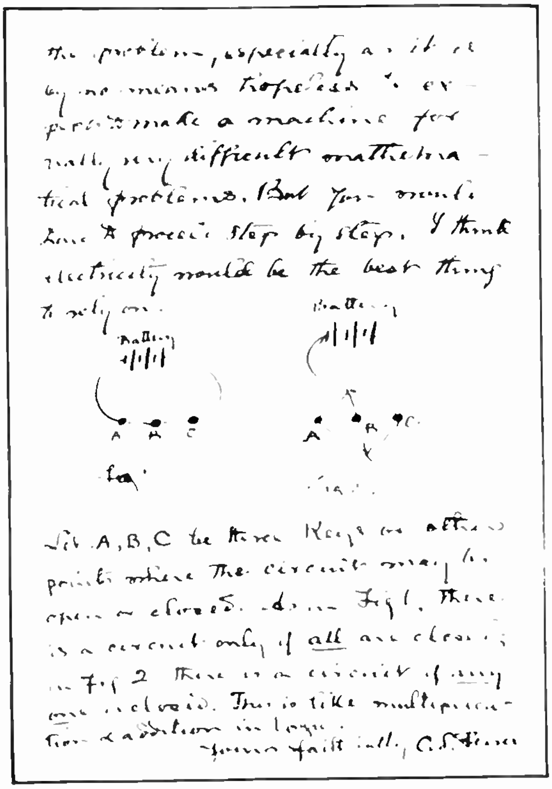
Later theoretical studies concentrated on ways to ensure that switching networks contained no more switch contacts than were absolutely necessary. Unnecessary contacts can easily be unwittingly designed into complex switching networks — the "spares" are called redundant switches. Shannon, in his M.Sc. thesis (Fig. 3) prepared at the famous Massachusetts Institute of Technology, realised ways to systematically set about analysing a given switching network in order to reduce the contact requirements to a minimum. Thus it was realized in the early 1940's that really powerful digital computers could be built using entirely electronic components.

Later in the course we will be dealing specifically with computer systems. They are, however, but a part of the total use of digital electronic methods — digital electronics finds use in an ever increasing number of instruments and devices.

BASIC LOGIC GATES

A quite satisfactory way to begin to comprehend basic switching algebra is to think in terms of mechanical switch contacts arranged in various different configurations. That we draw them and consider them as mechanical contacts that are either open or closed, does not imply that the contacts necessarily need to be mechanical — they are, today, more often than not the solid-state switches we discussed in the last part.

Groups of switches combining


the problem, especially as it is by no means hopeless to expect to make a machine for really very difficult mathematical problems. But you would have to proceed step by step. I think electricity would be the best thing to rely on.

Let A, B, C be three keys or other points where the circuit may be open or closed. As in Fig 1, there is a circuit only if all are closed; in Fig 2 there is a circuit if any one is closed. This is like multiplication & addition in Logic.

Yours faithfully, C.S. Peirce

Fig.2. The first known description of electrical switching carrying out logic is in this letter of C.S. Peirce written around 1880.

Fig.3. Claude Shannon published details of "modern" digital computing design in 1937.

digital signal levels are known as gates. We begin by considering the simplest possibilities where there are just two contacts to build with.

They can be placed in series or in parallel, as shown in Fig.4. In each case different conditions exist between the transmission made through them for the two positions of each of the switches. We denote the switch inputs as A and as B (and C,D, etc., if more are involved) and the transmission as Z, thus using mathematical symbols to represent a physical situation. Imagine that the switches are wired in series with a lamp: when a circuit is made the lamp lights.

In the series case we need switch A and switch B to be made to obtain a transmission function Z. In the parallel case either switch A or switch B will provide transmission.

The AND and OR are basic logical functions. They need not necessarily be used only to describe electrical circuitry. They did, in fact, as we have seen, arise originally from philosophical study of truths and falsities.

Note that switch contacts are always shown in their non-actuated condition and this brings us to another basic gate function which can be realised using only one switch. If, as shown in Fig. 5, the switch A is actuated, Z is NOT enabled. If A is not actuated Z is enabled. A single switch, therefore, can provide a NOT function if its contacts are closed in the non-actuated state.

Attempts to explain switching circuit action in words, as above, only applies for the simplest of situations. The descriptive method becomes prohibitive when, say, we have two switches in series, in series with two switches in parallel, as shown in Fig. 6. Describing the action of all possible switch combinations on the lamp Z using words, is an inadequate way with which to communicate the idea. And few digital systems are that easy: many contain literally thousands of AND, OR and NOT gates.

We designate an OR function by means of the '+' symbol. This does *not* mean the same as our normal understanding of addition. When applied to decimal numbers it means addition as we normally understand it. With binary numbers, however, it has a different meaning and still another meaning when designating an OR function. For example:-

In decimal addition	1+1=2
binary addition	1+1=10
OR addition	1+1=1

In Boolean algebra the OR meaning of addition is the one that applies. Thus A+B=Z means that A OR B switch closed will produce a transmission Z.

We designate an AND function with a dot. The dot means logical multiplication and is not to be confused with normal multiplication. However the truth tables for AND multiplication and normal multiplication are the same. Thus when we give the Boolean equation A.B=Z we mean that if switch A and switch B are both closed there will be a transmission Z.

The NOT function is designated as a line over the switches algebraic symbol giving Z = $\bar{A}$ to mean Z is NOT transmitted when A is actuated.

Each of these functions have a

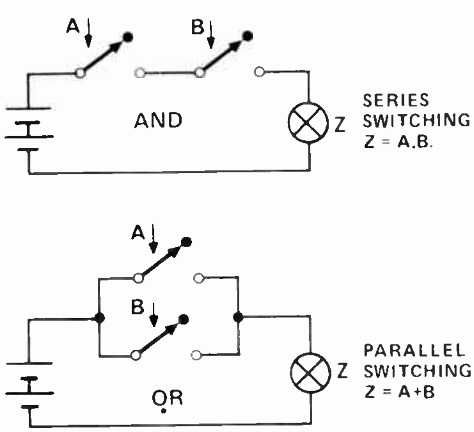


Fig.4. Two contacts wired in series or in parallel provide the basic logic functions of AND and OR.

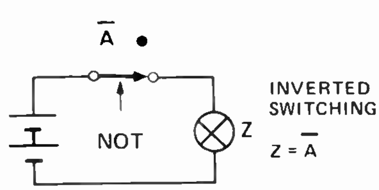


Fig.5. The NOT function is obtained by reversing the state of the switch operation.

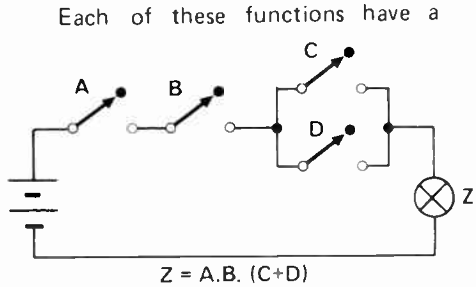


Fig.6. More complex switching functions are best described in terms of logic algebra than by words.

ELECTRONICS -it's easy!

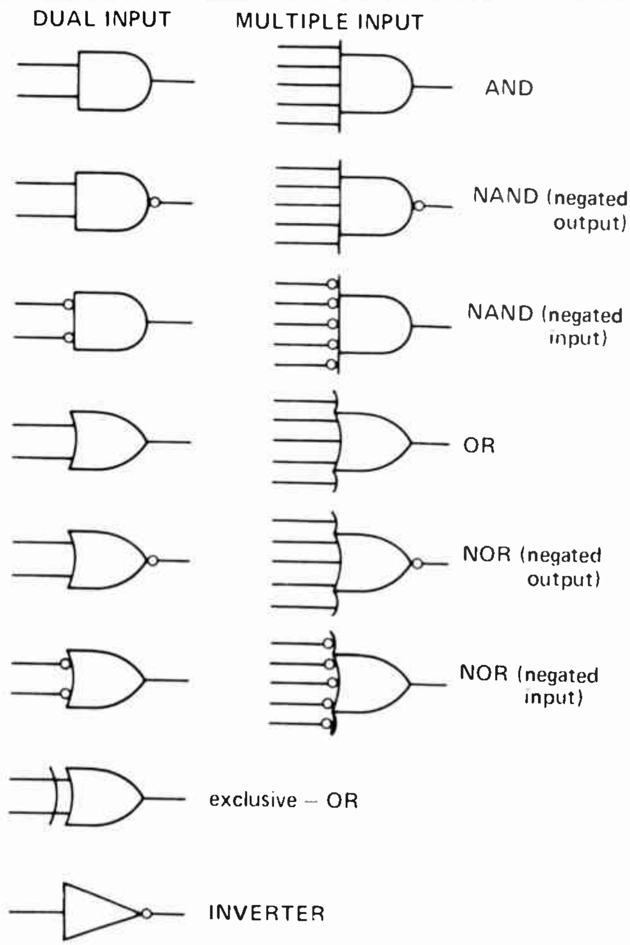


Fig. 7. Two-state logic gate symbols used in this course.

symbolic representation as black-boxes with inputs that act in certain ways to give the output. The shape of the box (or the designation within a square box) tells the viewer the function of the box.

Unfortunately there still exists more than one conventional way to draw these symbols. For this course we will use those given in Fig. 7, which are also those used in projects in Electronics Today.

The NOT function bar can be applied to any function to signify that it is negated. For instance, an OR such as A+B=Z becomes $\overline{A+B}$ which is called a NOR function. Similarly so $\overline{A.B}$ is a NAND function.

The OR, AND, NOR and NAND functions can each have more than two inputs, for example, A+B+C+D=Z. When a function is negated its graphical symbol is also altered in some way to signify this. The convention used is the convention of the addition of a small round circle. If the circle is at the output the output is negated; if at the input the inputs are negated. The inverter (that provides negation) is basically an amplifier providing 180° phase shift so its symbol is that of an amplifier with the circle added.

TRUTH TABLES

Before we discuss more complex gate networks by studying their inter-connection, we need to understand the concept of a truth table. This is a simply drawn table that lists the output state for the various combinations of input states.

Rather than write on and off, or high and low, true or false, it is simpler to express the two states merely as '0' and '1'. The *positive* logic convention considers a high-voltage level as a '1' and the low level as a '0'. Fortunately, today, just about all logic circuits used are now in integrated circuit form and they nearly all work between just two levels – which are the same for any devices from a particular logic family. This provides a compatible arrangement whereby gates and other logic system boxes (that are yet to be introduced) can each be intercoupled without having to worry about matching voltage and impedance levels. However, when transferring logic signals between devices from different logic families translator circuits will be needed to make voltage levels compatible.

Occasionally, but not commonly, it is more convenient to reverse the levels calling a 1 the lower voltage and 0 the higher. This is denoted *negative* logic. Such a system is however seldom used in modern integrated-circuit logic families.

Consider then the series contacts of Fig. 4. Assuming we use the positive logic convention where 0 represents an open contact and 1 a closed contact; it is easy to draw up columns as given in Fig. 8.

When A and B are both 0 then so also is Z, for no contacts are made. Similarly, if either A or B are open. When both A and B are closed, that is a 1 each, then Z is made. This is called a truth table.

Fig. 9 is the truth table for the parallel contacts of Fig. 4. In this case Z is 1 when A or B are 1.

An interesting property of the AND and OR functions is their dual nature when negated. For example, if we negate the inputs of the OR gate the truth table becomes that of Fig. 10,

INPUTS		OUTPUTS
A	B	Z
0	0	0
0	1	0
1	0	0
1	1	1

AND
Z = A.B

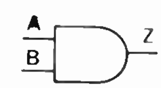


Fig.8. Truth table for A.B.=Z, the AND function.

A	B	Z
0	0	0
0	1	1
1	0	1
1	1	1

OR
Z = A + B

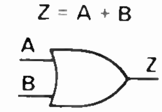


Fig.9. Truth table for A+B=Z, the OR function.

POSITIVE LOGIC		NEGATIVE LOGIC		NAND	AND (IN POSITIVE LOGIC)
A	B	$\bar{A}$	$\bar{B}$	Z	$\bar{Z}$
0	0	1	1	1	0
0	1	1	0	1	0
1	0	0	1	1	0
1	1	0	0	0	1

Fig.10. Truth table showing that negative logic (or negated positive logic) input to an OR gate provides NAND output.

the output of which is the NAND function. Hence a negated input OR gate is a NAND gate. Also, by similar reasoning, a negated input AND is a NOR. Put another way, in negative logic an OR becomes an AND and vice versa.

UNIVERSAL GATES

Using the basic gates, AND, OR and NOT, we can build a logic circuit for any given Boolean expression. Where there is a plus sign (+) we use an OR gate, where there is a dot we use an AND gate and we use an NOT gate for those functions that are negated.

However it is interesting that the NAND gate can be used to obtain any desired function. It can be used to build AND, OR or NOT gates. In other words it is a universal building block, as is the NOR gate also.

Thus the majority of gates used in modern logic systems are NAND gates with the occasional use being made of NOR gates and inverters (NOT) to minimize complexity. The use of one major form of gate simplifies manufacture and reduces costs.

FAN OUT

There exists a finite number of circuits that can be safely connected to the input, or the output, of logic elements. This number is called the fan-in and fan-out respectively, and gives the number of standard loads that can be accommodated. Fanouts of 10 and 30 are typical load factors.

EXCLUSIVE OR

One other important gate is a special class of the OR – the exclusive OR. The logic action of this gate is seen by studying its truth table which

A	B	Z
0	0	0
0	1	1
1	0	1
1	1	0

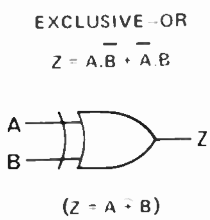


Fig.11. Truth table for two input exclusive OR gate.

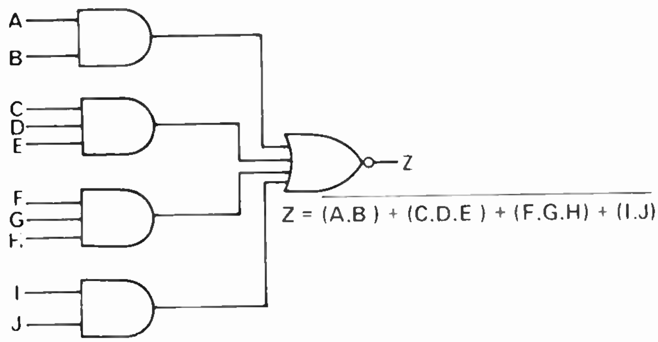


Fig.13. (a) Logic network for function $Z=\overline{(A.B) + (C.D.E.) + (F.G.H.) + (I.J.)}$ (b) same logic packaged in IC flatpack.

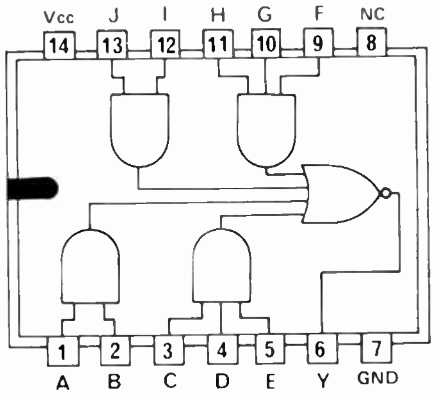


Fig 13(b).

is given in Fig. 11. In this variation of the basic OR gate the output is 1 for either A or B but not when both are 1 simultaneously. Written in Boolean algebra symbols this gate performs $A.\bar{B}+\bar{A}.B$ Z. (Symbols written as AB imply that a dot exists between them; it is common practice to omit the AND dot).

MORE COMPLEX LOGIC

The exclusive OR gate is more complex than the other gates discussed above because it contains more than one basic gate – it is a small logic system in itself. Fig. 12 shows how two inverters, two AND gates and one OR gate can be interconnected to achieve the exclusive OR requirement.

A second example is given by considering a function

$$Z = \overline{(A.B)+(C.D.E)+(F.G.H.)+(I.J.)}$$

The problem might be to realise a logic network that performs this logical task – imagine trying to describe it in words! Brackets are used to ensure that sub-connections are made in the correct way; as in linear algebra operations in brackets are dealt with first as individual units.

The first step in realising the network is to form the dot AND functions of Z. We need two two-input AND gates and two three-input AND gates. (It matters not if a gate has more inputs than needed – the unused terminal is ignored). The outputs of these four AND gates are then fed into the inputs of a four input OR gate so that the function under the negation bar is achieved. At this point we could select an OR gate followed by an INVERTER or make use of a NOR gate direct.

When drawn as a system of interconnected schematic blocks it appears as in Fig. 13a. Also given in Fig. 13b is how a 14 pin dual-in line IC would appear that performs this function.

As a third example the exercise is to devise a logic network that will add (in binary system) two binary inputs producing the binary sum output plus a carry output. This function, called the half-adder, forms the basis of digital computation with binary numbers.

Back in Part 5 the concept of the binary number system was introduced showing that the counting base is 2 instead of the more commonly encountered 10 of the decimal system. At any digit position in the binary number, the value can be only 0 or 1 so addition of two binary numbers gives a value at each digit position that alternates as 0 1 0 1, etc., as counting progresses. When 0 and 0 are added we obtain 0; when 0 and 1 are added we get 1. When 1 and 1 are added we cannot have 2 in a binary system so it returns to 0 with a carry of 1 going to the next higher digit position. Fig. 14 illustrates this idea – try adding the two numbers! A half-adder does this operation for one digit position. The truth table for the half-adder is, therefore, as given in Fig 15a.

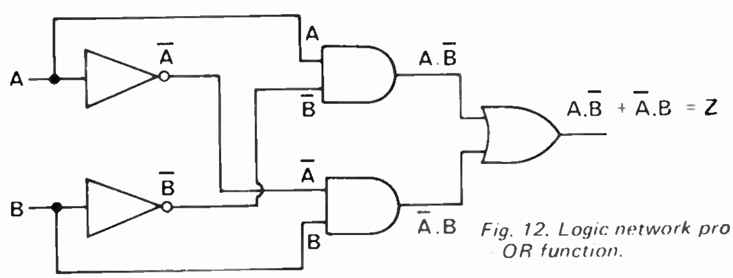


Fig. 12. Logic network providing exclusive OR function.

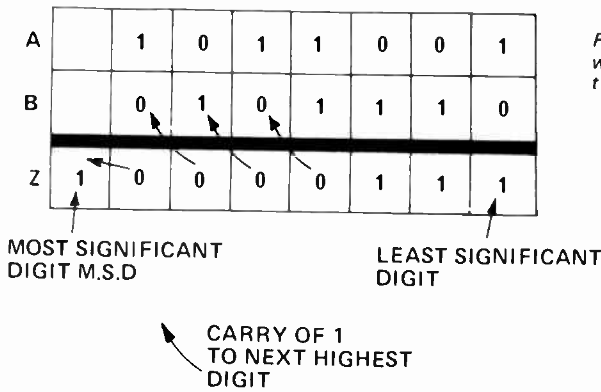


Fig.14. Addition of two binary numbers proceeds with a carry as for decimal arithmetic but with only two states 0, 1 in each digit.

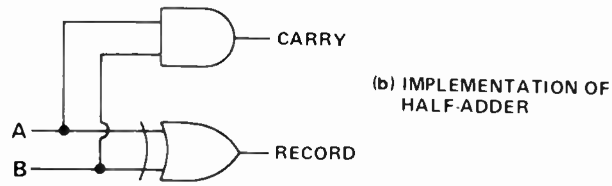


Fig.15 (b). One form of half-adder logic network.

INPUTS		OUTPUTS	
A	B	CARRY	RECORD
0	0	0	0
0	1	0	1
1	0	0	1
1	1	1	0

(a) TRUTH TABLE OF HALF-ADDER

AND CONDITION   EXCLUSIVE-OR-CONDITION

Fig.15. (a) Truth table for half-adder logic.

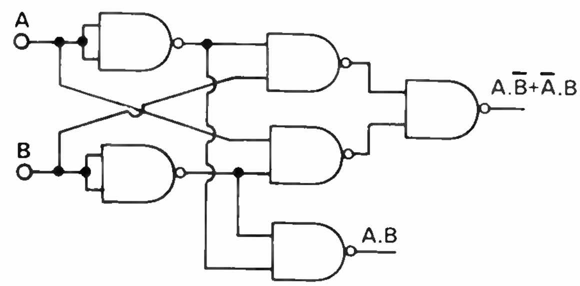


Fig.15 (c). The half adder using NAND gates only.

The sum column shows we need an exclusive -OR to provide the sum value – hence its importance in computer design. A carry is to occur when both A and B appear so an AND gate is needed. From these we can develop one form of the half-adder system – given in Fig. 15b. Note how the complexity is growing. Such a circuit requires around 30 or more passive and active components and hundreds of such circuits are needed in a digital computing circuit. A version of the same circuit only constructed using NAND gates is given in Fig. 15c. Note that NAND gates 1 and 2 have both inputs tied together, they therefore perform the NOT function. Try your Boolean on this as follows – .

SOME LAWS OF BOOLEAN ALGEBRA

When devising systems of logic the situation soon arises which calls for knowledge of the rules for manipulating Boolean expressions. Possible reasons for this may be that a limited range of logic functions are available, so conversion of an expression is needed, or that a large expression may not be in its simplest state. Reduction to its non-redundant state means use of less elements.

A number of axioms (truths based on experience) exist for relationships between Boolean statements. There is little point in dwelling on their individual proofs and historical development – for that see the reading list. The following relationships are summarized to assist when needed:

de Morgan's rule 1 : $\overline{A+B} = \bar{A}.\bar{B}$

de Morgan's rule 2 : $\overline{A.B} = \bar{A}+\bar{B}$

Commutative laws : A + B = B + A
A.B = B.A

Associative laws : A.(B.C) = (A.B).C = A.B.C
A + (B + C) = (A + B) + C = A + B + C

Distributive laws : A.(B + C) = A.B + A.C
A.C + A.D + B.C + B.D = (A + B) . (C + D)

This is as for linear algebra but with extra cases:–

A + B.C = (A + B) . (A + C)
and (A + B) . (A + C) . (A + D) = A + B.C.D

Absorption laws : A + (A.B) = A
A. (A + B) = A

Double negation : not $\bar{A}$ = A

Universe class laws : A + 1 = 1
A.1 = A

Null class laws : A + O = A
A.O = O

Complementation laws : $A + \bar{A} = 1$
$A.\bar{A} = O$

Tautology laws : A + A = A
A. A = A

Expansion laws : $(A + B) . (A + \bar{B}) = A$
$(A.B) + (A.\bar{B}) = A$

MINIMIZATION

To save components the network first realised by inspection from a valid truth table may well not be in its simplest or so-called minimal form. In simpler cases, application of the above Boolean algebra laws by a well-practiced person can often come up with simplifications.

Beware, however, of applying linear algebra rules of factoring. It is quite wrong to cancel or subtract equal terms in both sides of a Boolean equation.

Unfortunately, no direct way is known with which to arrive at a minimal network by a routinely declared simple procedure. The nearest we can get to this is by means of a Karnaugh mapping procedure which we do not discuss in this course as few readers will be required to be expert in this facet of digital electronics.

An example will show how a simple system can be minimized by inspection. Consider the expression Z = (A + B) . (A + C) . (A + D). This is readily seen to be the logic network given in Fig. 16a. From the distributive laws given above this can be rewritten as Z = A + B.C.D which represents the logic configuration of Fig. 16b. This minimal form requires two less gates (provided a three input AND gate is available).

THE VENN DIAGRAM

In the early days of logical algebra development, John Venn developed a system of overlapping circle diagrams as an alternative way with which to express the concepts contained in the truth table. Venn's diagrams consist of overlapping circles contained in a rectangular box. Each circle represents one of the required number of independent input variables – A, B, C, etc. If the output variable Z is a 1 (assuming that is the convention

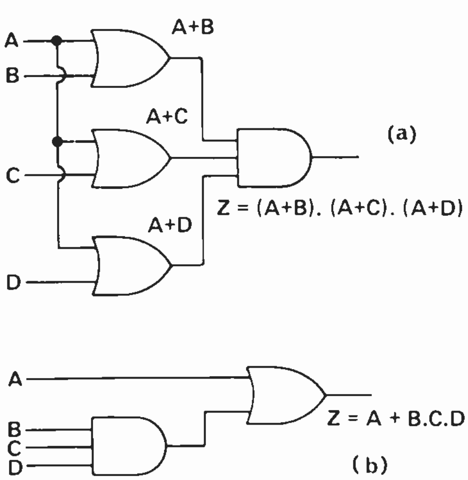


Fig. 16. Logic network realising Z = (A + B). (A + C). (A + D)
(b) Simplified network.

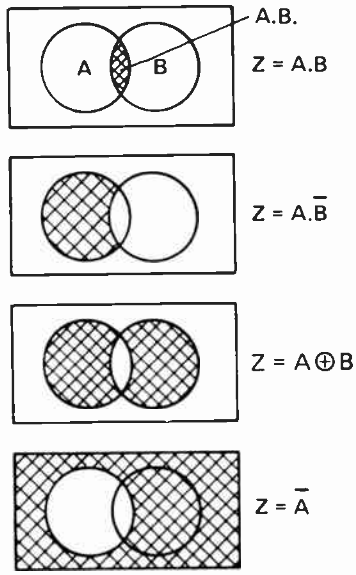


Fig. 17. Venn diagrams represent logic states in topological form. Some people find these easier to use than truth-tables.

chosen) the appropriate area of the circles is shaded. The rules are that inside a complete circle its variable is not negated, outside it is negated. Overlapping area of common circles represents their AND combination. The examples given in Fig. 17 illustrate the use of Venn diagrams in various simple logic situations. The concept extends to as many circles, that is, inputs as are needed.

LIMITS OF BOOLEAN

There are a number of limits to the use of Boolean algebra. In the logic combination we have considered so far, there has been no mention of time or of any feedback around the circuit. In practical systems, time delays always occur and, further, other elements such as counters, multivibrators and memory devices are generally present whose state depends, not only on the logical inputs at any given time but, on what has happened previously! Boolean algebra is unable to deal with such situations.

In addition, if a function is minimized by means of Boolean it does not follow that the derived circuit is the cheapest possible. The minimized circuit may call for 3-input AND gates, say, but it could well be cheaper to use the more readily available NAND gates – even if more gates are required to achieve the same function.

Thus it can be seen that Boolean algebra is far from an infallible means of arriving at the cheapest possible solution. In fact it may not give any solution at all! Engineering skill and ingenuity are still the most important factors in efficient logic design. It is of value however, and does give a good insight into the function of straightforward gate circuits.

In the next part we will look at practical circuitry of logic gates and introduce several other basic digital circuit building blocks. We will then be ready to discuss digital systems in some degree of depth.

REFERENCES

Most books on digital computer design include a chapter on Boolean algebra and binary arithmetic.

"Electronic Computers – Made Simple", H. Jacobowitz and L. Basford, W.H. Allen, London, 1967.

"Electronic Instrumentation Fundamentals" A.P. Malvino – McGraw-Hill, 1967.

"Numbers" R. Froom, Electronics Today International, July 1973, p. 84-89

For the historical development of computers and other data processing equipment see

"A Computer Perspective" C and R Eames, Harvard University Press, Massachusetts, 1973.

23 Integrated circuit forms of logic functions

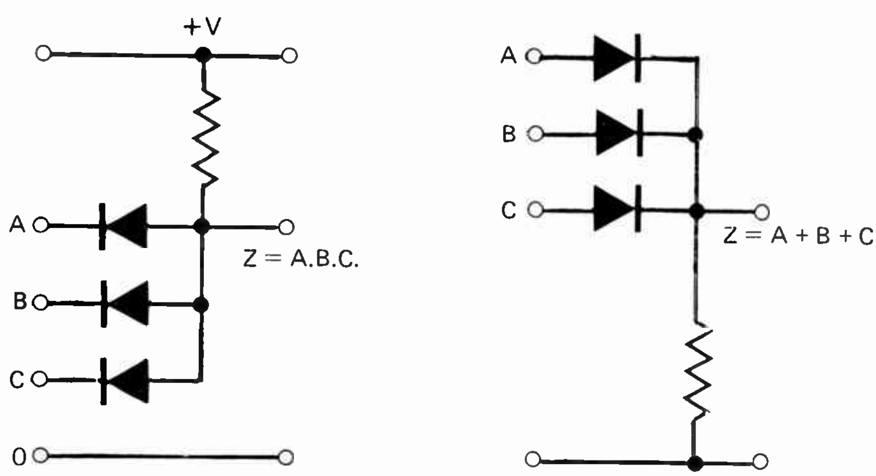


Fig. 1. The implementation of AND and OR gates using diode logic.

NOW that we have a basic understanding of switching circuits and the algebra used to mathematically describe logical operations we can look at the modern methods used in solid-state circuitry to produce the various logic functions in integrated circuit packages.

LOGIC FAMILIES
DIODE LOGIC – DL:

In part 21 it was explained how a single diode could be used as a switch by altering the bias or the input signal level. If more than one diode feeds the bias node the circuit becomes a gate. Three input AND and OR gates are shown in Fig. 1. The circuits are the same; the different class of gate arising because of different bias conditions in each case.

Diode logic is the simplest form of solid-state logic, it is not available in integrated form but is often used in discrete designs to obtain logic functions at high, or unusual (compared to standard IC logic) voltage levels.

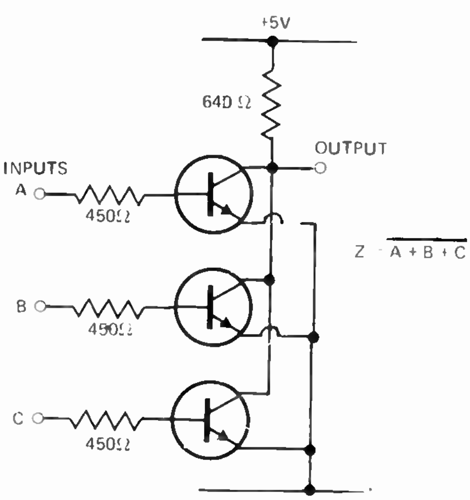


Fig. 2. Typical RTL circuit of a NOR gate: any logical '1' input will give a logical '0' output.

RESISTOR-TRANSISTOR LOGIC – RTL:

Logic gating operations can also be obtained using transistors acting as switches in various ways. Fig. 2. shows a typical RTL NOR gate for which Z = $\overline{A + B + C}$. Base current appearing at A, B or C will cause the respective transistor to switch to the ON state taking the output to ground which is a '0' in a positive logic system.

This family was the first to be used in the now more usual integrated-circuit form based on the planar manufacturing technique (one in which a mask is used to selectively diffuse impurities into a pure substrate in order to produce and separate active device junctions).

RTL is based on a supply of 3.6 V. Propogation delay is 12 nS for a medium power gate and 40 nS for a low power gate. It is a reasonably economic family to use but needs more space than the alternatives developed since. This form of logic was very much in vogue in the early 1960s but, although still manufactured by some companies for replacement purposes, is an obsolete type not used in new design.

DIODE-TRANSISTOR LOGIC – DTL:

This was the next family developed. The devices of the family use resistors, diodes and transistors. Initially DTL logic was constructed with discrete components. These designs were then integrated as shown in Fig. 3. Later devices used transistor input logic instead of diodes, thus reducing the input current requirement and allowing higher fanouts. Typical noise immunity (for a 5 V supply level – the standard used) is around 1 V. The delay time for a pulse signal to travel through, that is, the propagation delay between input change causing output change, is around 30 ns. Output is > 3.5 volts for a '1' and < 0.4 volts for a '0'.

It has a generally lower speed and lower noise immunity than other families. The advantages of DTL are the reasonably high fanout of 10 and the ease of interfacing or coupling a stage to the TTL family to be considered next.

A similar family is HTL (high threshold logic) which uses 15 V supply lines and zener diodes. This is useful in situations where high noise levels occur because this logic is more immune to noise effects than is DTL.

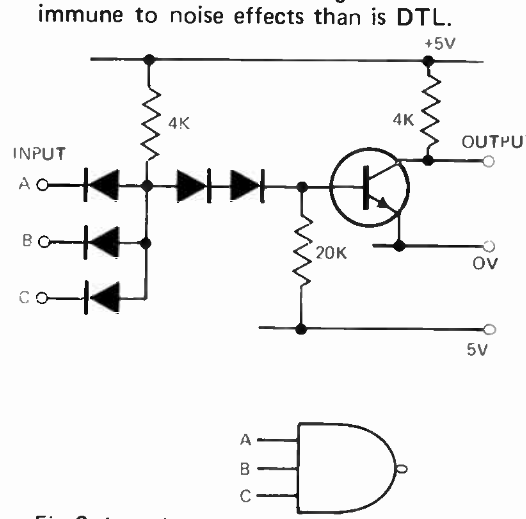


Fig. 3. An early integrated circuit design for a NAND gate in diode-transistor logic – DTL.

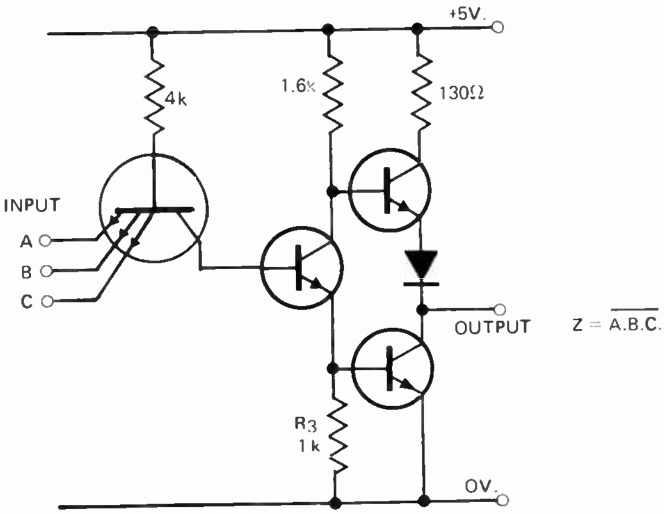


Fig. 4. In the TTL (transistor-transistor logic) form of NAND gate a multi-emitter transistor replaces the diodes of DTL.

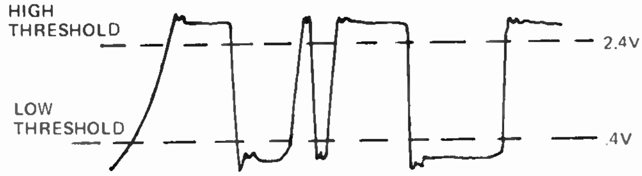


Fig. 5. Typical output signal from TTL, note the transient ringing at each transition.

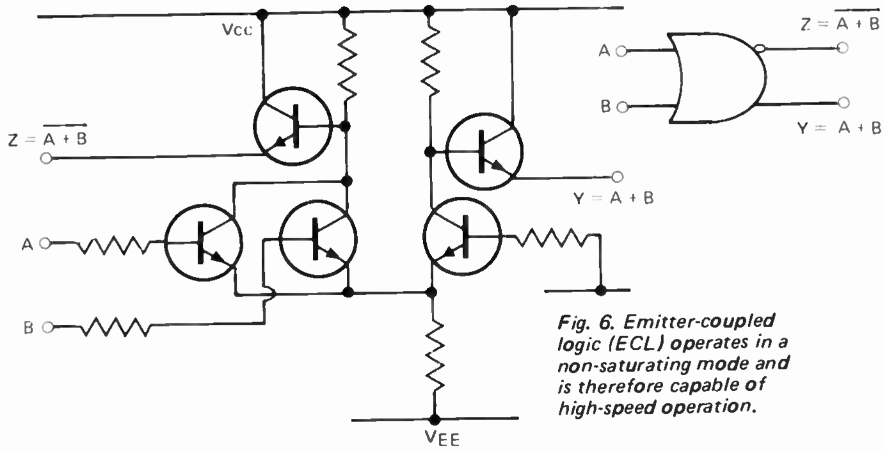


Fig. 6. Emitter-coupled logic (ECL) operates in a non-saturating mode and is therefore capable of high-speed operation.

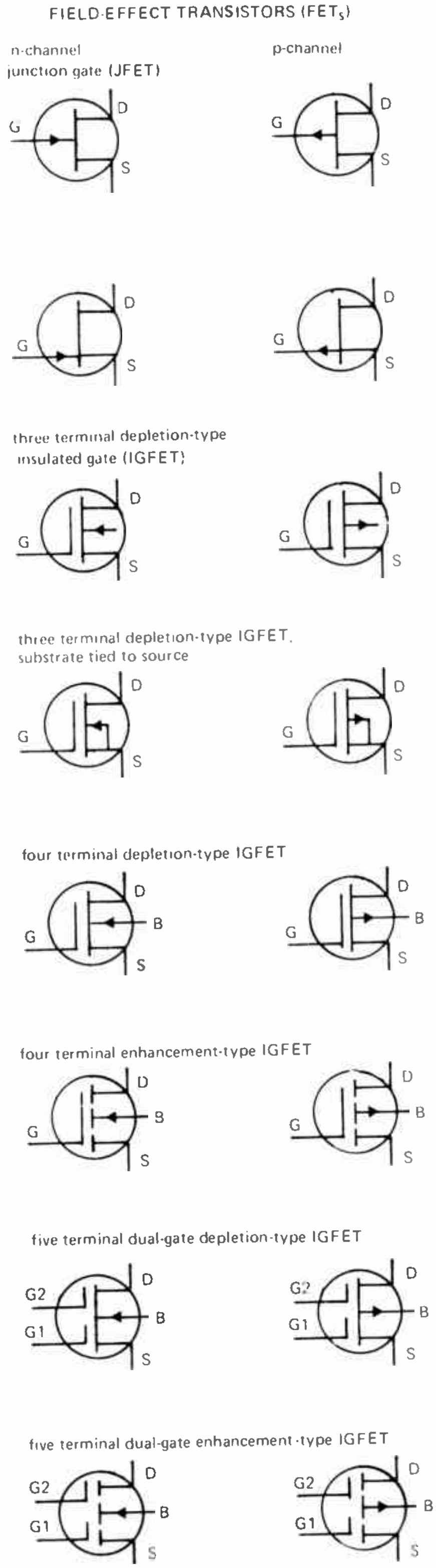


Fig. 7. The symbols used for the various types of FET device.

TRANSISTOR-TRANSISTOR LOGIC – TTL:

This is the most popular logic family in current use. It has higher speed and driving capability than DTL. The propagation delay is around 12 ns which is quite fast enough for the majority of computing applications.

A typical TTL gate circuit for Z = $\overline{A.B.C.}$ is shown in Fig. 4. Note how the diode input gate has been replaced with a multi-emitter transistor. This multi-emitter technique reduces the input capacitance thus speeding up the switching time, as well as simplifying manufacture. In TTL the supply is +5 V, the output switching between around > 2.4 V for a 1 to < 0.40 V for 0 (in positive logic). Fig. 5 shows a TTL signal switching state with time.

For all of its popularity TTL is not ideal, especially for the fastest circuit operation or where the lowest power consumption is required. Another difficulty with TTL is that switching transients occur (see Fig. 5) at the transitions. It is also not particularly suitable for large-scale integration by virtue of the relatively large amount of space and power required by each gate function.

EMITTER-COUPLED LOGIC – ECL:

A typical ECL stage is shown in Fig. 6. As this operates in the linear mode, that is, without allowing the active devices to go into saturation, it gives high-speed 2-3 ns switching times. It, however, needs a moderate power requirement, is not particularly noise immune and needs an extra power supply line. Supply voltages used for ECL vary but when typical supply rails of 0 V and –5.2 volts are used the output is –1 volt for the '1' state and –1.6 volts for the '0' state.

Each of the above logic families is based on the use of the transistor semiconductor junctions – the so-called bi-polar technique. Around 1970 ECL emerged as a possible future contender to TTL and at the same time another quite different kind of semiconductor active device became freely available – the field effect transistor FET. A variation of this is the insulated-gate field-effect transistor IGFET. Fig. 7 lists the symbols of the basic FET structures used in logic. This technique is manufactured using metal-oxide-semiconducting materials;

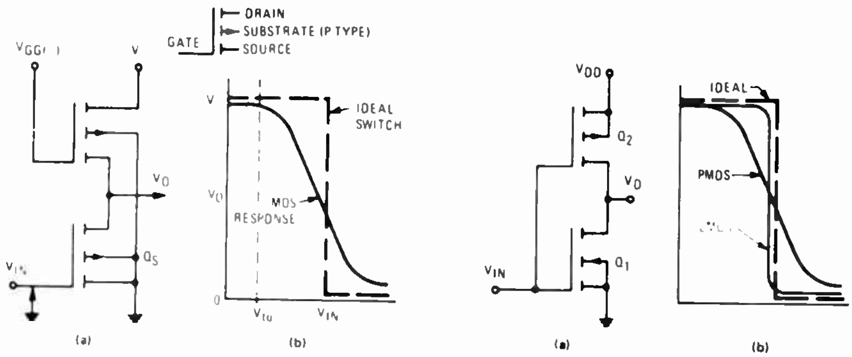
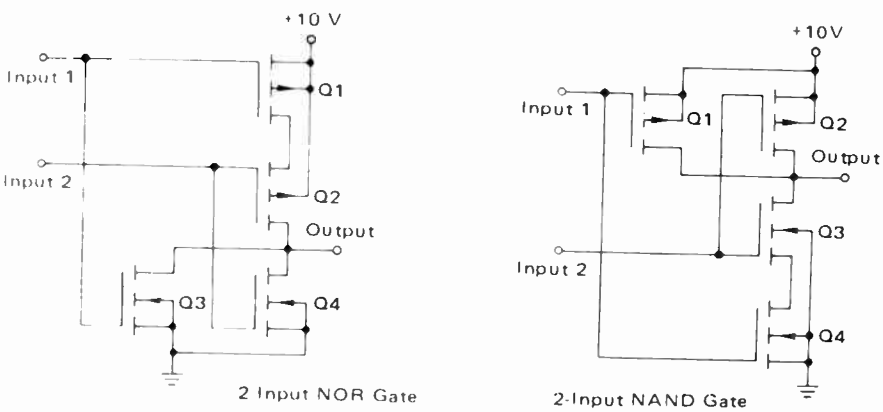


Fig. 8. A comparison of PMOS and CMOS inverters.



"0" 0 V	"1" +10 V	
Input 1	Input 2	Output
0	0	1
0	1	0
1	0	0
1	1	0

"0" 0 V	"1" - +10 V	
Input 1	Input 2	Output
0	0	1
0	1	1
1	0	1
1	1	0

Fig. 9 (a). A CMOS 2-input NOR gate.
(b). A CMOS 2-input NAND gate.

abbreviated to MOS. Hence, the MOSFET is a metal-oxide-semiconductor field-effect-transistor. They can be made in the two complementary ways, P or N, of which the first is most conventional giving the PMOS technique. With improvements in technology the NMOS technique is also being used for very complex circuit blocks.

The attraction to manufacturers and users of MOSFET devices is that the packing density of the active devices is the highest of all types — much better than TTL. It is, however, not as fast as TTL but adequate for a large part of the consumer market. FETs have extremely high input impedance, 1.0 TΩ (1 000 000 M Ω) is common.

THE CMOS FAMILY

IC device designers went a step further in the early 70's to produce yet another family — the complementary metal oxide semiconductor logic — CMOS. This combines both the P and N, MOS technique in a complementary manner to produce a more ideal switching action than PMOS — Fig. 8 illustrates the difference between the two. Fig. 9 is the schematic of a CMOS two-input NOR gate. One of the most significant advantages of CMOS is its low power dissipation. Because of the extremely-high off-resistance of MOS transistors, and because only one transistor of the series-inverter CMOS pair is ever on at the one time, the dc current drain of a CMOS inverter is in the low nano-ampere range as compared to 0.2 milliamps for a low-power TTL inverter. At the system level a CMOS system typically requires one-twentieth to one-thirtieth the power required by an equivalent TTL system. In fact torch batteries are adequate to run quite complex CMOS devices.

Because of the complementary configuration CMOS has high common-mode noise immunity (such as power supply variations). It will operate with supplies from 3 V to 16 V and needs only a single positive supply. In addition the high packing density has allowed the building of low cost large-scale-integrated (LSI) packages. Interfacing to conventional transistor logic is easy because it has a low output impedance. It generates low noise because of nicely conditioned rise and fall times. The fan-out factor of 50 is the highest of all logic because of the extremely high input impedance. Its speed is better than PMOS but not quite up to that of TTL — propagation delays being from 12 to 60 nanoseconds depending upon supply voltage used.

We have merely glossed over these various kinds of device because we are mainly concerned here with digital systems in general. To design systems requires little in-depth understanding of the manufacturing method used for the actual logic element. The logic IC is merely a black-box with certain input-output characteristics as stated on a data sheet.

Why do new families keep emerging? The facts are that there is still a cost saving to be had and the market is huge. The estimated value of the total market for IC devices in 1980 runs to around $1,000,000,000. CMOS offered new horizons in cost savings in manufacture. As a bonus from the power requirements of CMOS systems, the so-often neglected power supply cost drops remarkably. One example published in 1973 gave the comparison that a given transistor-based logic (called bipolar to distinguish it from MOS) would have a power supply cost of 33c per MSI bipolar function compared with 2c for CMOS. In fact one watt will power 50 CMOS devices.

We have all experienced the remarkable increase in the use of active devices in the last decade or so. A transistor radio now costs $1.50 to make commercially — and there the semiconductors are but a small part of the cost. (Just one common thermionic valve costs more than this today!) A calculator using hundreds of devices can be bought for $5.

LOGIC DIAGRAMS AND TERMINAL CONNECTIONS FOR SN7400 SERIES

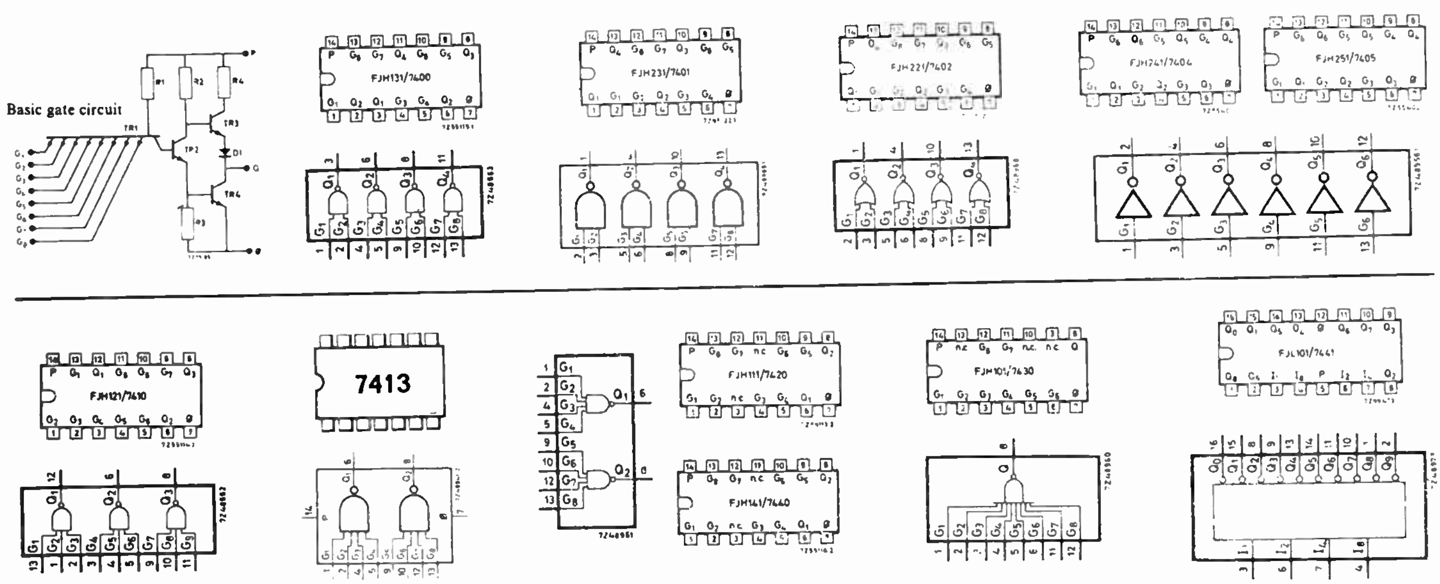


Fig. 10. The basic logic gates of the TTL family.

Minicomputers and microprocessors are rocketing down in price – tens of thousands of elements for a few hundred dollars. It might be argued that reducing the price will soon reduce the makers' total income but as prices fall, applications for digital circuitry widen at an even faster rate, thereby keeping up an expanding demand.

THE DUAL-IN-LINE PACKAGE

The most commonly used form of IC logic package in small batch production is the dual-in-line arrangement with 8, 14 or 16 pins. Large-scale integration LSI used by specialist manufacturers will vary in number of connection, but systems based on these require very large volume sales to make a large special system economic. Thus LSI chips are largely restricted to computers or very high volume things such as calculators and digital clocks.

The number of connections decides the available combination of functions or inputs and outputs. Assuming a need for two power-supply terminations, of which one is the common for all functions, a 14 pin device will have 12 pins available to produce various combinations of input. Fig. 10 shows the main units that are marketed. Available are the sextuple inverter (6 inverters with one input and one output each); the quad 2 input NAND (four two-input NAND gates); the quad two-input NOR, triple three-input NAND, dual four-input NAND, single eight-input NAND plus other more special combinations such as a more powerful dual four-input NAND buffer and others to be discussed below. Each gate function is a quite separate entity on the substrate; when schematic circuits are drawn the individual gates can appear anywhere on the system schematic, as shown in Fig. 11, where part of a digital-voltmeter circuit is given. Fig. 12 shows a component overlay and actual circuit boards for the same voltmeter.

As well as gates there are several other basic digital-system building blocks. These are the flip-flop, (more correctly called the bistable), the monostable, the astable and the Schmidtt trigger. Let us look at each in turn.

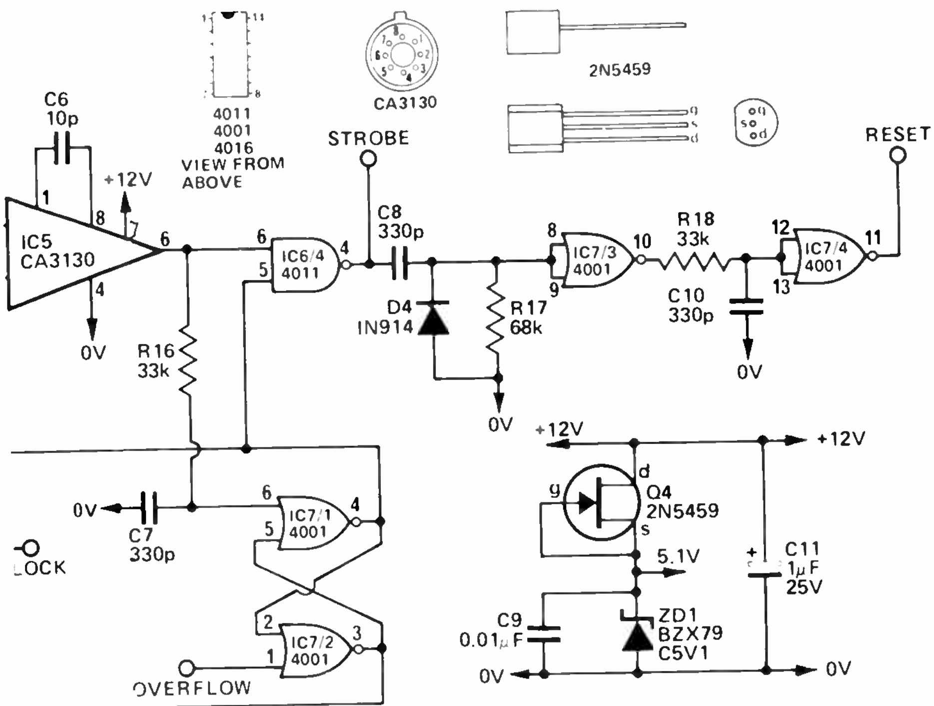


Fig. 11. In a schematic diagram the individual gates in a common IC package are drawn to suit the system layout – not the actual wiring layout. This is illustrated above by part of a digital voltmeter circuit.

THE FLIP-FLOP

Gates are used to perform logical arithmetic, such as, allow event A to occur when B or C have operated. Digital systems can be greatly enhanced by the addition of blocks

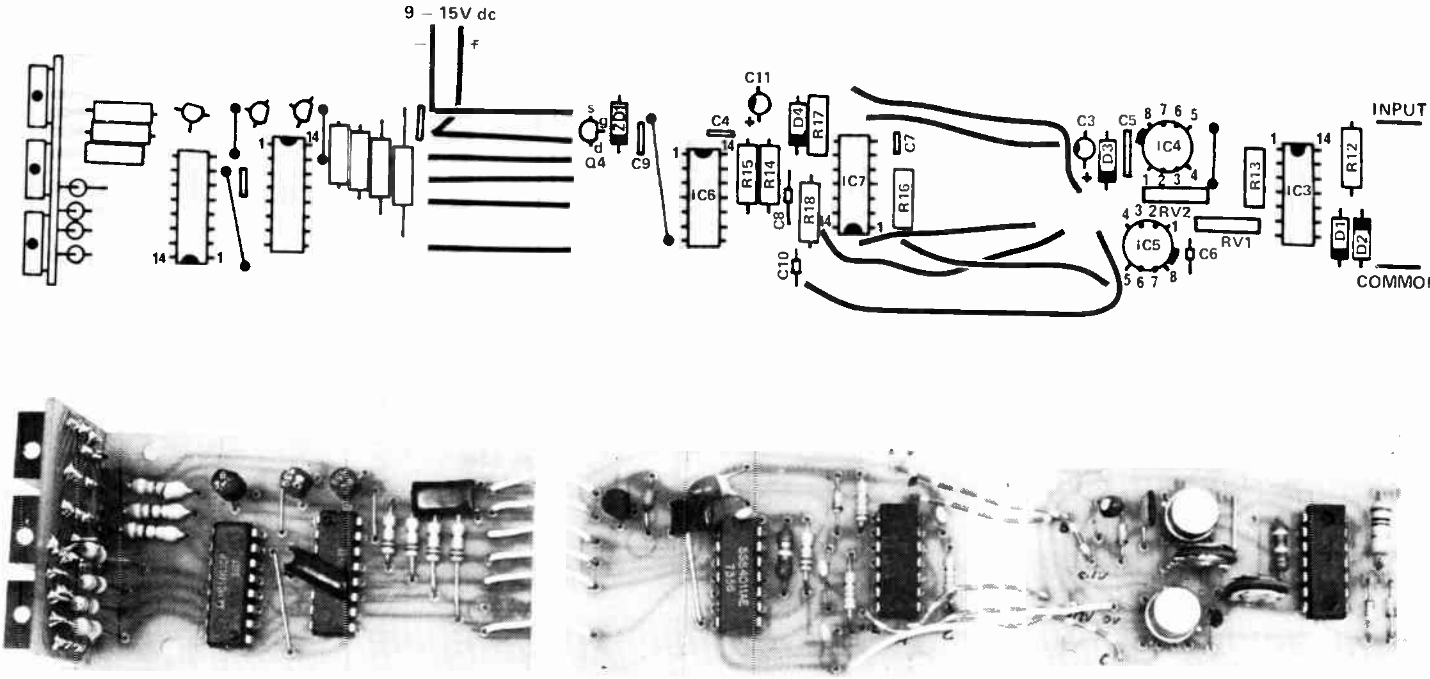


Fig. 12. Wiring layout and actual circuit boards of the DVM shown in Fig. 11.

that remember and count the digital numbers. It is quite possible to use the gates we have discussed to form counting stages but it is more economical to build specific circuitry that will count pulses and/or store values in a binary form. The basic block of modern counting technique is the flip-flop – FF or bistable – one of the multivibrator family.

The most elementary flip-flop circuit is repeated from Part 18 as Fig. 13. It consists of two-transistors and several resistors which are dc coupled in such a way that each stage provides a signal which controls the state of the other.

In essence the characteristic behaviour is that one stage will be held ON when the other is OFF, or vice versa. To reverse the situation a pulse or change in input signal level (which can be applied to either the base, emitter or the collector) will cause the system to toggle over to the other state. Because of the heavy degree of positive feedback provided, the circuit does not dwell in the in-between state. State change is very rapid – nanoseconds in well designed IC flip-flops.

A decade ago flip-flops were built from two, or in the faster toggling circuits, four transistors. Considerable effort was expended to provide a fast, reliable flip-flop action. Today a typical IC equivalent, see Fig. 14, uses many active elements for less cost than two discrete transistors. As discrete designs play no real part today we will only give the characteristics of the flip-flop, not its design details.

Flip-flops provide a counting action because each pulse at its input causes the system to switch over one state – the output, therefore, switches state at each *second* input pulse providing a divide by two action. The output is a switched level which can then be used to pulse a following stage dividing by two again and so through cascaded flip-flops.

The most commonly used flip-flop symbols are given in Fig. 15. Outputs are denoted by the symbol Q. A flip-flop has two outputs, one of which is the inverted value of the other, that is, Q and $\overline{Q}$. The pulsing input is denoted T for trigger. As well as these connections we need a set and reset input denoted S,R. (Although often only R is provided). These enable the flip-flop to be set up on demand with the output Q set to either 0 or 1 state as is needed. This is essential firstly because a flip-flop can come up in any state when the power is energised, and secondly because it may be necessary to set counter stages to a given binary number.

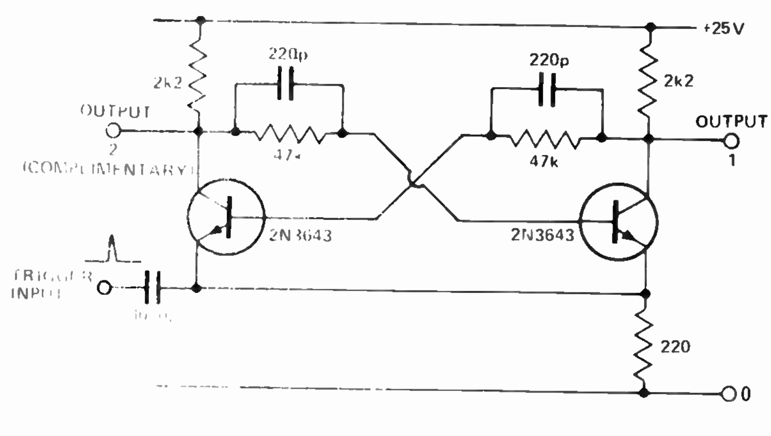
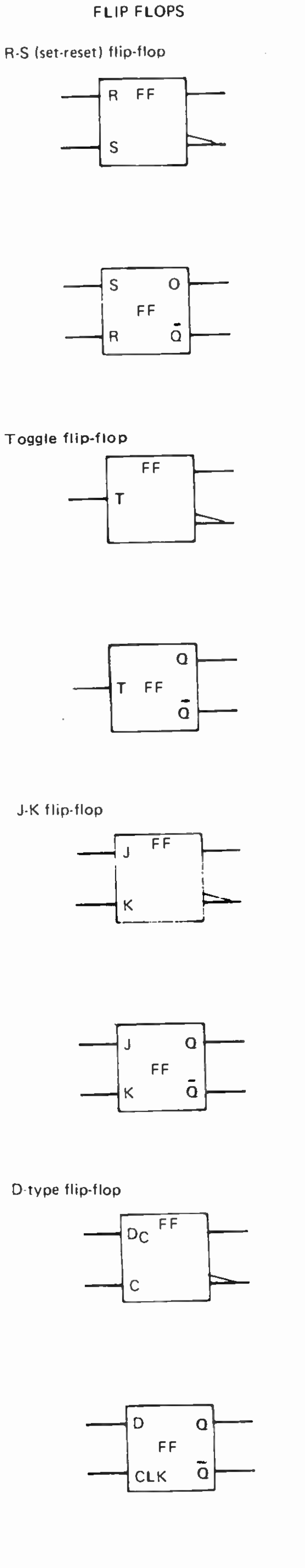
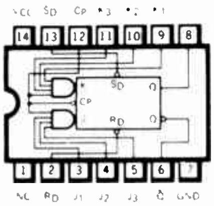


Fig. 13. A basic flip-flop built with discrete components.



DIP (TOP VIEW)



SCHEMATIC DIAGRAM

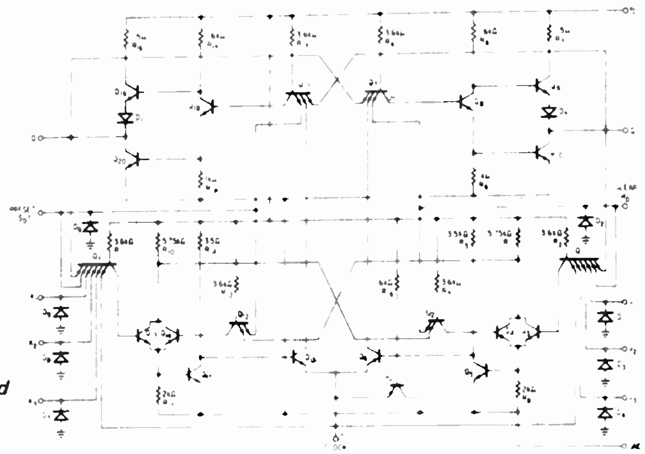


Fig. 14. Schematic and logic diagrams of an integrated circuit flip-flop using bipolar devices.

Component values shown are typical.

LOGIC DIAGRAM

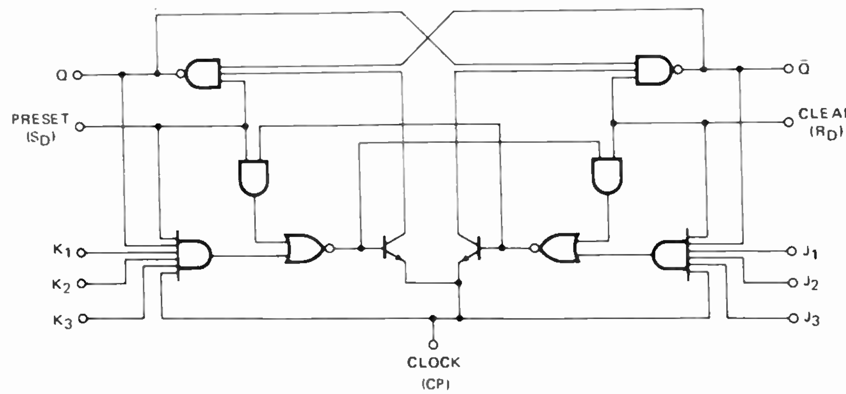


Fig. 15. Schematic symbols used for different types of flip-flops.

Some flip-flops provided in IC form also may contain three-input AND gates that feed the set and reset inputs. These are known as the J and K inputs giving the name J-K flip-flop. The JK flip-flop overcomes an ambiguous output state, when both inputs are high, that occurs with the RS flip-flop. We will see, when counting systems are discussed, why the reset ability is vital in many counters. Fig. 16 gives the schematics of some of the available flip-flop ICs.

THE MONOSTABLE

If one of the bias resistors of the bi-stable circuit is replaced by a capacitor, as shown in Fig. 17, the circuit provides a different action from the normal flip-flop. When triggered the circuit changes state, but only stays toggled in the changed state for a time decided by the product of the values of the capacitor and its "charging" resistor R. ($T = 0.7RC$). Hence a pulse input will cause the output to change state and remain there for a chosen time interval before it triggers back to the original state.

The monostable, also called a one-shot or single-shot provides, as we saw in Part 18, an output pulse having a designed length and height which remains the same irrespective of the input pulse shape. It, therefore, finds application as a pulse reshaper. As the duration is fixed it can also be used to generate a pulse that is delayed from the triggering pulse by the length of the monostable pulse. Monostables are available in IC form, and can provide pulses of duration from 20 ns upwards to minutes or more by appropriate choice of values.

THE ASTABLE

If both feedback paths use capacitive coupling the circuit becomes self-toggling with the stages alternating in state without being externally driven. We considered this circuit in Part 18 when discussing signal generation. The astable is important in digital systems for it provides the square wave signal that increments the digital system along pulse by pulse. It acts as the 'clock' regulating a digital system's sequential operations.

Astables are not usually produced directly in IC form, for the same action is obtainable with other elements, for instance, with the next element to be

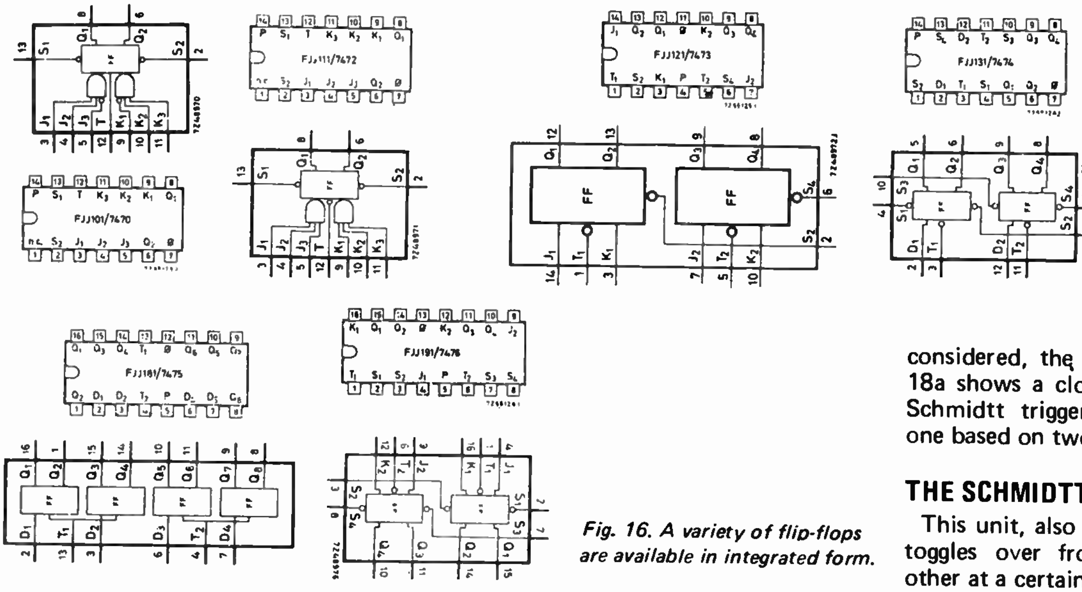


Fig. 16. A variety of flip-flops are available in integrated form.

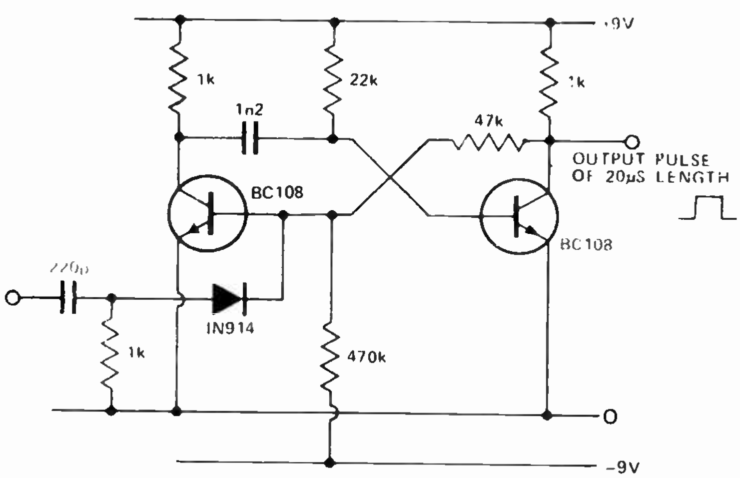


Fig. 17. A basic monostable constructed with discrete components.

considered, the Schmidtt trigger. Fig. 18a shows a clock source based on a Schmidtt trigger IC. Fig. 18b shows one based on two NAND gates.

THE SCHMIDTT TRIGGER

This unit, also introduced in Part 18, toggles over from one state to the other at a certain input voltage level. It remains in the opposite state until the voltage falls below the threshold level. The Schmidtt trigger is used to produce digital signals from analogue signals providing the two necessary binary levels at the output which indicates whether the analogue signal is above or below the threshold and, which are compatible with the rest of the digital system.

The Schmidtt function is available in IC form as a dual-in-line pack. It has a four input AND gate feeding the actual trigger circuit and is buffered with an inverter. The Schmidtt trigger is readily identified in Fig. 18. Its preferred symbol is given in Fig. 19.

YOUR LIBRARY

There are a bewildering number of digital ICs and to identify them correctly it is wise to have a good range of manufacturers' catalogues and application notes.

REFERENCES

The reading list given in Part 22 is also relevant to this part.

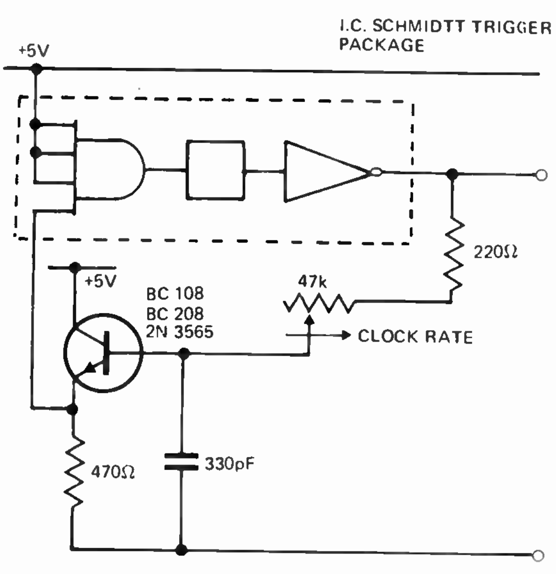
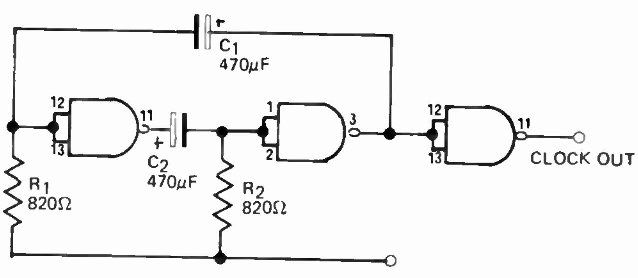


Fig. 18(a) LEFT: A Schmidt trigger combined with a transistor provides an astable multivibrator action.
Fig. 18(b) BELOW: Astable action may also be produced with two NAND gates.



BASIC DIGITAL ELECTRONIC TERMS . . .

test yourself on these. Each is explained in this chapter.

AND-gate	Circuit that performs logical AND operation.
Bi-stable (function)	Two state circuit that is used for binary counting.
Boolean algebra	Mathematical algebra for logic operations.
Boolean statement	Mathematical way of writing a logic situation.
Clock	Source that causes a digital circuit to step on bit by bit.
Digital circuit	Electronic circuit that processes information in digital form.
Digital IC	Integrated circuit unit using digital techniques of operation.
Exclusive-CR gate	Special kind of CR gate: basic to digital computing.
Fan-in	Number of inputs that a digital circuit can have and operate correctly.
Fan-out	Number of output circuits that can be driven by a digital stage without altering operation.
Flip-flop	Same meaning as bi-stable but more generally used.
Gate	Any circuit that performs logical algebra AND, OR, NAND, NOR function.
Inverter	Circuit for changing state of a logic level to its negated form.
Indicator	Device used to show logic state.
Karnaugh map	Matrix representation of logical statement; used in minimising a logic circuit.
Logic (circuit)	Collective circuitry for performing logic operations.
Minimization	Process of reducing degree of redundancy in a prototype logic system.
Mono-stable	Circuit for generating a known interval; square pulse; used to regenerate digital signals or to provide a delay.
Multi-vibrator	A general name for the family of bistable, monostable and clock circuits. Also used to describe the clock source itself.
NAND-gate	Circuit that performs logical NAND function.
Negative logic	Circuit in which logical 1 is represented by zero or negative polarity voltage.
NOR-gate	Circuit that performs logical NOR function.
Positive logic	Circuit in which logical 1 is represented by the positive polarity signal; usually 5 V in TTL.
Redundancy	Part of logic circuit that could be removed without altering logic operation. Often used to ensure total reliability — sometimes there by accident!
Schmidtt trigger	Circuit used to provide digital signals from analogue levels. It changes state at a predefined constant level.
State of circuit	In digital terms, the logic level at a point in the circuit.
Truth table	Matrix table displaying states of logic circuit in terms of input and output states.
Venn diagram	Overlapping circles drawn to represent a logical situation.

24 Digital sub-systems counters and shift registers

IN THE preceding articles of this series we have described the basic building blocks of digital systems. To summarize, there are the various gates, the flip-flop, the monostable, the Schmidtt trigger, the inverter and the square-wave clock source — a surprisingly few basic elements from which all the countless different forms of digital equipment are constructed. We are now in a position to examine how digital signal systems are put together using these basic building blocks.

DIGITAL NUMBERS INTO DIGITAL SIGNALS — COUNTER SUB-SYSTEMS

Not long ago digital systems were invariably built up from individual blocks where each of the above functions could be clearly identified in the system. But not so now. Many of the blocks now marketed as basic building elements are complex systems in themselves. The most extreme example is probably the micro-processor system (it provides the bulk logic requirement of a powerful computing system) which is now available as a 'throwaway' element for around $10 or less. It would take to the end of this course to begin to appreciate the complete system — operation of such a building block!

We have already become involved in small systems — the exclusive OR and the half-adder of Part 22, for instance. The next step to take is to form sub-systems with the fundamental blocks that provide us with the facility to form and manipulate digital numbers, because many (but not all) digital systems operate with numbers either to provide means of calculation, or to provide a display of numbers. Thus we need to know something of digital counters and the somewhat similar units known as registers.

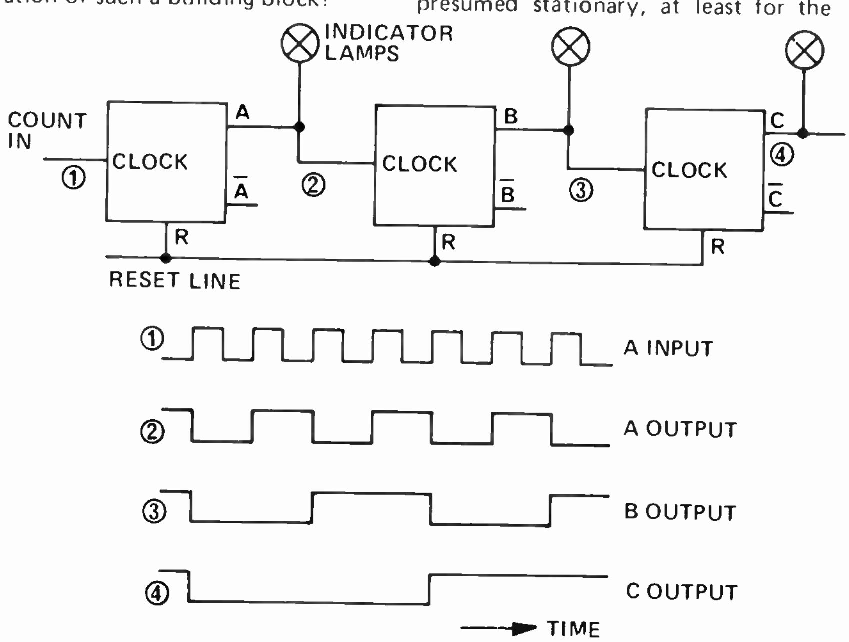
We saw in the previous part how the flip-flop provides a counting action by virtue of its ability to switch states for each pulse appearing at its input. However nowadays it is a level change rather than a complete pulse which causes the transition.

BINARY COUNTING

Cascaded flip-flops, such as shown in Fig. 1, form the simplest type of counting system. Each time the input changes state, the A flip-flop toggles back and forth delivering a state change to flip-flop B for each second input change — and each fourth change to C and eighth change to D and so on.

At any instant (where the input is presumed stationary, at least for the interrogation period) the outputs A, B, C, etc. will either be at a 0 or 1, as shown in the truth table. For example, a count of five (decimal) will be registered at 101 for CBA respectively. Note that the truth table appears to be written back to front — the reason is simply that we write numbers (by convention) with the most significant digit to the left hand of the number and this corresponds with the furthest right-hand flip-flop, its position on the schematic arising from the drawing convention used for the signal-flow through information systems. Thus a stream of input pulses with time are converted into a multi-element digital number. This form of input is often referred to as a "crazy-digital" number system when applied to systems incorporating measuring sensors. Such sensors generate pulses not having any clearly obvious relationship with time — examples are digital-position sensors wherein a pulse is generated for each unit displacement occurring.

Clearly, if the state of A, B, C, etc. is to represent a number the count must start from some clearly defined initial condition for each stage — often, but by no means always, stages are reset to 000 — by applying a level to the reset R line. In some applications the number must begin at



INPUT	C	B	A
0	0	0	0
1	0	0	1
2	0	1	0
3	0	1	1
4	1	0	0
5	1	0	1
6	1	1	0
7	1	1	1
8 = 0	0	0	0

TRUTH TABLE

Fig. 1. The asynchronous or ripple-through binary counter is the simplest to implement with flip-flops.

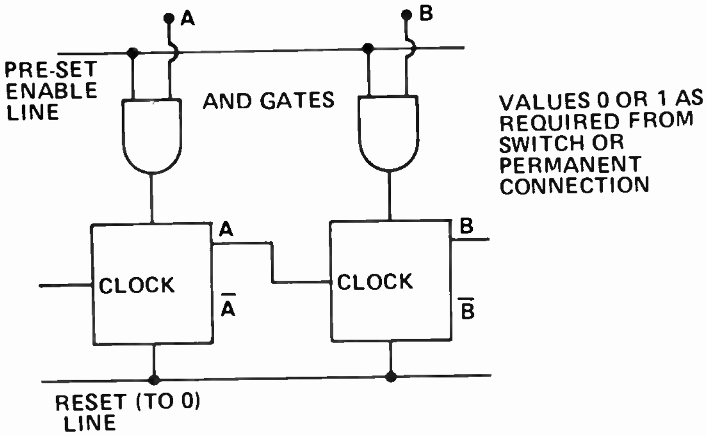


Fig. 2. Counting stages require their initial conditions be set – either to 0s or to desired values.

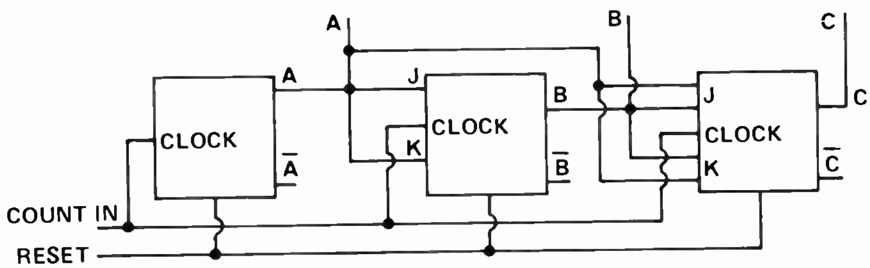


Fig. 3. Synchronous binary up-counting with JK flip-flops.

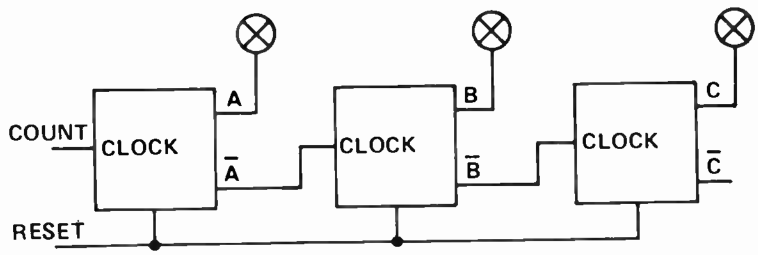


Fig. 4. Reverse counting is simple – use complement outputs instead to trigger the next stage. (a) Ripple-through count-down.

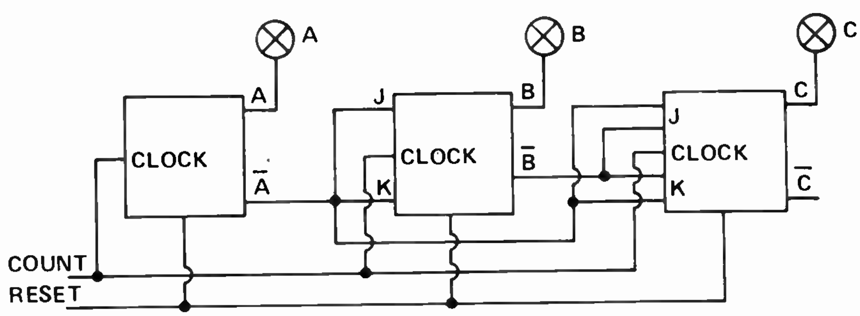


Fig. 4(b) Synchronous count-down.

a specific value. This is achieved by operating the preset lines accordingly for each stage, the value being arranged, say, by the setting of numeric dial switches usually operating through solid-state gates. Once these are set, a single command signal can preset the whole counter to that chosen number. Figure 2 shows how this is achieved using gates.

It is often desirable to display the digital value for visual interrogation. An indicator lamp (light emitting diodes LEDs are now used) driven from the A, B, C, etc., will show the positive logic binary number. Driven from $\bar{A}$, $\bar{B}$, $\bar{C}$, etc., it would show the negative logic number.

Waveforms for this form of counter (called the ripple-through type) are shown in Fig. 1. Although we refer to pulses flowing through, the waveforms are actually square-wave trains in the dynamic state, or levels in the static state, that have frequencies divided down in the ratio 1:2:4:8 etc. This 'pulse' form of expression is a hangover from the early days of digital technique (just a decade ago) where actual pulses, not levels, were used to trigger flip-flops. Each square-wave transition was differentiated by a simple RC circuit to provide a pulse of energy that triggered the next stage. Today, this method is unnecessary and is seldom used.

ASYNCHRONOUS AND SYNCHRONOUS COUNTING

Once an input pulse has occurred, it sets the chain in action, each flip-flop passing each second input pulse on, to the next stage. As each stage has a finite delay time – nanoseconds with TTL, microseconds in older forms of logic – each stage triggers at a later time than the one before it. Hence the form of action of each stage is said to be asynchronous with the others. Thus whilst the pulse is rippling through the counter the outputs are in an undefined and changing state between the previous and next correct states. The output cannot therefore be read until the whole thing has settled.

It is quite common to have binary counters with over 20 stages – at 100 ns delay in each, the maximum input pulse counting rate (if the outputs are to be used whilst it is in a counting transition mode) would be limited to 2µs between incoming pulses, that is, 500 kHz. Faster logic is available that provides around 20 ns delay but this still seriously restricts the data-transmission rate where the ripple-through design is used.

This disadvantage can be overcome by increasing the circuit complexity somewhat to form what is known as a SYNCHRONOUS counter. Each stage in the cascaded chain is fed a clock pulse simultaneously via control gates. The control gates for each stage also receive inputs from all previous stages such that the particular stage only operates at the correct count.

By this means all stages operate synchronously and the propogation delay is reduced to that for a single stage only. Synchronous counters are essential where the outputs of all stages must be decoded in parallel, eg, where a display of the count is required. However for straight frequency division applications, where the output is taken only from the last stage, a ripple counter is normally faster than the synchronous type.

The logic-gate inputs of the JK flip-flop allow them to be connected for synchronous counting as shown in Fig. 3. It is not important to know how JK flip-flops work internally for counters are now built by cascading ICs as per application note instructions — it takes a specific type of mind to realise digital counting systems without effort! Fortunately for those of us without this ability it is rarely needed except, of course, by IC designers.

UP, DOWN AND REVERSIBLE COUNTERS

So far we have only looked at counters that increase up the binary number scale for each additional input. To make the same system count down is incredibly simple — we merely re-connect them so that the complemented outputs are fed to the next stage -- see Fig. 4 — instead of using the normal output, that is, feed A, B, C, etc. to the count inputs. It will then count down in binary sequence. Intuitively we would expect this because of the two-state complementary nature of binary numbers. If you are worried about numbers passing through zero, try your digital arithmetic on a count-down case starting at 000.

In many applications needing counters, one-way counting is satisfactory. Examples that come to mind are nucleonic pulse-event counters, counter-timer units and counts of objects passing a given point. In some requirements, however, the need is to add or subtract pulses to provide at any time the instantaneous sum or difference between two inputs, or to give the nett value from a single measurement parameter that alternates in sign. Examples here are digital-position indicators where the direction of movement reverses, integration of reversible variables such as the flow of solid or liquid past a point the number of vehicles in a car park, and situations where the difference between two pulse-train variables is needed.

Several methods may be used to

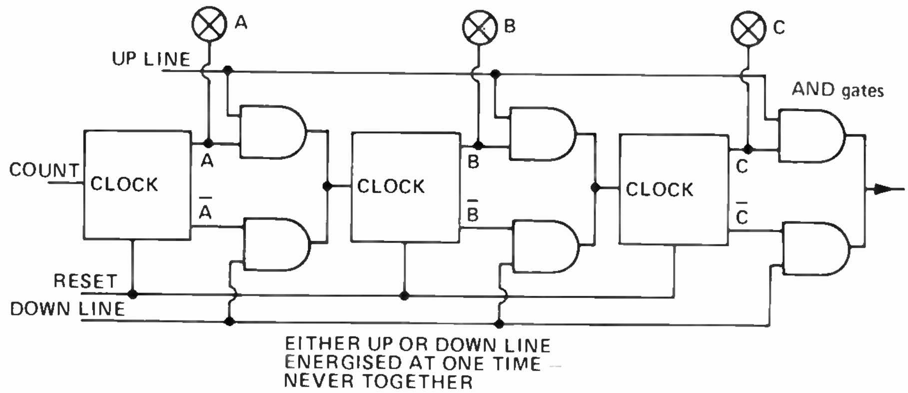


Fig. 5. Reversible counting uses fast switches to select which output of each stage feeds the next. (a) Asynchronous — line controlled.

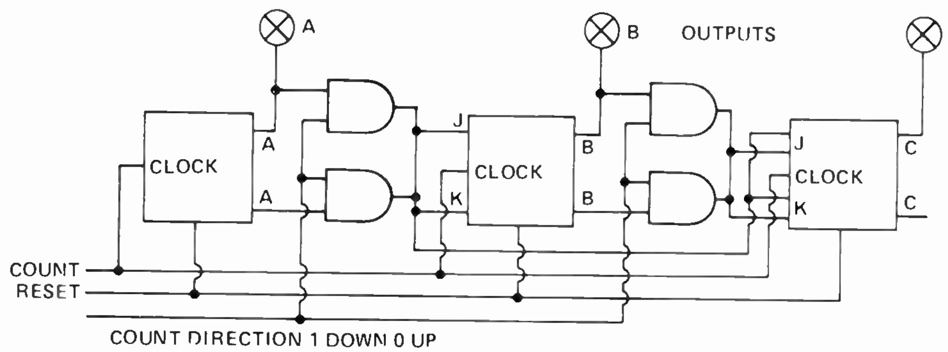


Fig. 5(b) Synchronous — line controlled.

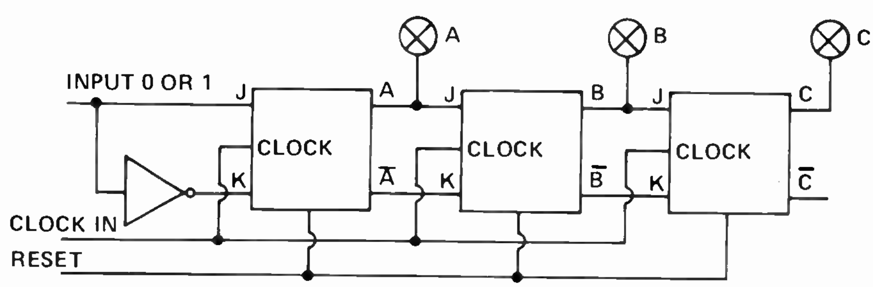


Fig. 6. The shift-register is also built from flip-flops but with different connections to counters. (a) JK type

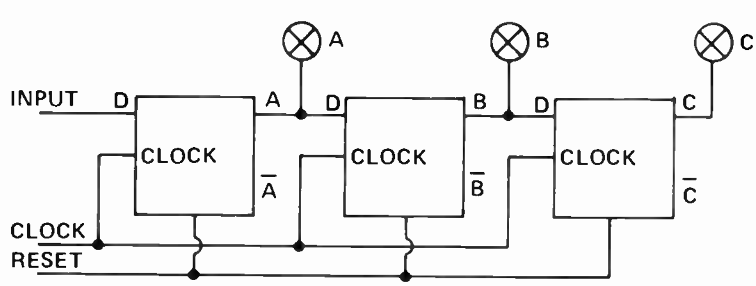


Fig. 6(b) D type.

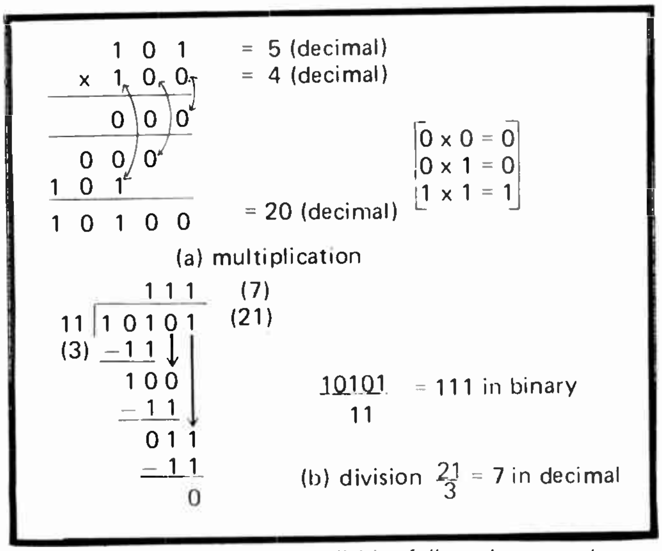


Fig. 7. Binary multiplication and division follows the same rules as decimal numbers but note how simple multiplication becomes — a process of shifting and adding 0 or 1.

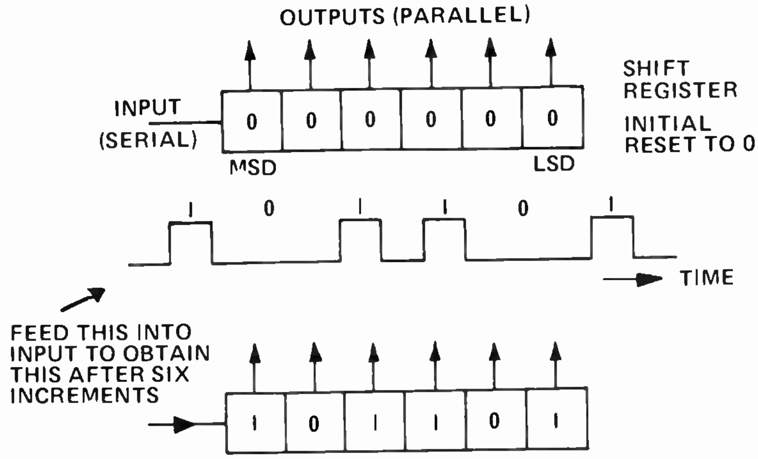


Fig. 8. Registers can provide serial to parallel word conversion, the reverse also applies.

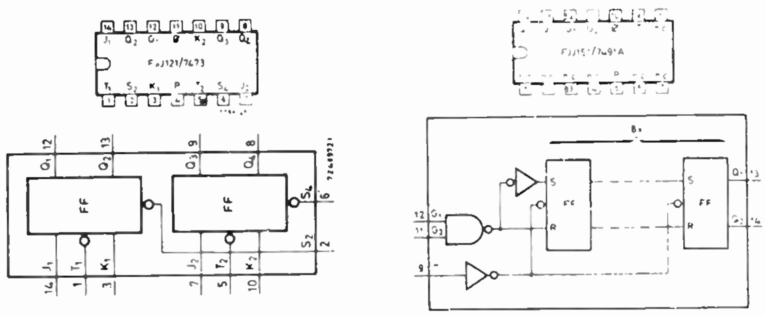


Fig. 9. Schematics of IC units. (a) Binary stages for counting or other purposes. (b) 8-bit shift register.

construct up-down or reversible counters. The most common method is to use a common pulse-count input that accepts both 'up' and 'down' pulses, the decision to add or subtract each individual pulse being decided by the simultaneous voltage levels applied to control lines. The control lines select whether the pulse is routed through the A, B, C, etc., or $\bar{A}$, $\bar{B}$, $\bar{C}$, etc. paths. Switching is accomplished using logic gates as shown in Fig. 5a and 5b — these provide adequately fast switching. In the ripple-through variety the direction-line commands must be held stationary until the counter stages have settled in order to preserve accurate counting. The delay is less pronounced in synchronous designs. Within certain limitations it is also possible to design counters that will accept a new pulse while the system is still in transition. This is done by 'holding' the pulse (by applying a delay) until the counter is ready to accept it.

SHIFT REGISTERS

Although not a counter in the same sense as above, the shift register also consists of a cascaded chain of flip-flops but with different connections. The purpose of the register is to hold a binary number but allow it to be shifted as a whole to the left (toward the most significant digit — called a forward shift register) or to the right (toward the least significant digit — the reverse shift register) one step for each input pulse.

As shown in Fig. 6, the incrementing signal, which is more usually a free-running clock signal than a one-at-any-time instruction, feeds the count inputs causing all stages to toggle in synchronism. The state of each following stage, being tied to the output of that preceding, goes to that of the one before with each clock pulse. Thus a number can be fed into one end in serial fashion and will be caused to pass through the register. The whole is cleared to zeros or reset to any desired value via the reset input line. Using D-type counters outputs A etc., go to D inputs. Whereas when using J-K counters A, etc. go to J inputs.

Registers perform three main functions in digital systems. Firstly a digital number can be delayed in time by the additive propagation time (divided by the clock frequency) of the number of stages it passes through, or stored indefinitely (provided the power is held on).

Secondly, one digital number can be successively offered up to another for digital summation of the two — a

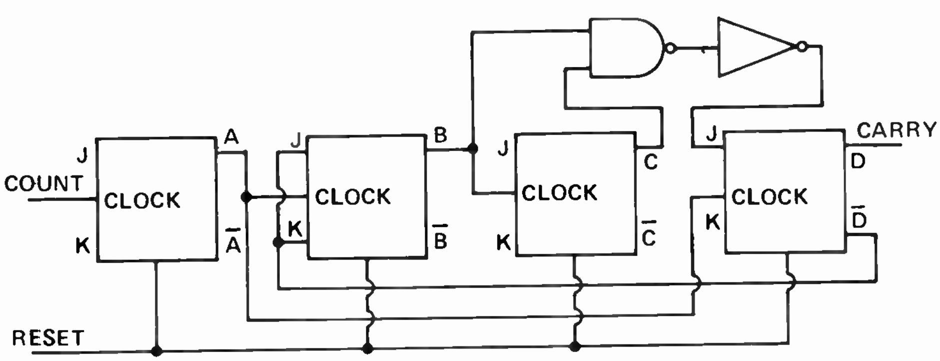


Fig. 10. Four JK flip-flops connected to count as a normal binary coded decimal NBCD stage.

basic step of multiplication by digital means. One number remains in position, the other increments across, the two being added at each step by the half-adder (part 22). Binary multiplication and division follow decimal number procedures but are much simpler – see Fig. 7.

A third use for shift registers is for the conversion of digital data from serial to parallel form and vice versa. For example it may require a group of eight pulses to define the position of a shaft with sufficient accuracy. The group of eight pulses is called a word and each individual pulse is called a bit. To transmit this word we would normally need eight lines. However a single line may be used if the information is transmitted one bit at a time, with synchronizing pulses to tell the receiving equipment where each word starts and ends. If the data when received is fed into a shift register serially, eight bits at a time, each word will appear in parallel at the output of the shift register, providing synchronism is maintained, where it may be decoded.

In reverse, a register can be set up to the desired number by the appropriate choice of stage inputs. Once set the register is incremented to feed out the number in serial manner to a single line.

Many forms of binary counters and registers are available in integrated-circuit packages. Figure 9a shows the schematic of a binary counter having two bits on the chip – that is it contains two flip-flops ready to be connected to count. Figure 9b is an eight-bit shift register. Both can be cascaded with like units to extend the bit capacity to virtually any length of binary word needed.

COUNTING IN OTHER THAN BINARY

In the binary counting system each bit position requires a counter element that has two stable states. In the decimal counter each digit needs 10 states to describe the individual numbers. Similarly three states are required for a ternary system, five for quinary, eight for octal and so on. Ten is not the limit – we use 12s and 60s in time and angle subdivision.

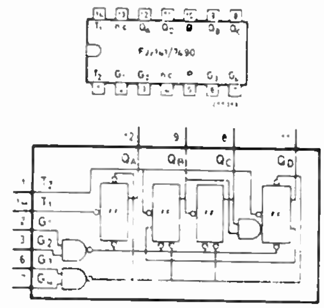


Fig. 11. Ready-made IC decade counter (SN 7490)

From the hardware realisation viewpoint, binary numbers are the simplest to hold because of the existence of the two-stage flip-flop. **(Three-state devices exist but have not** gained favour). From the user's viewpoint, however, we are more familiar with decimal numbers – a few people can read and work with binary displays but more would agree that values like 1025 decimal, which when displayed in binary form required 10 bits (1000000001), is difficult to read and interpret as a magnitude.

Flip-flops ideally count in powers of two – 2, 4, 8, 16, etc. – so any other value such as three or 10 requires modification to the counting procedure.

The flip-flop chain must be made to skip the required number of unwanted states in the truth table in order to return to the zero position. It is probably evident that this implies a waste of counting capacity when states are not used. In 10s counting, four flip-flops are needed to create 10 states – the other six of the 16 possibles go unused. Similarly, with three's (called modulo-3) two flip-flops are needed wasting one of four states, and with modulo-5 counters three flip-flops are needed with three wasted states.

Although a decade counter uses only 10 out of the possible 16 states, of its four flip-flops, this is not necessarily as wasteful as it may at first appear. For example if we wish to count up to 9999 with decade counters we will need a total of four counters containing 16 flip-flops. To count to the same number with straight binary counters we need 14 flip-flops – only two less. Additionally the ease of obtaining an easily interpretable display results in a system cost that is less than if an all-binary system were to be used.

In computers, information handling capacity is at a premium and the need to display the internal numbers negligible. In such cases the octal-number system comes into its own because three flip-flops provide eight states without any waste of states, and without need for the extra components required to skip unwanted states. The octal range is 0, 1, 2, 3, 4, 5, 6, 7 and then back to 0. A number 312 (octal) is $3.8^2 + 1.8^1 + 2.8^0 = 3.64 + 1.8 + 1 = 202$ decimal. Note that the decimal number requires roughly the same number of bit positions as the octal number but to implement decimal in digital hardware would need four (compared with three for octal) flip-flops for each bit position. However, where output is needed for human use – printouts and readout in numbers – the decimal system is best.

DECIMAL COUNTING

To obtain the 10 states 0 to 9 we must begin with enough flip-flops to provide them. Decimal or decade-counting stages, therefore, need four flip-flops which count in some form of code over just 10 states. The most straight-forward realisation is to let them count through the normal binary code and to apply interconnections between stages which prevent illegitimate states occurring and, often

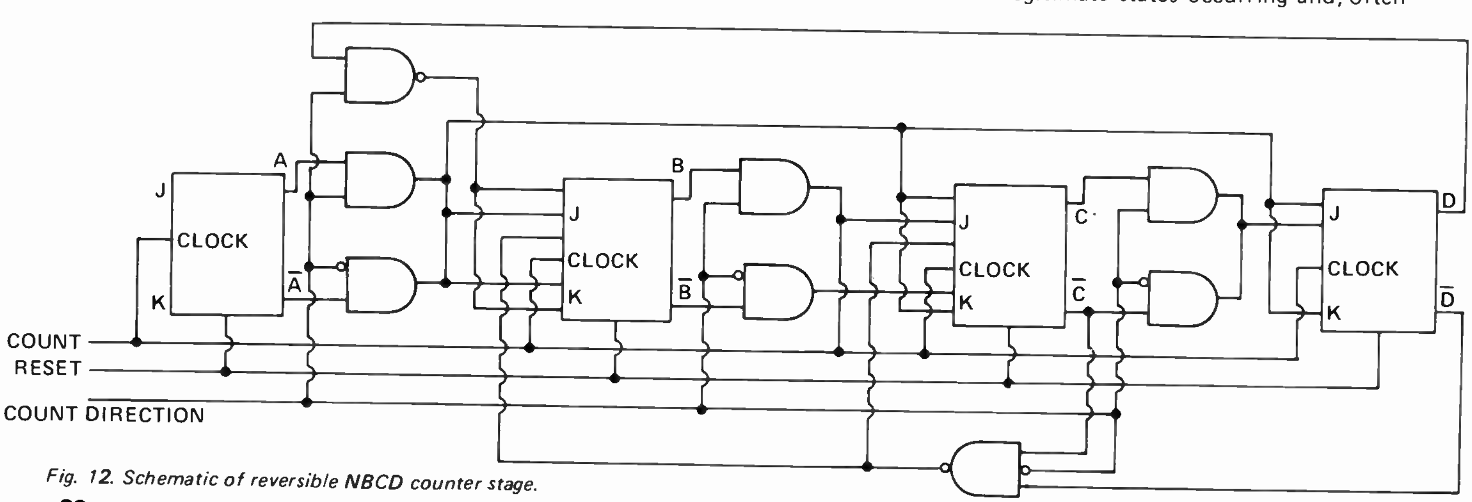
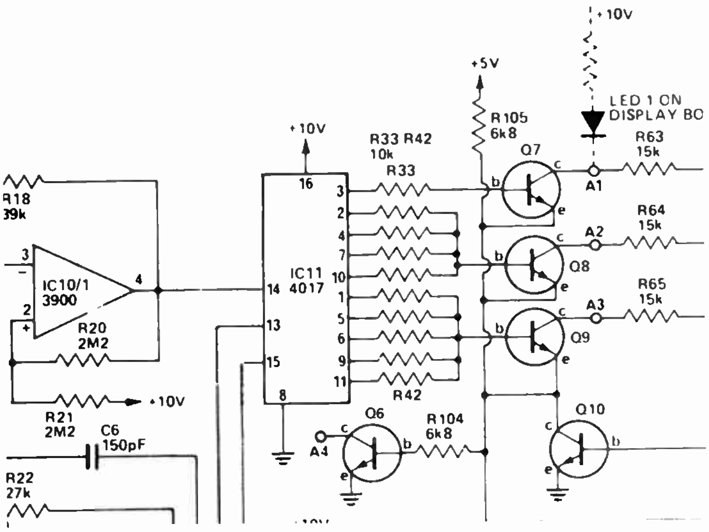
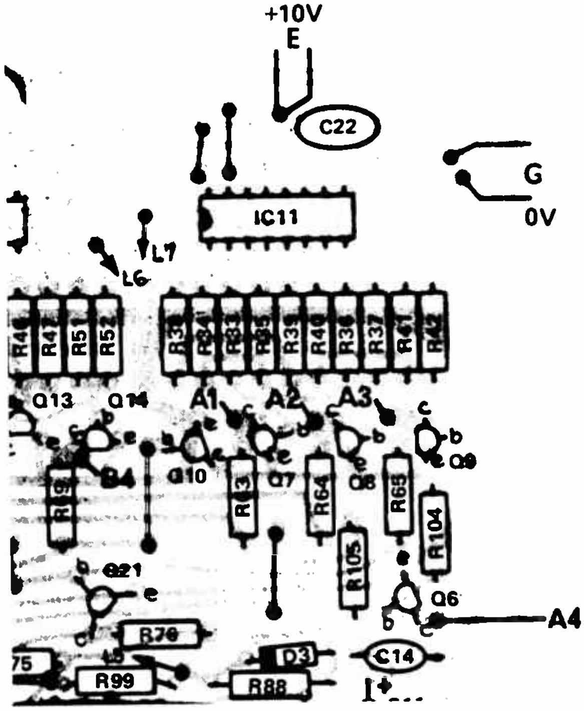


Fig. 12. Schematic of reversible NBCD counter stage.

ELECTRONICS -it's easy!



(a)



(b)

Fig. 13. In use IC counters are used as black-boxes. (a) How it appears in a circuit schematic. (b) How it appears on a circuit board.

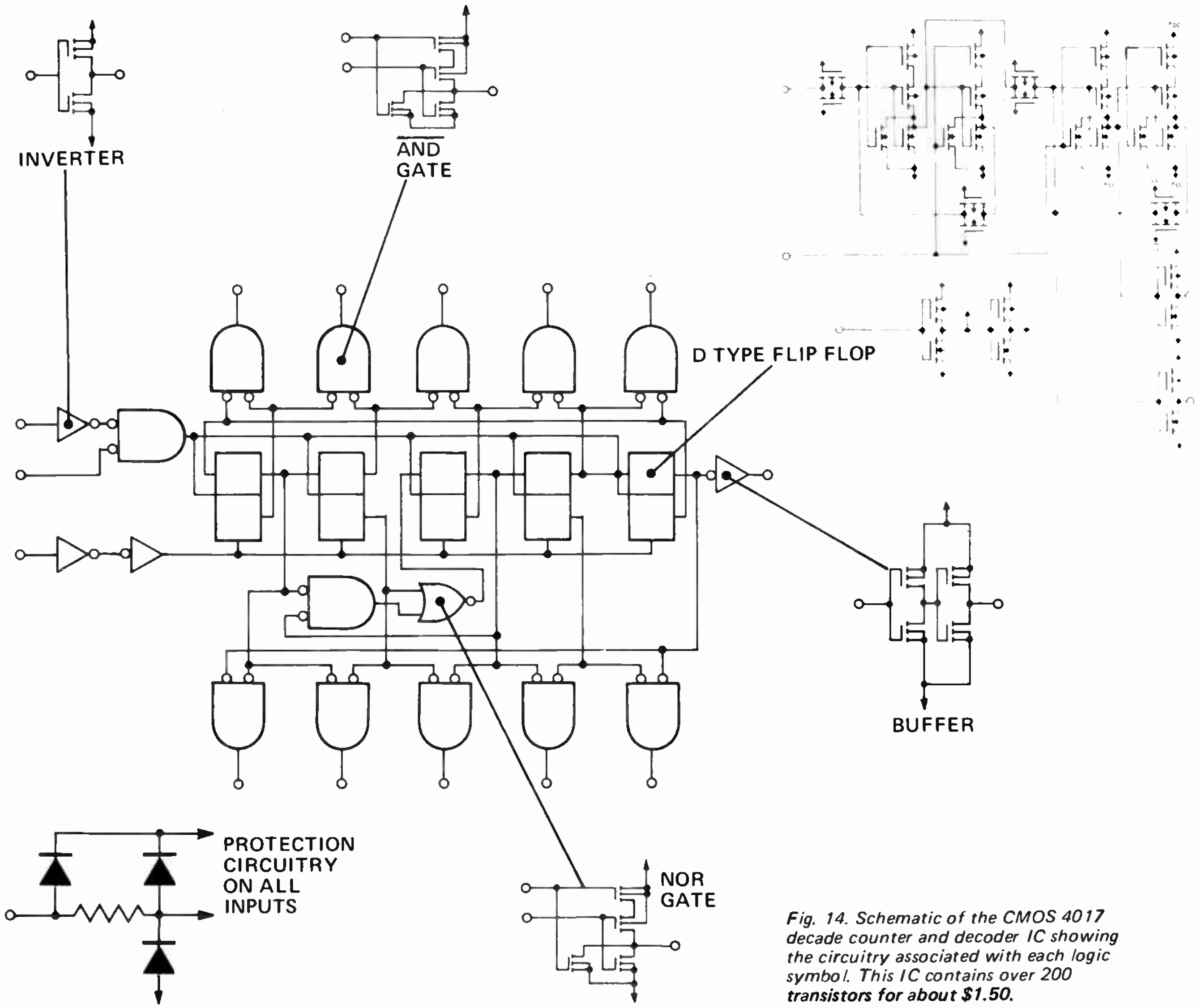


Fig. 14. Schematic of the CMOS 4017 decade counter and decoder IC showing the circuitry associated with each logic symbol. This IC contains over 200 transistors for about $1.50.

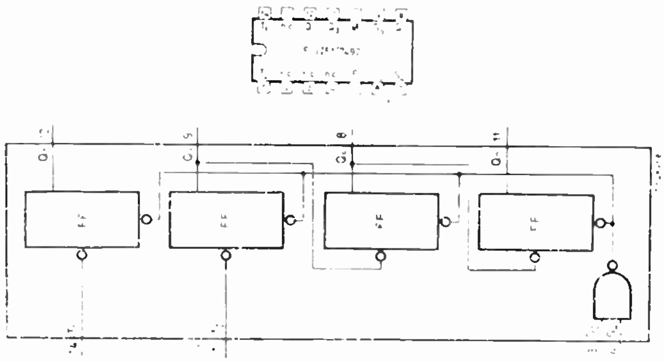


Fig. 15. Schematic of a divide by 12 counter the TTL 7492.

equally important, prevent the system locking up (as can happen in some counting systems).

Decade counters using straight binary are denoted Binary Coded Decimal – BCD. As there is no real reason why a stage should start with decimal 0 and binary 0 coinciding, it is feasible to construct a large number of different BCD counter stages using different code sequences. If the zeros do coincide it is called normal BCD (NBCD).

As we shall see in the next part, it is important to know that BCD code is used when the system needs a display in decimal. This is because the circuitry of the decoder depends upon the logic sequence of the counter as well as the requirements of the display method – more of that later.

At this stage design becomes a matter for the expert but fortunately ICs are available with the whole counter and BCD unit constructed ready to act as one decade. To obtain multiple-position decade numbers we merely cascade decade counter ICs in accordance with manufacturers' data sheets. Figure 10 shows how four JK flip-flops can be connected to provide NBCD counting. Note the need for an additional NAND gate and inverter. Comparing costs, it would usually be illogical to build a counter this way for the cost of ICs to build such a system is higher than that for a ready-made decade stage such as that shown in Figure 11.

Many applications require visual display with each decade: it is now possible to purchase an IC with count and decode functions combined in the one chip.

The question of synchronous versus asynchronous counting in decimal stages still applies. As there are, however, only four stages in each decade, the ripple-through time is not usually as vital an issue as with binary counting. However, this time must be compounded with the ripple-through time for all decades – 12 or more are common on calculators. The fastest decade system is one where all decades are synchronised, not just individual stage flip-flops. The practical catch is, however, that synchronous designs require many extra connections as the number of stages increases.

Reversible decade counters are built in a similar fashion to the binary types – using A or $\bar{A}$ etc., as needed. The problem is complicated by the need to lose states but once again few people would be called upon to design one from scratch as single –IC systems are marketed ready-made. Figure 12 shows the schematic of a reversible synchronous NBCD counter.

USING COUNTERS IN PRACTICAL CIRCUITS

Application Notes explain the connections and any special conditions to be observed in using IC counters. The electronic designer today regards the appropriate IC as a black-box with pins which are wired accordingly – what is inside is of little consequence. Figure 13a shows how an IC counter is represented in a circuit diagram and Figure 13b shows how it is wired onto a printed-circuit board. The internal complexity is seen in the circuit schematic given in Figure 14.

OTHER COUNTERS

The register becomes a counter by joining the output to the input to form a ring-counter. This system passes a pattern around the loop one step at a time for each input pulse. Hence the logical state of the elements at any time represents the number of counts accepted. This is also a convenient way to recirculate a digital word which needs constant re-use. A faster version is the twisted ring-counter.

Counters are also used frequently as frequency dividers. For example, a 1 MHz clock source passed through one BCD decade provides a source at 100 kHz; through two decade counters 10 kHz and so on. As it is more convenient to provide stable high frequencies than stable low frequencies, precise low-frequency pulse trains are best produced by this subdivision process.

As four flip-flops will provide up to 16 states they can also divide by 12 by the use of stage interconnections. Figure 15 is the schematic of a divide by 12 counter. These are used in timing systems to provide two seconds, minutes and hours units. A divide by five plus divide by two is also available in the same IC.

Just twelve years ago my technician and I built one of the fastest up-down decade counters reported at the time. It used discrete components, it took several months to build, could reverse at about 400 kHz, cost about $300 in components alone, needed a shoe-box size container and a hefty power supply. Today a match-box size unit, including a battery, virtually indestructible in normal use, can be reversible at (at least) 10 MHz rates and is available 'off the shelf'. It costs a mere few dollars and is vastly more reliable. We have reached the point where the mechanicals – the knobs, dials, case and boards cost more than the electronics circuitry. At the time of printing, $30 buys a 100 step programmable pocket-size calculator.

REFERENCES

An excellent inexpensive book worth purchasing is: "Digital Instruments Course – Part 1, Basic Binary Theory and Logic Circuits" – A.J. Bouwens, N.V. Philips Gloeilampenfabrieken, Eindhoven, Netherlands, 1974. (Available from Philips offices).

This gives a little more depth than this course can allocate and is recommended for those people who are involved with building digital systems routinely.

An excellent book for basic digital circuits is "Pulse, Digital and Switching Waveforms" by J. Millman and H. Taub, McGraw-Hill, New York, 1965.

Although now several years old, a useful work is "Digital Computer and Control Engineering" by R.S. Ledley, McGraw-Hill, New York, 1960.

Instruments are discussed in "Digital Instruments" by K.J. Dean, Chapman and Hall, London, 1965.

Counting circuits are dealt with in "Electronic Counting Circuits" by J.B. Dance, Iliffe Books, London, 1967 and in "Introduction to Counting Circuits and Transistor Circuit Logic" Chapman and Hall, London, 1964.

Logic is the subject of "Binary Arithmetic and Boolean Algebra" by A.C. Gillie, McGraw-Hill, New York, 1965.

More recent books are "Digital Electronics for Scientists" by Malmstadt and Enke, W.A.Benjamin Inc, 1969 and "Digital Electronics" by Price Ward, TAB Books Inc., 1972.

BASIC TERMS USED IN DIGITAL COUNTING OPERATIONS

Asynchronous counter	Type of counting system in which the count passes through the circuit stages in sequence.
Batching counter	Counter used to provide control signals at preset numbers of counts. They may count up from zero or down from a preset value to give a signal at zero.
BCD code	Abbreviation for binary-coded-decimal; is a code which counts in binary for the sequence of one decade of ten, each unit operating as a distinct decade.
Binary counting	Counting in binary number form.
Binary number	Number expressed in the two-state, binary, form.
Bit	The name given to the least size number unit of a binary number
Bus	Circuit path which conveys signals from one stage to others; a common interconnecting link.
Carry	The number that results, to be carried over to the next higher number stage, when the stage reaches the capacity limit of the stage.
Clock input	Terminal to which the clock signal is applied.
Control line	Line along which control signals are sent to actuate a circuit, such as a counter stage, when logic desires it.
Crazy digital system	Counting system in which the counter fills or empties as pulses are produced by an external source. No clock is used to pace the counter on.
Decimal code	Counting code based on decimal numbers.
Digital counter	Counter that totals digital signals. As all numbers are discrete entities, all counters must really be digital in form.
Direction line	Circuit line along which control signals are set in a bi-directional counter to set the counting direction.
Down counter	Counter that has its flip-flops connected so that each pulse input causes the number to count down.
Gray code	Digital number sequence in which each step causes only one digit to change each time. Many alternatives exist. They are used to reduce the risk of error.
Latch	Facility often incorporated into counters that enables the number to be frozen whilst the counter continues to count on.
NBCD code	BCD code that starts with its zero at binary zero. Counting on in a binary sequence from there.
Octal number	Number with base of eight.
1.2.4.8. code	Alternative name for normal binary number code, the ratios representing the weights that each stage of a four stage counter will have.
Overflow	Carry quantity that has no place to go and is therefore lost. An indicator of this condition is often provided.
Pulse	The basic entity that feeds a counter. Originally they were short duration, fast rise-time pulses but today they are more likely to be a fast level transition going up or down in polarity.
Pulse event counter	Counter used to total physical processes that provide a pulse representing some kind of event. An example is nuclear radiation pulse counting.
Quinary code	Code that recycles in steps of five bits.
Register	A counter that holds a digital number in short term storage so that it can be operated upon.
Reset facility	Terminal to which the appropriate polarity signal will cause the flip-flops to change to a given desired state.
Reversible counter	Counter arranged to total pulses, adding or subtracting them as directed by control line levels or the nature of the pulse.
Ring counter	Counter in which the end of a counting chain connects back to its input. A number set into it can be sent around as needed without losing it out of the carry output.
Ripple through counter	Alternative name for asynchronous counter.
Set facility	Similar facility to reset. Two names used to help distinguish between the two alternatives existing for setting up a flip-flop.
Shift register	Register in which a number has been set up for the main purpose of moving it digit by digit for presentation to another number usually for binary addition.
Synchronous counter	Counting system in which all stages are updated together when an input appears. This overcomes the delay of ripple-through systems.
Ternary counter	Counting in the number three system.
Up counter	Counter wired to cause the contents to move up the number sequence for each input pulse.
Up-down counter	Alternative name for a reversible counter.

25 Digital displays – their development and forms

THE HUMAN operator of a piece of digital equipment often needs to know the results of a measurement or to monitor progress in a process etc. The most effective way of conveying the required information is by means of a visual display. The display must provide the information in a readily understandable form — that is, in the decimal numbers, alphabetic characters and symbols (such as pi, decimal point, plus and minus signs etc) of common visual experience.

Everyone will be familiar with common examples of digital display such as the now ubiquitous calculator, digital clock radios and cash registers etc. (Fig. 1). As the amount of information to be displayed increases we find much more complex displays, such as CRT terminals, which can not only display a full page of alpha-numeric information but can generate graphical data and line drawings.

In this section we will study the various types of device that are used to generate the displays of calculators, digital multimeters and similar instruments.

Decoding techniques used to convert those numbers held within the system (in the binary or binary-coded-decimal form) into decimal numbers and alphabetical characters are discussed in the next section. Displays are dealt with first because their requirements partly dictate the decoding techniques that must be employed.

HISTORICAL DEVELOPMENT

Originally the display of the digital numbers held within the circuitry was performed using single lamps for each binary bit position in the counter or register of interest. A light ON represented a 1 and OFF a 0. Occasionally this may suffice where binary readout is needed but the inconvenience experienced where decimal display is really needed overwhelms the relatively small additional cost required to convert the binary code to its decimal equivalent.

Decimal displays using several quite different means were developed. An obvious starter was the columnar method which used ten individual lamps placed in a vertical column behind graticules of the figures 0-9. The decimal number being displayed showed up as a single value in each decade digit position. The non-illuminated numbers do not show if a suitable contrast graticule is used. This technique was certainly better than a pure binary-display and many instrument systems of the late 50's and early 60's used this method. A considerably reduced peak power-supply demand results with decimal displays as only one lamp in each decade needs power at any time. Disadvantages of the columnar display were the sheer mechanics of arranging numerous light-globe holders on a panel (Miniature Edison Screw MES lamps were used), the non-aligned number appearance which was awkward, and the character size that had to be small to keep the panel within reasonable proportions.

To overcome size and alignment difficulties a number of manufacturers marketed ingenious opto-mechanical modules. One method used a moving-coil meter movement to rotate a graticule, containing the 0-9 numbers, inside a projection system. The filament lamp illuminated the rear of the projection screen to make a quite large number appear at the front. In this way numbers were generated at the same position giving a visually well-proportioned multidigit decimal number. A sketch of this method is shown in Fig.2. Watching such a display is somewhat disconcerting, for the individual numbers wobble into position with changing values. Another method used ten individual-lamp projection systems each with its image plane set on the same viewing location on the rear of a screen. Provided only one lamp was energized at a time the number was easily read.

Projection systems were, by today's standards, delicate, costly and far from ideal. Their continued use and acceptance occurred mainly because they did not require high voltage levels to operate them. Voltage-wise they were quite compatible with early solid-state circuitry although considerable current gain was needed between the counter circuitry — the flip-flops — and the lamp.

NEON INDICATOR TUBES

In parallel development were indicating display tubes based on the principle of the neon tube. These have become universally known as "Nixie" tubes, a name really applicable only to tubes of Burrough's origin. As these are still designed into new equipment today — projection devices are seldom used now — we will study how they operate in some detail.

The basic neon lamp has two electrodes, see Fig.3a, passing into the interior of a glass envelope filled with neon gas. With a low voltage difference applied between the electrodes the gas acts as an insulator and no current flows between the electrodes. At around 70 V the gas conducts producing a red glow on the wires. If the voltage rises above this value the neon continues to glow; if it falls below, the discharge extinguishes. As there is no concept of rectification in this double electrode system the neon will light with ac or dc excitation of

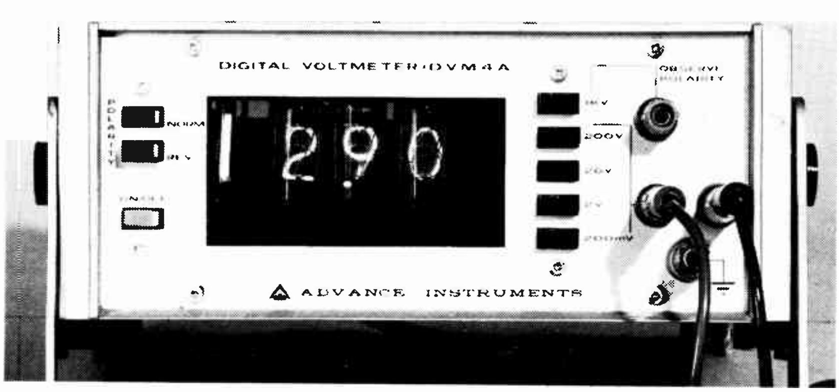


Fig. 1. This typical digital multimeter uses a neon-indicator tube display.

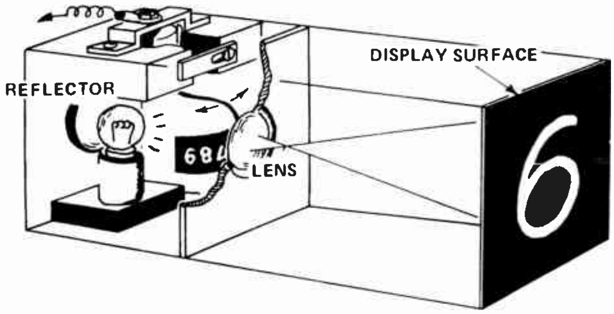


Fig. 2. The moving-coil numerical character display is based on an optical projection system. The numbers are on a graticule which is rotated behind a lens by a meter-line movement. (Drawing courtesy of Electronics Australia).

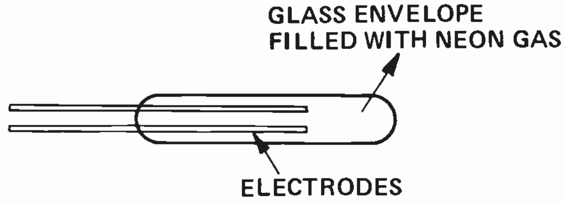
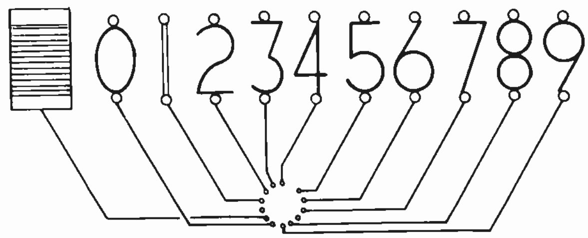


Fig. 3a. Construction of the basic neon lamp. b. How a multi-digit neon indicator is constructed from an anode grid and ten separate character-shaped cathodes.



sufficiently high voltage magnitude. Single neon indicators are used extensively for "mains-on" indication in instruments, power points and appliances, in which case a series resistor is added to obtain operation at 240 V.

The neon-indicator tube, developed from the basic neon lamp, incorporates 10 cathodes (when numbers are to be generated; letters and other symbols are available) one for each 0-9 number, which are stacked on top of each other behind a fine mesh. Each is insulated from the others and has a connection lead brought out through the glass envelope as shown in Fig.3b. The mesh acts as a common anode electrode for whichever cathode is selected. The tube displays just one of its number set. Non-energized grids remain dark and are unseen because they do not glow.

Numerical neon-indicator tubes are made such that the numbers appear either at the side of the glass cylinder or at the end. Character sizes ranging from 10 to 50 mm are available. This form of display has remained popular for reasons of the very acceptable readability, nicely shaped character format and low-costs. They require a relatively high voltage supply (180 Vdc is typical) and are not as robust as the solid-state devices described later.

The format and connections of a typical neon-indicator tube are illustrated in Fig.4. Note that only one input drive signal is required to energize any particular display character. The majority of all other displays in use require several inputs to be energized in order to produce the desired character. We will see later, however, that the amount of decoding circuitry needed for neon-indicator systems and the solid-state alternatives is similar.

It is possible to construct neon-indicators needing lower input — command voltages. In the Mullard Digitube, for example, the discharge remains on continuously. The trigger voltage, a 5 V level change, causes the discharge to transfer from an out-of-sight cathode to a visible one. This single-bit principle has been applied to a 10 step unit in, which individual separate numbers are illuminated as needed. This form of neon display has not become popular, probably because the numbers are arranged in a circle, giving small numerals which do not line up when several displays are used to form a multi-digit decimal number. (One early variety produced a dot glowing at the side of the numbers printed around a circle).

Neon-indicators radiate red light, which (more by chance than design) happens to be at a wavelength of reasonable sensitivity to the eye. Red is particularly suited to strong ambient daylight viewing.

MULTI-SEGMENT FORMATS

Each of the above displays uses characters generated by the application of a single signal that provides the character complete. This is said to be of simple format. An alternative method is to produce the character from individual segments or dots arranged to build-up the shape needed.

After the very active development period of the 60's designers and suppliers are now settling on the use of seven-segments, hexa-decimal 7 by 4 dot and 7 by 5 dot matrix formats.

Seven-segment format — This is the simplest and most used composite matrix method. It consists of seven equal-size bars placed to form the 0 through to 9 series of numbers. Several distinct alphabetical characters and a

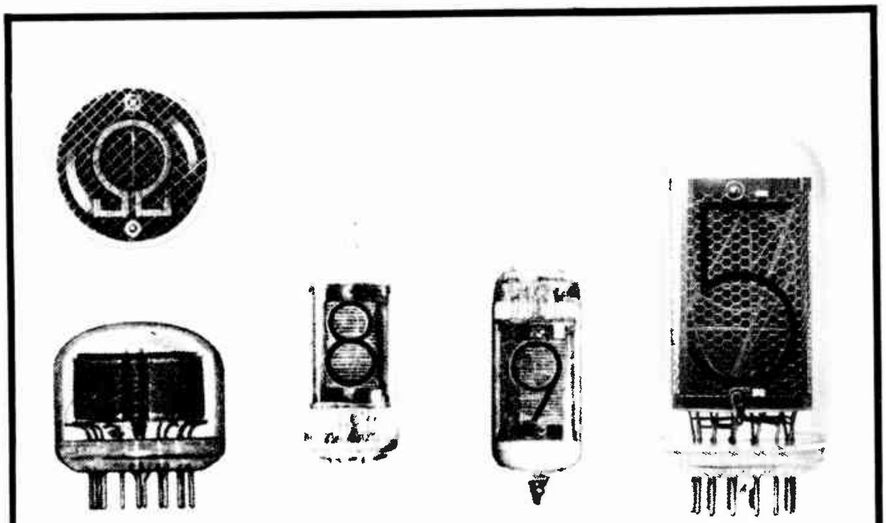


Fig.4. Different types of neon display.

minus sign are also possible. The appearance of seven-segment numbers and letters is as shown in Fig.5. This system is based upon a stylised figure of eight. Of particular note is the requirement that the individual characters are generated with different combinations of bars being illuminated.

Methods for illuminating a bar include separate filaments for each, separate incandescent bulbs, luminescent phosphors lit by filaments, light-emitting diodes (LEDs) and liquid crystal indicators – more of these later.

Hexadecimal format – these rely on the formation of a character by illumination of the necessary dots (or small squares) of a 4 by 7 dot matrix. Figure 6 gives the appearance of number characters generated this way. Note again the need to energize selected positions to provide the required character.

Alpha-numeric matrix format – the above 7 by 4 matrix is limited in that whilst it can generate all numbers, it cannot provide all 26 alphabetic characters. If the matrix size is increased to 7 by 5 the full 36 alpha-numeric characters can be generated. Figure 7 gives the characters of the American Standard Code for Information Interface ASC11.

SOLID-STATE DISPLAYS

Incandescent lamps are very inefficient at converting electrical energy into radiant visible energy – conversion is generally only around 20-30 lumens per watt. Neon-indicators consume less power in general and deliver a brighter output but do require a high voltage that is not directly compatible with the now standard 0-5 Vdc TTL signal levels. The life and robustness of both filament lamps and neon devices is also far from ideal. The breakthrough came several years ago when light-emitting diodes (LEDs) were developed.

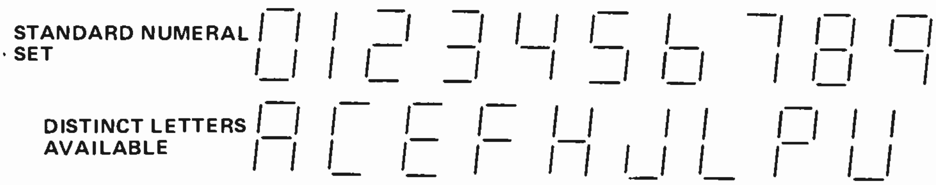


Fig. 5. Format of seven-segment numeric and alphabetic characteristics.

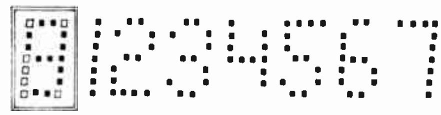


Fig. 6. Typical format of characters of the hexadecimal system using a 4 by 7 dot matrix.

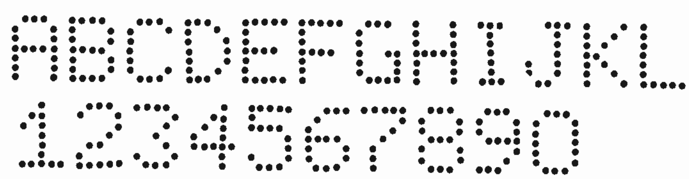


Fig. 7. Alpha-numeric characters as generated by 5 x 7 dot matrix.

Light Emitting Diodes – LEDs are semiconductor junctions (formed by the same processes used to make solid-state signal diodes) which emit radiation from the junction when current is passed through it. The basic materials used are gallium arsenide phosphide GaAsP and gallium phosphide GaP.

This form of light source generates relatively narrow wavelength energy centred on red yellow or green colours. (Typically 635 nm, 583 nm and 565 nm wavelength respectively) with high luminous efficacies of 140, 460 and 610 lumens per watt. Compare these efficacies against that for a typical tungsten filament lamp of 20 lumens per watt. The term efficacy should not be confused with efficiency. Efficiency is the percentage of radiant power compared to input power whereas efficacy refers to the effectiveness of the *radiant* power produced in stimulating the eye. For example an LED producing infra-red radiation will have an efficiency of say 3% but an efficacy of zero.

The high efficacy of LEDs means reduced power supply requirements, and high visibility is obtained even when LEDs are driven via a resistor directly from TTL.

Another feature of LED sources is the high speed of response – 100 ns is typical. The operating voltage is nominally 2 V and current requirement varies around 20 mA.

Single and multiple format LED displays are now available in a wide variety of forms and they are the most

This calculator from Hewlett Packard displays calculations and instructions from the calculator to the user by means of a full alpha-numeric 5 x 7 dot matrix display.

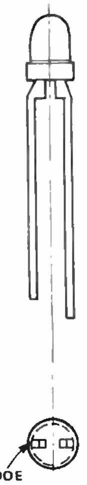


Fig.8a

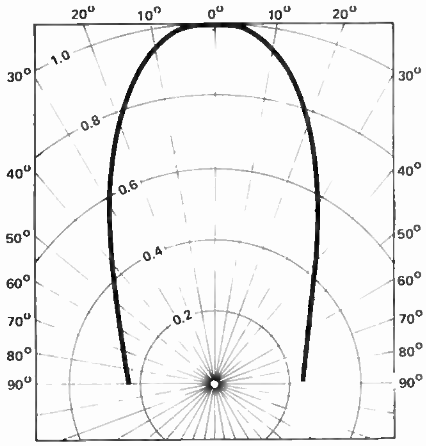


Fig.8b

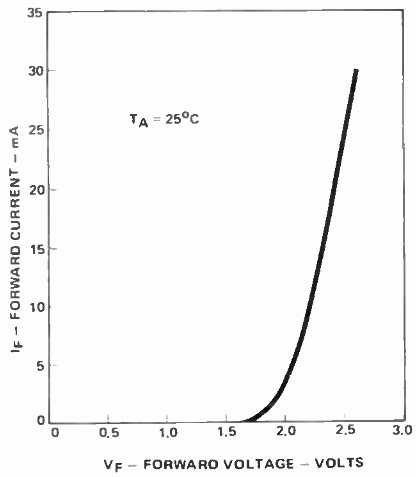


Fig.8c

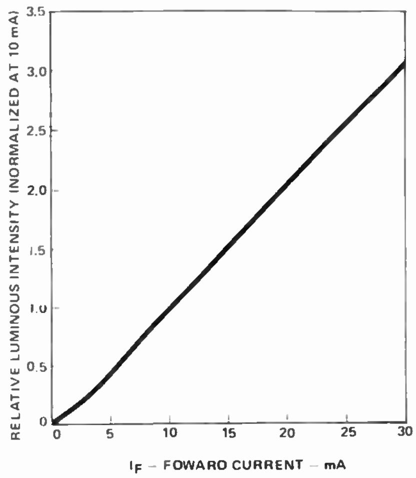


Fig.8d

Fig.8e

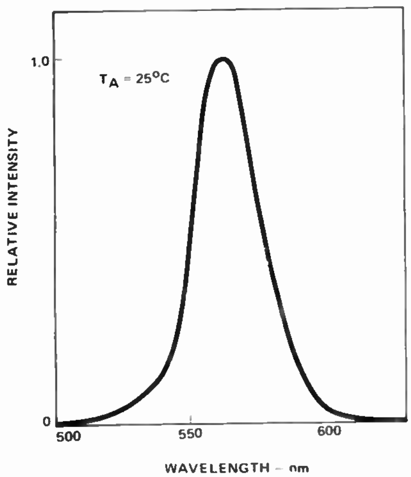


Fig.8f

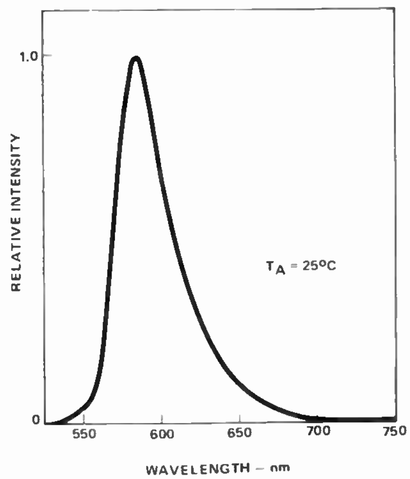


Fig.8g

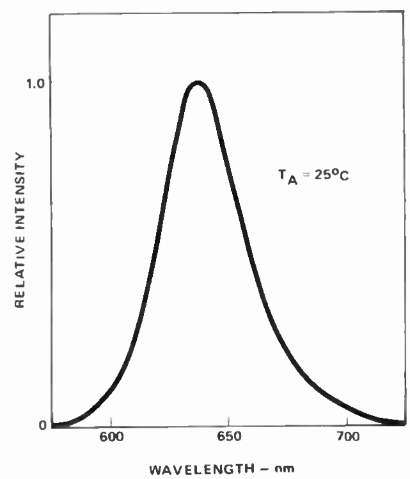
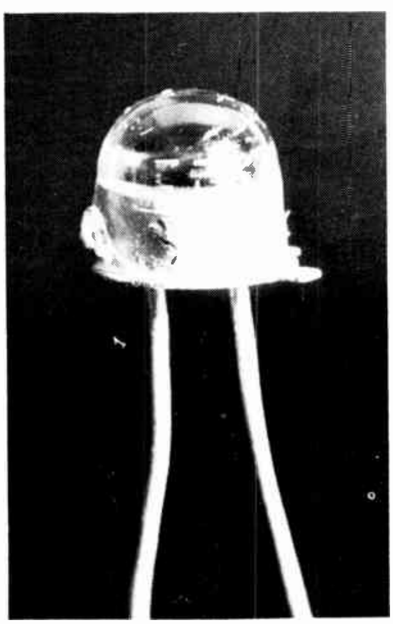
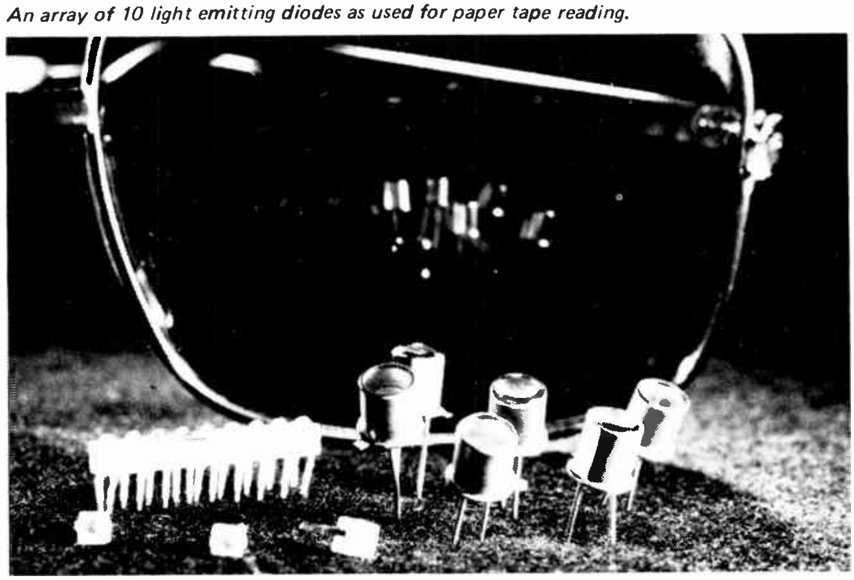


Fig. 8 Characteristics of the HP 5082 mini-LED series lamps.
(a) the shape of the lamp; (b) relative luminous intensity versus beam angle; (c) Forward current versus forward voltage. (d) Relative luminous intensity versus forward current; (e) Intensity versus wavelength of green lamp; (f) Intensity versus wavelength of yellow lamp; (g) Intensitv versus wavelength of red lamp.



◀ *A typical LED indicator lamp.*

An array of 10 light emitting diodes as used for paper tape reading.



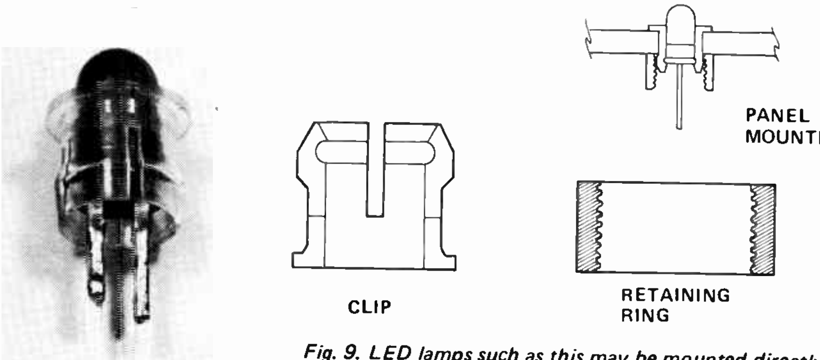


Fig. 9. LED lamps such as this may be mounted directly onto a PC board or onto a front panel by means of the clear plastic clip.

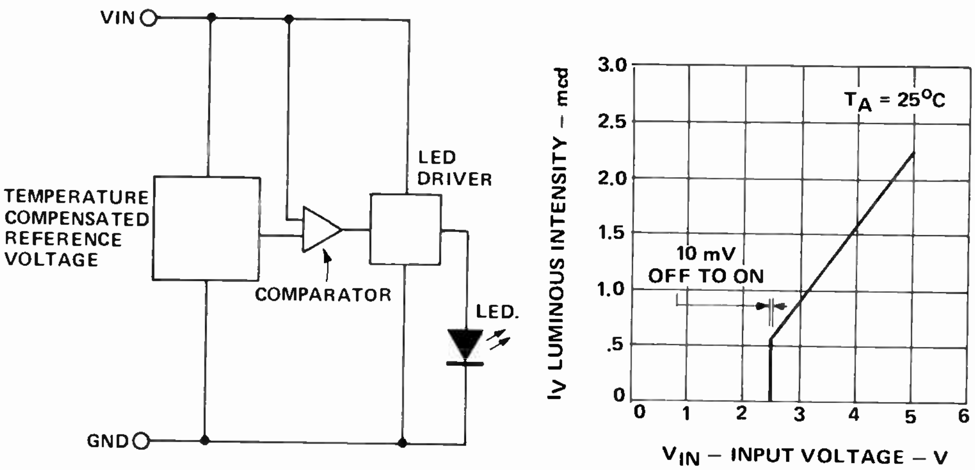
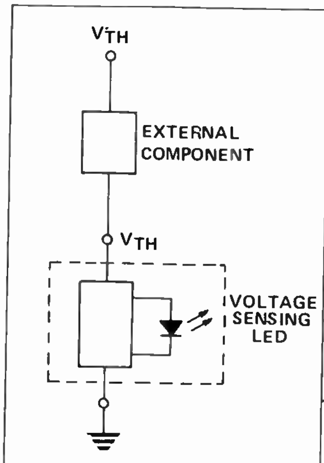


Fig. 10. LEDs with integrated voltage sensing amplifiers turn on when the applied voltage exceeds a built-in value.

LEFT: schematic. RIGHT: luminous intensity versus input voltage. BELOW: ways of increasing threshold voltage.



EXTERNAL COMPONENT	V'_{TH}	$TC = \frac{\Delta V_{TH}}{\Delta T_A}$ (mV/°C)
SCHOTTKY DIODE (HP 5082 – 2835)	V_{TH} + 0.45V	−2
P – N DIODE (1N914)	V_{TH} + 0.75V	– 2.5
LED	V_{TH} + 1.6V	– 2.9
ZENER DIODE	$V_{TH} + V_Z$	−1 + ZENER TC
R SERIES RESISTOR	$V_{TH} + I_{TH}R$ NOTE 1.	NOTE 2.

NOTES:
1. I_{TH} IS THE MAXIMUM CURRENT JUST BELOW THE THRESHOLD, V_{TH}. SINCE BOTH I_{TH} AND V_{TH} ARE VARIABLE, A PRECISE VALUE OF V_{TH} IS OBTAINABLE ONLY BY SELECTING R TO FIT THE MEASURED CHARACTERISTICS OF THE INDIVIDUAL DEVICES (E.G., WITH CURVE TRACER).
2. THE TEMPERATURE COEFFICIENT (TC) WILL BE A FUNCTION OF THE RESISTOR TC AND THE VALUE OF THE RESISTOR.

used display medium. Figure 8 gives the various data of a typical unit. Figure 9 shows how a single lamp can be mounted in practice.

Developments arising out of the basic single LED lamps are units incorporating an integrated resistor (for direct TTL connection) those having an integrated voltage sensing amplifier (Fig.10) which provides a lamp that triggers on or off as the input level passes up or down through a 2.5 V level and the opto-electronic relay or isolator discussed in a previous section. Hermetically sealed units and military approved units that will operate from -65°C to +100°C with very high reliability over a life measured in years of operation are also available.

Given a matrix of LED lamps it is quite practicable to generate numbers and characters by what is called an addressable system in which decoding logic decides the diodes to be illuminated. LED character displays are marketed as single unit 7 segment modules and as 4 by 7 and 5 by 7 dot matrices. Integration has gone as far as incorporating a complete decade counter stage (Fig. 11), with the necessary decoders, buffer amplifiers and LED display all integrated on a single LSI unit. As LED manufacturing techniques are the same as conventional integration methods it is possible where large quantity production is economic, to integrate the display with the circuitry — examples are to be found in some styles of IC wristwatch.

Seven segment LED displays have the eight diodes placed on a common transparent GaP substrate. (The eighth diode provides a decimal point). A typical single unit is shown in Fig.12 — they are available in red, yellow and green colours. The 7.6 mm letter size is visible at 3 m; a larger 11.0 mm size can be readily seen at 6 m. Another series, shown in Fig.13 includes an integral optical magnification technique that provides improved readability for low drive power (1 mW per segment). These are available as 3, 4 and 5 character units which are mechanically compatible with standard printed-circuit board hole spacings.

To meet the demand for portable calculators manufacturers also supply special units with 8 or 9 digits mounted on a small plug-in printed-circuit board.

The range of dot generated character displays is also extensive. A 39 mm high character is available that can be read from 20 m. This, as can be seen in Fig.14, is based upon a large size

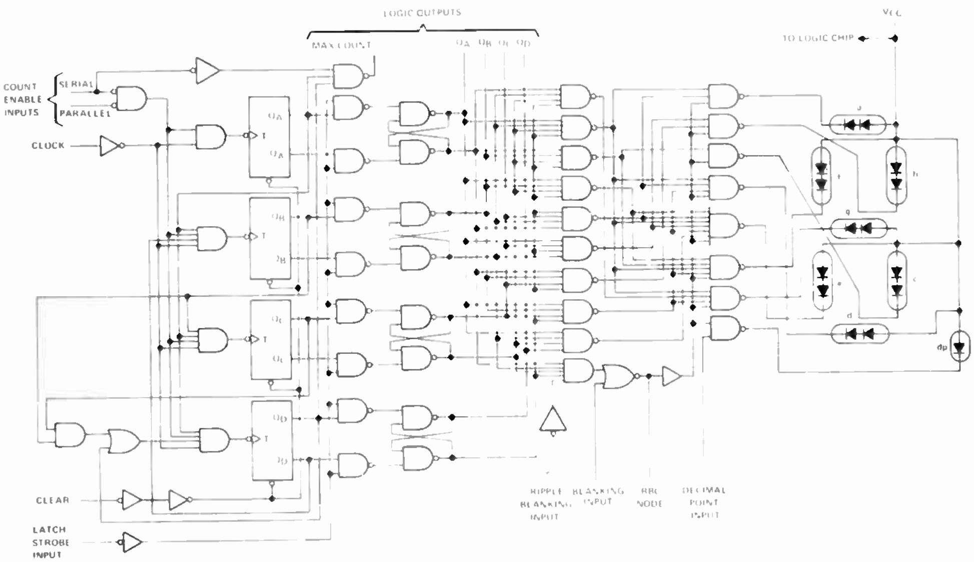


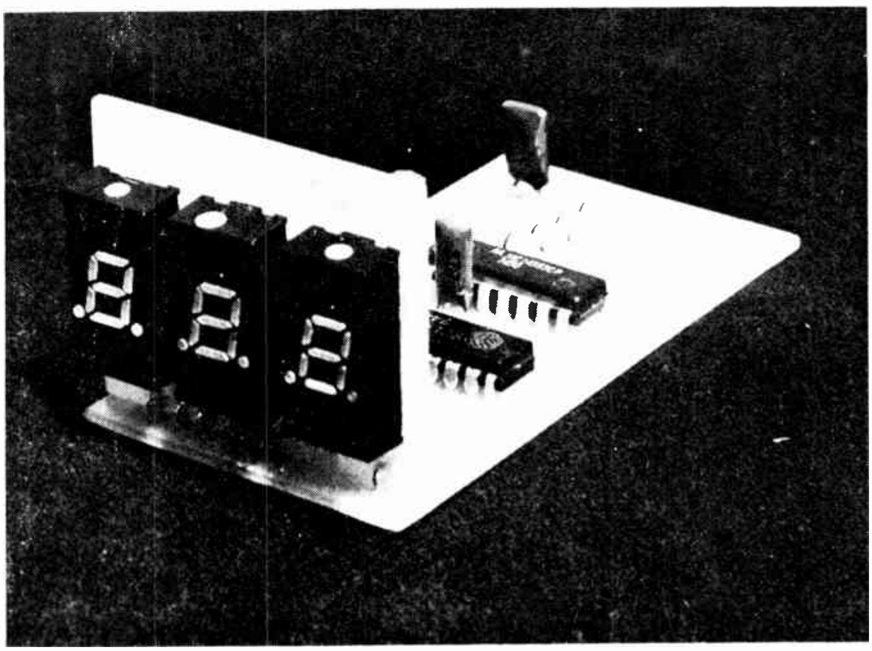
Fig. 11. The Texas Instruments TIL306 display integrates all the logic of a complete decade counter onto the same chip as a 7-segment display. The circuit shown is the schematic of the device.

5 by 7 dot matrix and includes the decoder/driver unit for the most commonly used BCD code – the 8421-logic input (decoders are discussed in the next part). Dot matrix displays with characters as large as 45 cm height are produced. These, however, are not usually solid-state but use electromagnetic drives to rotate reflective dots into or out of the viewing aperture. Such units, given adequate ambient light, are visible at 300 _ m. Multi-digit dot matrix solid-state displays are also made.

Liquid Crystal Displays. Although LED displays consume little power compared with earlier filament displays very little of the power used is actually transmitted as radiant energy. Efficiencies of visible diodes are typically only 0.1%! Thus an LED display often consumes considerably more supply power than the rest of the associated digital system. Indicators of all types, except liquid crystal, require about 300 to 500 mW per character (all segments illuminated).

The power requirements of the display could be reduced considerably if the circuit could switch available ambient light rather than actually generate light. Naturally such a method will only work when ambient light is available.

In the dark, displays which generate radiation would still be required. Displays are available which do switch ambient light. They are known as liquid crystal displays and by virtue of their mode of operation consume very little power.



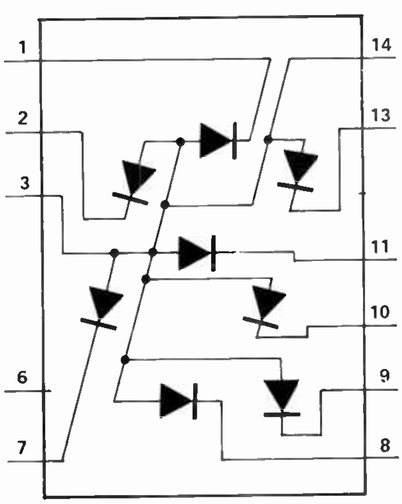


Fig. 12. ABOVE: three seven-segment LED (DL704) displays mounted on a common PC board. LEFT: internal diode positions for a right-hand decimal point seven-segment display module.

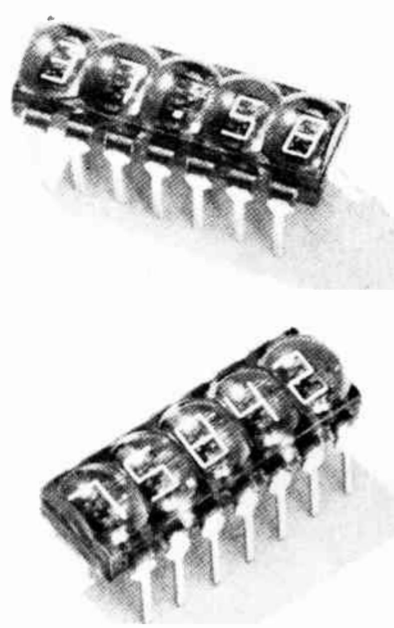
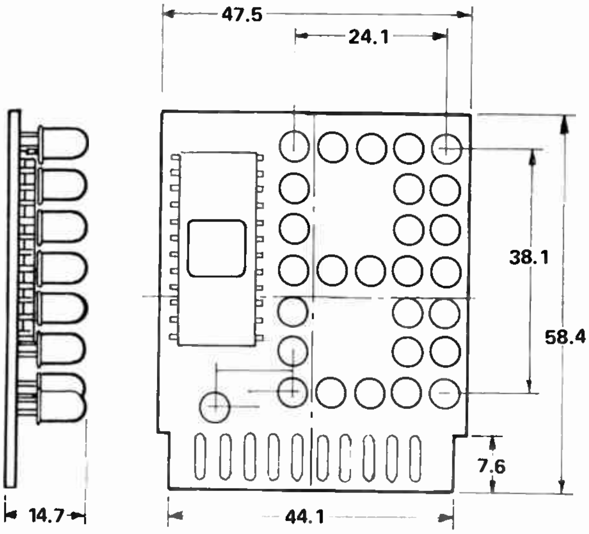


Fig. 13. Some seven-segment displays suitable for calculators etc are assembled in groups and have plastic lenses to increase character size.



MECHANICAL DETAILS

PIN	FUNCTION
1	V_{CC}
2	DP
3	X1
4	X2
5	X4
6	X8
7	GROUND
8	NC
9	NC
10	V_{LED}

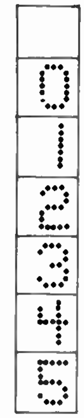


Fig. 14. A large size (38 mm character) alpha-numeric display constructed from discrete LEDs and a decoder/driver IC. (HP5082-7500).

FORMAT

Basically liquid crystal displays consist of a minutely thin layer of liquid-crystal material placed within two thin glass covers. The glass covers have transparent electrodes deposited on them in the shape of the characters or segment needed. This is shown in Fig. 15. With no excitation the whole unit appears transparent, for the liquid crystals remain stationary allowing light to pass through virtually unattenuated, that is, no light is reflected. When an alternating voltage (40-1000 Hz) is applied to the electrodes forming the character shapes, the resultant electric field causes the liquid layer to become turbulent, scattering light between the confines of the deposited areas. The display then shows an optically dense character because the ambient light is reflected. In simple terms application of an input signal causes the liquid crystal in the vicinity of the transparent electrodes to act like a mirror.

The power requirement for the circuit driving liquid crystal displays is around 20μW per segment (compare this with the lowest 100 μW per segment but more usually 20 000 μW for LED characters). Response is not as fast as for LEDs – 20 ms rise-time and 100 ms fall-time, but that is not a serious shortcoming in visual observation applications. In some instances faster response is needed – consider, for example, the use of photographic recording of a character display. With LED displays the display, when being photographically recorded, can be cycled considerably faster than the eye can follow.

Liquid-crystals are the most recent solid-state display to be developed and it is still too early to state with certainty if they will eventually compete seriously with LED techniques. At present the life of the display is inferior to LED units. Although manufacturers quote 10 000 hours minimum life (just over a year) experience has shown that units often fail after only a 1000 hours.

Seven segment displays are also made using neon lamps, self contained filaments and separate incandescent bulbs. It is to be expected that these will not be in use in new designs of the future for the price alone of solid-state devices will usually undercut the available alternatives.

Regardless of the display used it is necessary to decode the binary logic of digital circuits into a code suited to illuminate the required number and combination of characters in the system used. The next section will look at the schemes used and at more efficient methods of driving multiple character displays.

REFERENCES

The use of solid-state displays is straight-forward in simple applications. In each case, however, design information is vital to ensure that the displays are used within ratings. Advanced display design has become a high-level art and generally Application Notes are the essential guide to their successful use.

Hewlett-Packard produced an "Opto Electronics Designer's Catalog" in 1973 and 1975. The former included several applications notes, the latter a list of the range of relevant application notes now available from HP: both contain a wealth of practical data.

"Digital display systems", written by E.G. Breeze and available as Fairchild Application note 212/1, 1972 is also worthwhile having.

Many other manufacturers – Texas Instruments, RCA, National, Hawker Siddeley Electronics, Monsanto, Mullard, Atron, Litronix, Siemens – also provide service data that gives practical advice on how to use their display products to best effects. ●

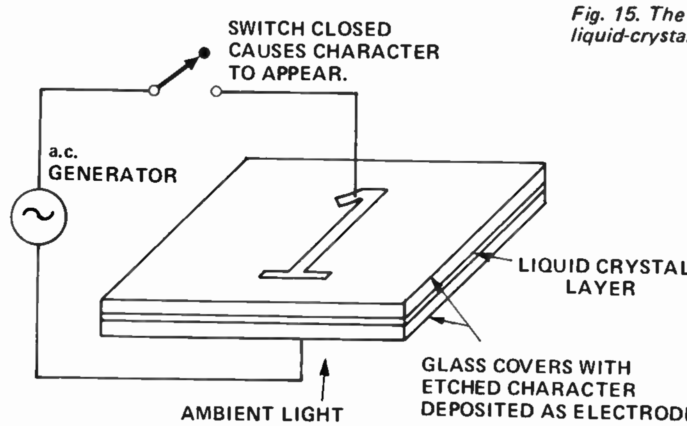


Fig. 15. The construction of a liquid-crystal display.

26 Code converters and display systems

IT IS both necessary and convenient to transfer data between sub units of a digital system, by means of some kind of code. We have seen in the previous section how counting stages are sometimes arranged to count in BCD (Binary Coded Decimal), and how this form of code must then be 'converted' into another form that is suitable for the particular kind of display device used. Thus codes and code converters are of great importance in digital instrumentation.

There are a multitude of digital codes in use for communication, data interchange, and for numerical manipulation and display. Although many of the earlier codes used have now been discarded, there are still dozens in use. In this section we will not discuss codes like ASCII, Baudot, Excess 3 etc, which are computer and communication codes, but restrict ourselves to those codes and converters which are concerned with counting and display.

The main counting codes used in instrumentation are binary and BCD. Octal and Excess 3 are other counting codes used in computers but seldom in instrumentation. Converters are needed to change from any one of these codes to any other, and between any one code and decimal or vice versa. In addition the counting code in use needs to be converted into a form suitable for driving particular kinds of display (seven-segment, dot-matrix and neon tube, etc).

For example we find converters for binary to BCD, binary to decimal, BCD to seven-segment, and BCD to decimal, to mention just a few of the possible combinations. Figures 1 and 2 show two such variations.

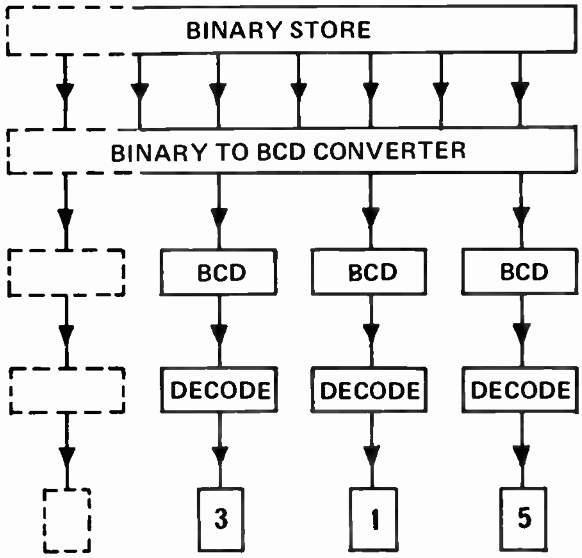


Fig. 1. Basic arrangement of binary to BCD converter and display decoders.

BINARY TO BCD CONVERSION AND VICE VERSA

To convert binary to BCD a common method is to set a binary counter – see Fig. 3 – to the desired number either by direct counting upward from zero, or by transferring the value across from another stage with a parallel converter. Clock pulses are then fed into both this binary counter, now set to count back down to zero, and to an up-counting BCD counter. A detector senses when the binary counter reaches the 0000 state upon which any further changes in the count state of both units are inhibited. The BCD equivalent of the binary number is now held in the BCD counter.

At this point the BCD number is cleared into a store or is available for any other system need, the binary counter is reloaded and the process repeated to convert the next number from binary to BCD.

The reverse, BCD to binary, is accomplished in the same manner except this time the roles are reversed as shown in Fig. 4. The BCD counter is set to the desired number, the clock, when enabled, clocks the BCD counter down to zero and at the same time the binary counter upward. When the BCD counter reaches zero state its outputs logically inhibit the clock input to both counters. The process is repeated for each new number after clearing and resetting the two counters to the correct starting conditions.

This serial method is fairly slow and a much faster method is to use logic gates in a parallel arrangement. The

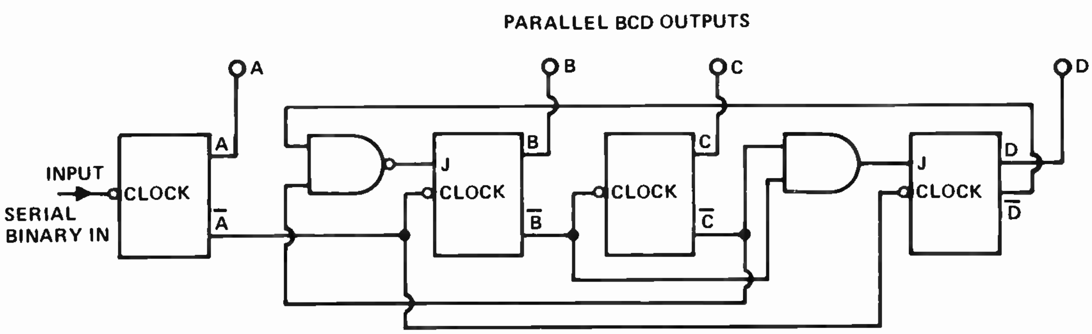


Fig.2. An asynchronous BCD counter provides a parallel BCD coded output which corresponds to the number of pulses fed to the input.

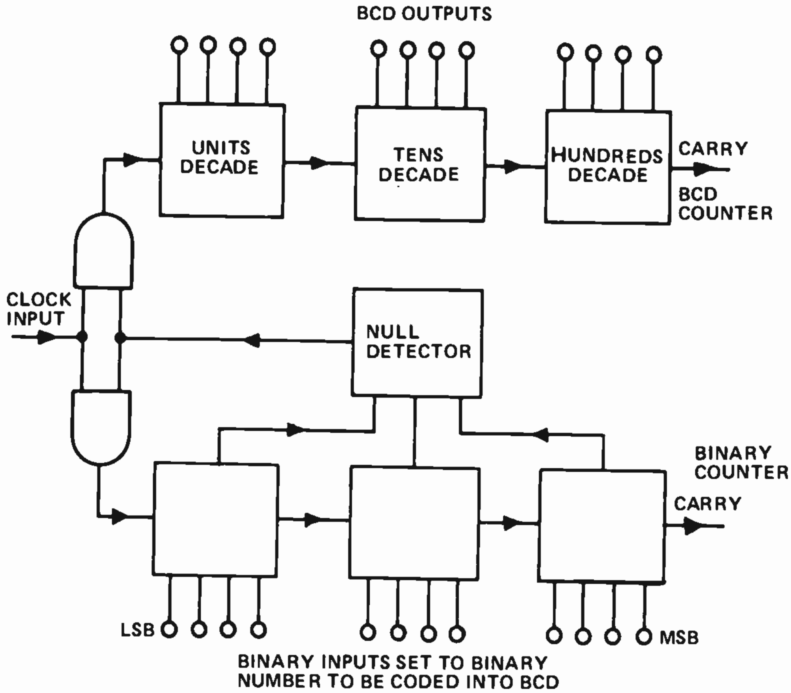


Fig.3. System required to convert a binary number to its BCD equivalent.

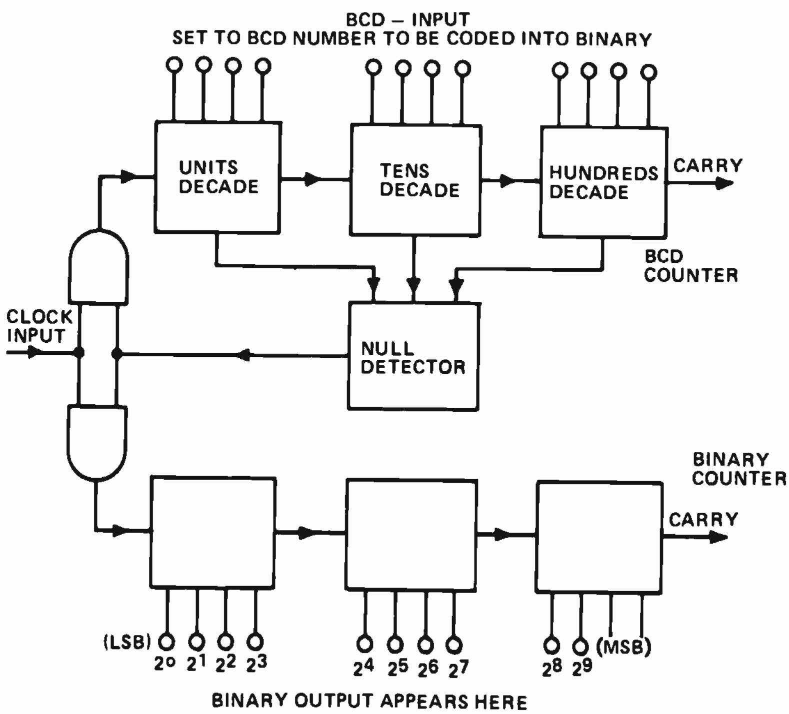


Fig.4. To convert a BCD number to its binary equivalent the same circuit as Fig.3 is used but the BCD number is now the one pre-loaded.

parallel method works well for BCD to decimal and other cases but requires innumerable connections when set to convert decimal or BCD back to binary. It is little used in this reverse mode.

BINARY CODED DECIMAL TO DECIMAL AND OTHERS

Each BCD stage stores its decimal number as some form of binary code using four bits. The 1-2-4-8 weighting is the most usual form used but other codes such as the 'excess three' and the 'Aiken' variations, are also used. Thus, each of the digit values 0 to 9 (in decimal) is represented by four lines, each having a '0' or a '1' state. This is demonstrated by studying the truth table for equivalent BCD and decimal states given in Fig. 5.

When the output must finally appear as decimal indication, it is necessary to energise the display or character printer segments appropriately. Displays, such as the neon tube or the columnar style — see previous part — require one output for each of the 0 – 9 numbers. From the truth-table of Fig. 5 it can be seen that to energise, say, a decimal '4' output we must set up a logic gate that provides an output when the BCD state is 0100. For '8' we need a BCD state of 1000. It is not possible to totally economise by using only '1', as the logical indications for this leads to ambiguities between numbers. On the other hand there is no need to arrange for all code sequences and bit combinations as that introduces redundancies using up extra unnecessary gates.

By careful design, and the use of inverters to invert '0' to '1's, it is possible to find a minimum number of AND gates and interconnections that will produce the 10 decimal states (0 to 9) as distinct outputs from the four-line (A B C D) outputs of the BCD stage. One such scheme is given in Fig. 6. Thus, by the use of logic gates alone we can provide a parallel code conversion from BCD to decimal.

Getting from BCD to a format suitable to drive a seven-segment display requires more gates, see Fig. 7, but the technique is basically the same. A decoder suited for the BCD to decimal requirement of a neon tube is quite unsuited to drive a seven-segment display. As both these and other conversions are in great demand they are available as simple ICs. Further, in some options the decoding logic is integrated onto the same chip as the BCD counter stage.

In practice the need to understand the internal operation of the decoder arrangement rarely arises, for the ICs are clearly marked with the connections to be made — it is just a case of making correct connections between the counter-stage chip, the decoder chip, and the display.

THE NEED FOR DISPLAY DRIVER STAGES

The power levels available from decoder stages are rarely adequate for direct drive of display units. A buffer stage which raises the power level is normally required. Again, these are generally incorporated into the decoder IC stage. Such integrated units are known as decoder/drivers. Different displays, even of the same format, require differing power needs so it is important to select decoder/driver stages suited to the display being used.

The buffer stage of a decoder/driver obtains current (or voltage gain) by the use of a transistor stage such as a Darlington pair or an emitter-follower circuit. A method recommended for driving seven-segment fluorescent displays is shown in Fig. 8 – these displays require high voltage drives.

Sometimes the need arises to drive displays from special-purpose one-off decoding circuits. In such cases a suitable driver stage is obtained by using standard IC inverter chips. (Discussed in Part 23).

THE ASSEMBLED BASIC COUNTER WITH DISPLAY

An illustration of the actual construction of a complete counting stage is to be found inside the digital clock shown in Fig. 9a. The display – neon tubes in this case – is arranged to appear where required by using end-connected tubes which plug into the main printed-circuit board as shown in Fig. 9b. Immediately behind each neon tube is the decoder/driver IC which, in turn, is driven by the counter IC located behind that again. On the circuit diagram these appear as shown in Fig. 9c and on the component overlay of the PC board as in Fig. 9d.

The Texas Instruments TIL 306 LSI is an example of an IC unit that incorporates the counter, decoder, driver, and display in one package. No doubt this will be the trend as it provides reduced assembly costs and smaller packaging. It should not be unrealistic to expect complete single-IC, multi-stage counter units to be in general use before long.

LATCHES

Direct coupling of the display to the counter stage results in a continuously changing display value. If the input is sufficiently dynamic it is awkward, if not impossible, to read values. Addition of a latch overcomes this by sampling and storing the count to be displayed for fixed intervals whilst the counter continues to cycle.

This is achieved using a memory stage between the counter and the decoder/driver stage as shown in Fig. 10. This system, the digital end of a digital voltmeter, displays a steady value for a short period by transferring the instantaneous value of the divider stage (the counter) into a buffer-store stage. The transfer or updating process is initiated by a common display timing line which is actuated at appropriate intervals. Such latches are invariably placed between the counter and decoder stages.

Internally a latch is a bi-stable designed specifically for the purpose of storing and transferring the value of a digital bit. Integrated circuit units provide four such latches in a dual-in-line package – see Fig. 11, thus allowing the four line data from a BCD

B.C.D. Inputs				Outputs										Decimal equivalent	
A	B	C	D	0	1	2	3	4	5	6	7	8	9		
0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	Valid state of encoding
1	0	0	0	0	1	0	0	0	0	0	0	0	0	1	
0	1	0	0	0	0	1	0	0	0	0	0	0	0	2	
1	1	0	0	0	0	0	1	0	0	0	0	0	0	3	
0	0	1	0	0	0	0	0	1	0	0	0	0	0	4	
1	0	1	0	0	0	0	0	0	1	0	0	0	0	5	
0	1	1	0	0	0	0	0	0	0	1	0	0	0	6	
1	1	1	0	0	0	0	0	0	0	0	1	0	0	7	
0	0	0	1	0	0	0	0	0	0	0	0	1	0	8	
1	0	0	1	0	0	0	0	0	0	0	0	0	1	9	
0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	Illegal states that should not occur with BCD code.
1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	
0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	
1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	
0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	
1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	

Fig.5. Truth table relating four-line BCD with its decimal equivalents.

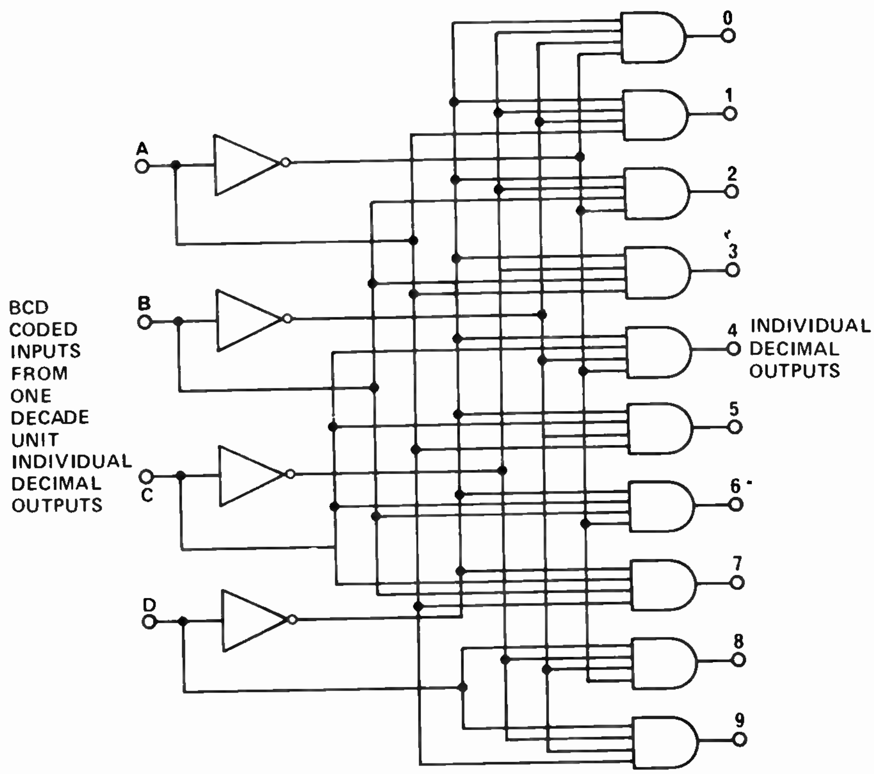


Fig.6. Logic arrangement for decoding BCD units into one-of-ten outputs.

counter to be sampled and stored by a single IC.

DECIMAL POINT

Facilities are usually provided in all displays to enable an input to energise a left-hand or right-hand point at each digit position. Obviously specific circuitry is required to energise just one of the total available in a multi-digit display in order to present the correct decimal number. The simplest arrangement is when the range switch of, say, a multi-range digital voltmeter decides the point to use. In autoranging voltmeters (etc) and in many calculators, solid-state switching is used to select the correct point position.

BLANKING

In normal writing practice we do not

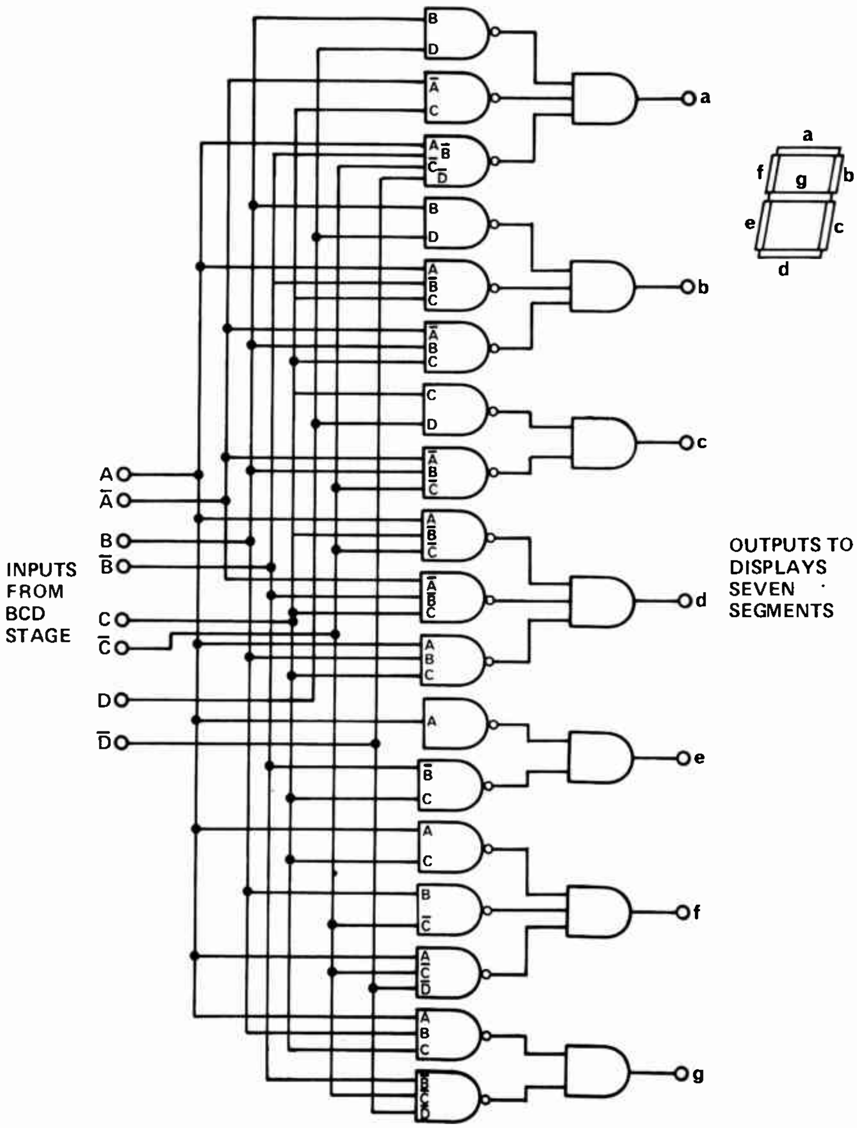


Fig. 7. Logic for converting BCD to seven-segment display format.

write the '0's that appear at either end of a number, for example, 0001357.0 as could be held in an eight digit display, is better presented as 1357 or as 1357.0. A facility is sometimes provided in display-counter systems that blanks unnecessary zeros. Leading zero suppression is performed within the decoder stages of Fairchild seven-segment decoders by connection of the ripple blank output RBO (ripple because each stage connects to the next) of the decoder stage to the ripple blank input RBI of the next lower decoder stage. Blanking of least-significant zeros is not usually included. The actual arrangement for blanking control varies from maker to maker. Fig. 12 shows a method using ripple blanking.

The blanking facility can also serve other purposes. It can, in certain applications, be used to blank-out illegal display values resulting from incorrect codes. The RBO output also provides a detection output indicating when the decoder stage is at the BCD zero state.

Blanking is also valuable as a way to save display power for it can be used to hold all displays off until there is something to display.

When no blanking input is provided it is also possible to blank the system by applying the spare BCD code states that result in no drive to the display. Yet another procedure is to disconnect the supply from the display itself.

INTENSITY CONTROL

Displays are usually manufactured to supply one value of output brightness. When brightness is to be tailored to

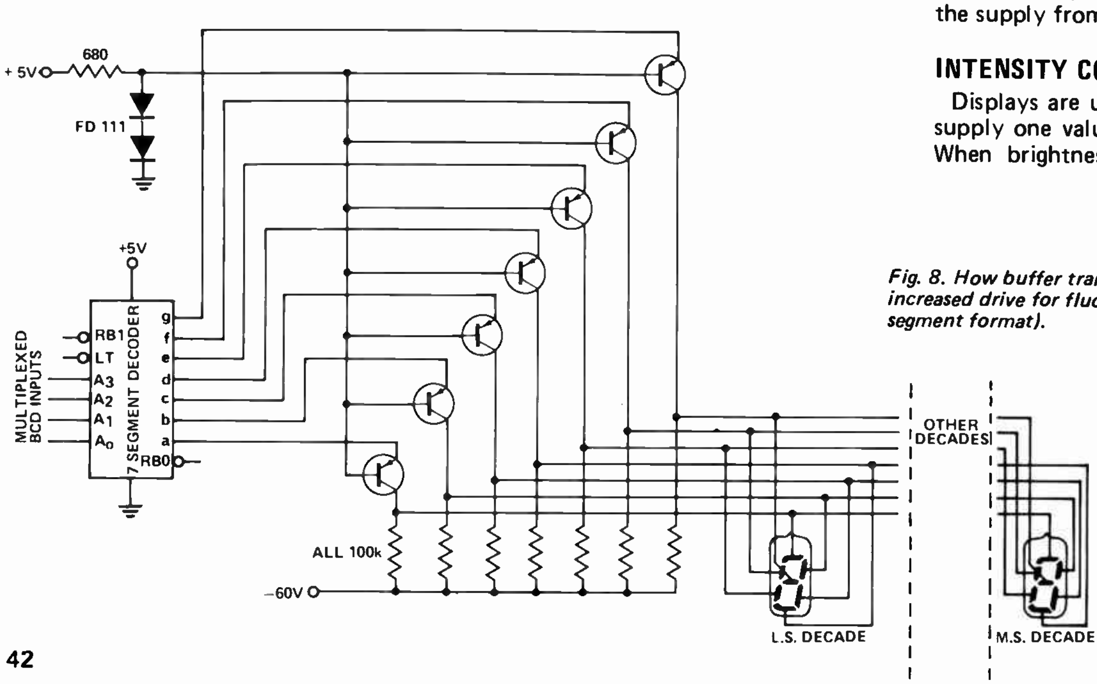


Fig. 8. How buffer transistors are used to obtain increased drive for fluorescent displays (seven-segment format).

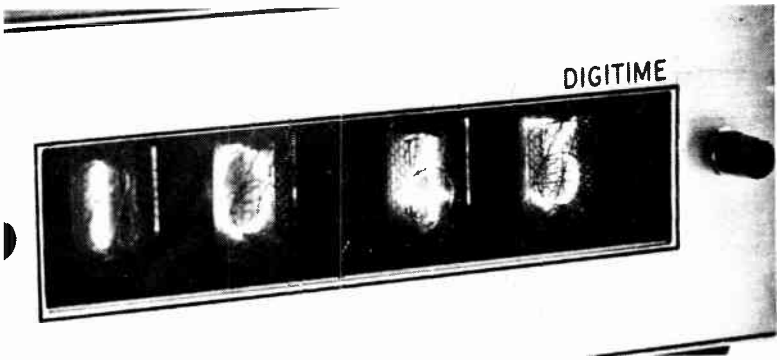
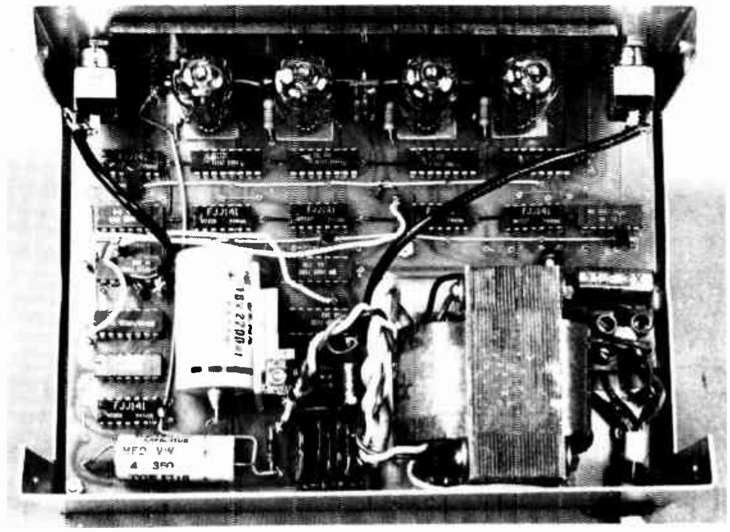
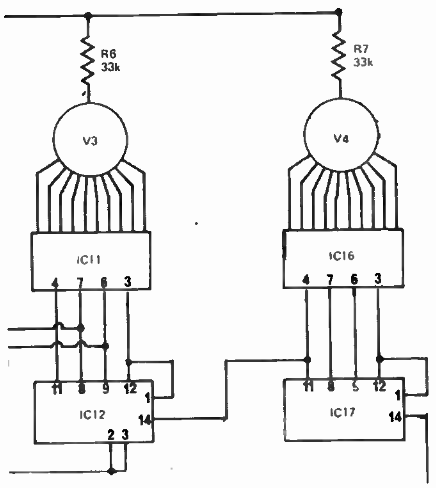


Fig.9. This digital clock illustrates use of counter stage subsystems.
(a) The frontal appearance of the Nixie display.
(b) Top view of component assembly.
(c) Circuit diagram of counter to display section.
(d) Component overlay. The V1 to V4 positions are the sockets for the Nixie tubes, behind are the decoder/drivers which have the dividers behind those again.

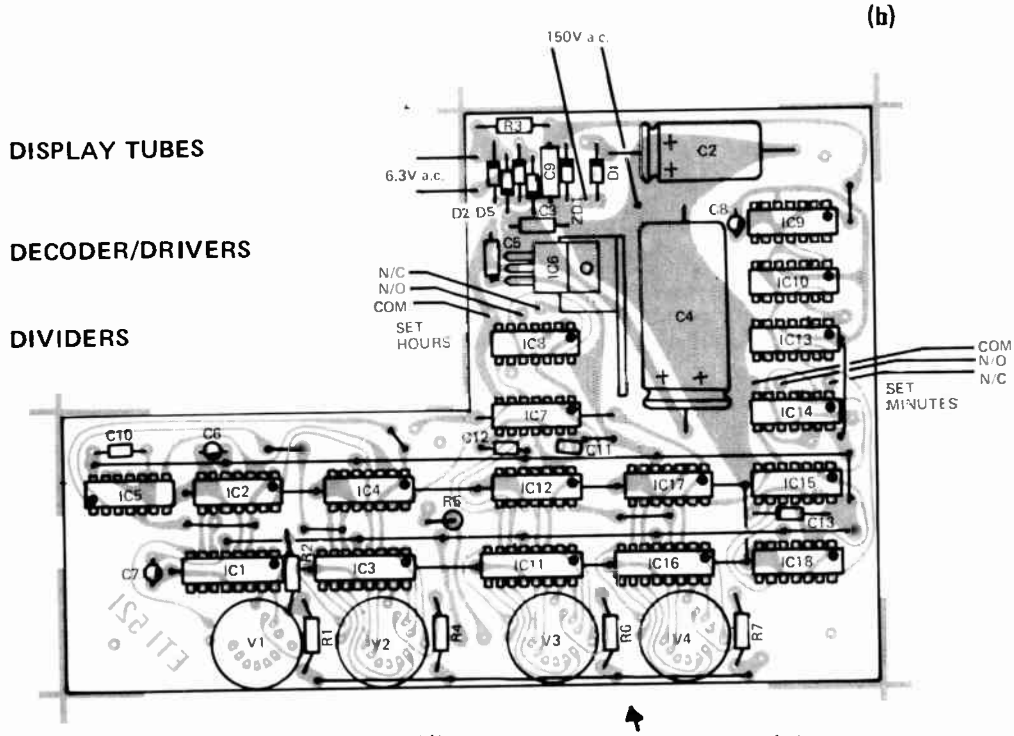
(a)



(b)



(c)



(d) CIRCUIT SHOWN IN (C)

particular ambient light conditions an appropriate kind of display can be selected that provides the desired luminance level. This however, does not always lead to a satisfactory choice when other considerations are taken into account.

Intensity of any display, however, can be controlled in a digital manner (that most desirable in digital systems) by turning the display on and off with an appropriate duty cycle (ratio on to off period). This is called pulse-duration intensity modulation. Provided the repetition rate exceeds 100 Hz the eye cannot detect that the radiation source is being modulated. Modulation may be achieved with any of the blanking methods given above. The schematic of Fig. 12 includes an intensity modulation facility.

With LED displays, intensity modulation can actually increase the apparent brightness. The human eye has a characteristic response to radiation that has greater sensitivity to the peak value of modulated light, rather the average or rms power. LEDs can be pulsed at high frequency with high peak currents because of their nanosecond response time. The net result is apparently higher brightness for a given amount of power.

STROBING OR SCANNING

Displays which generate characters in the 7 x 5 dot matrix or seven-segment formats require decoding logic which energises the correct dots or segments. If each character has its own decoder we would need 7 lines for each digit of a seven-segment display. And 35 lines for each digit of a 7 x 5 dot-matrix display!

Obviously a method is needed to reduce the number of lines and circuitry required for multi-digit displays.

One such method is called strobing where lines of dots or segments are illuminated sequentially. The 7 x 5 array can be either strobed as lines horizontally or as rows vertically as illustrated in Fig. 13. Each row is selected one by one in sequence and the appropriate diodes in the row energised. Provided each row is

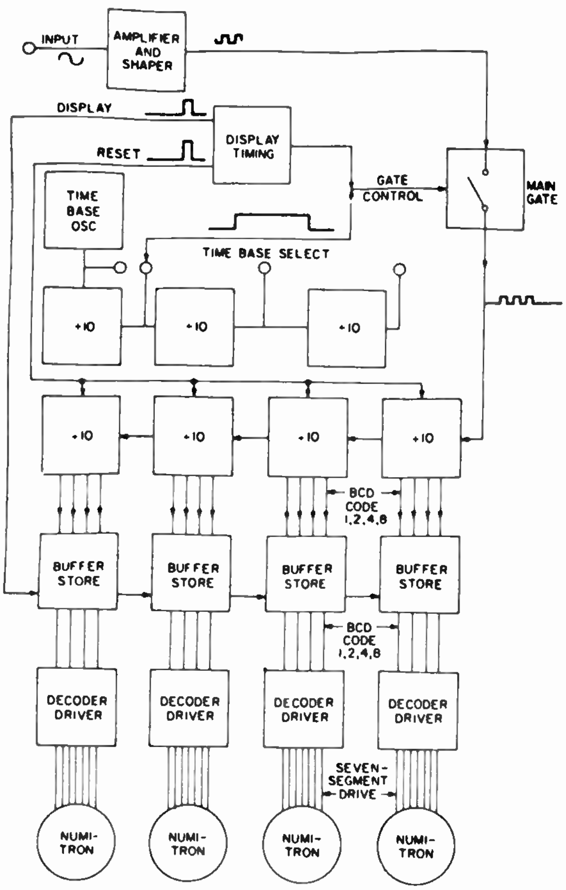


Fig. 10. Buffer stores are used to latch the display causing it to remain steady for selected periods. (circuit shown is a counter/timer using RCA incandescent displays).

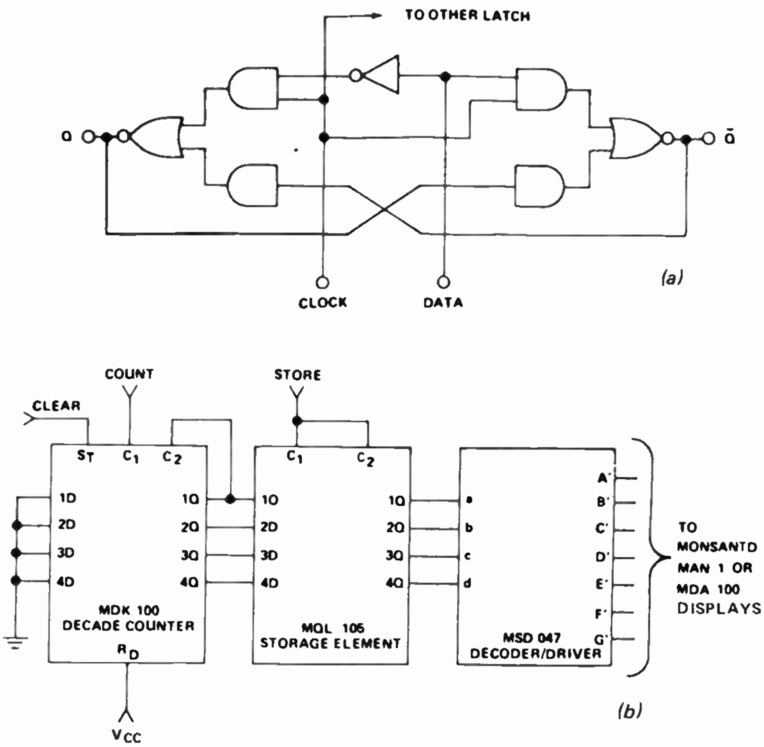


Fig. 11. Schematic (a) and pin connections (b) of counter, latch decoder system using the Monsanto MQL 105 four-bit bistable latch.

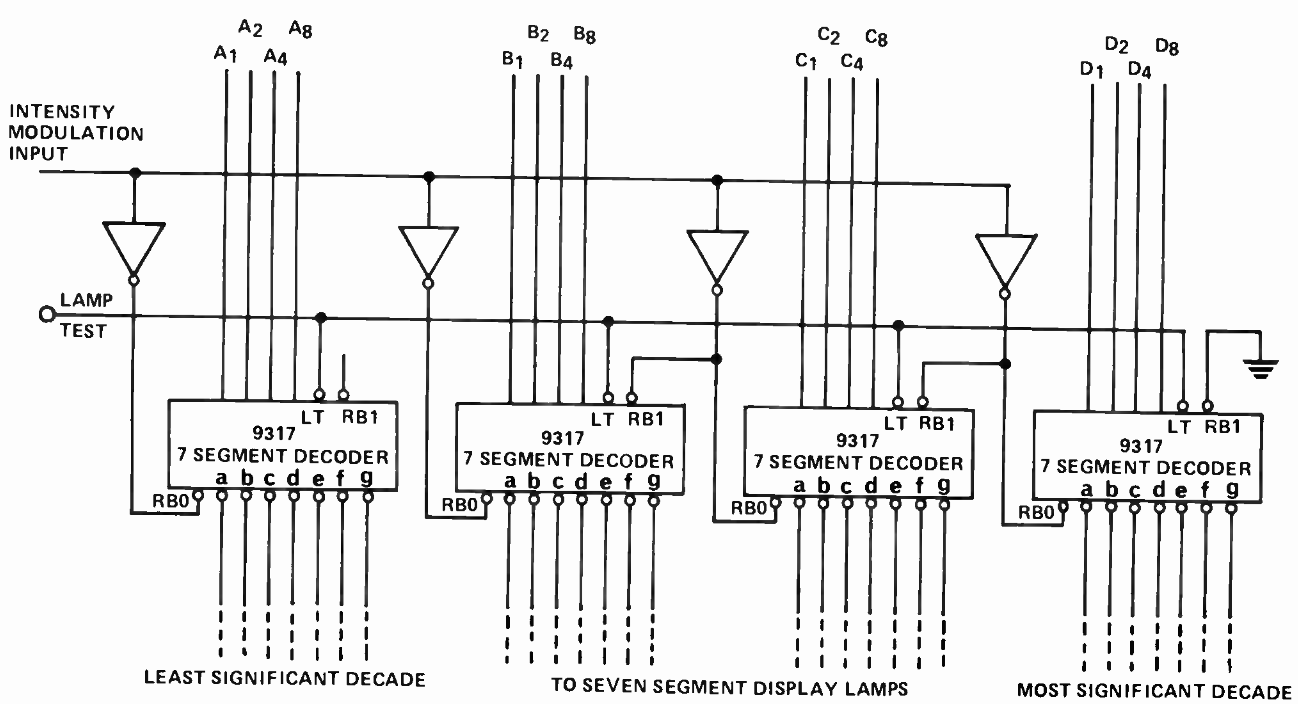


Fig. 12. Connections for ripple blanking in a four-decade display system. (RBO ripple blank output, RBI ripple blank input).

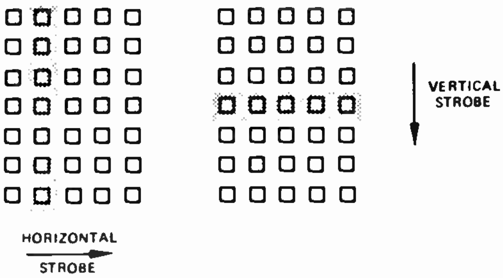


Fig. 13. Horizontal and vertical strobing of a 7 x 5 dot matrix display.

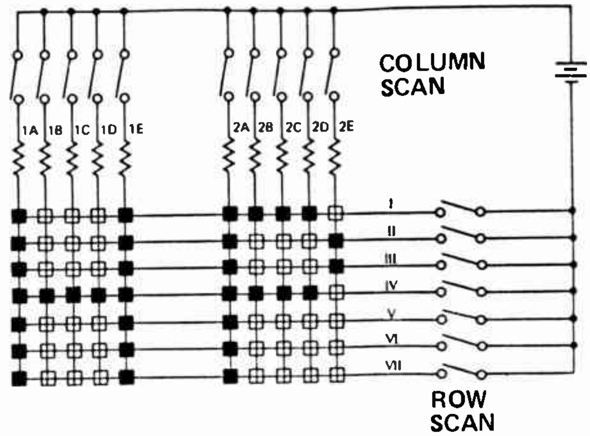


Fig. 14. Schematic of switches needed to address a 7 x 5 dot matrix.

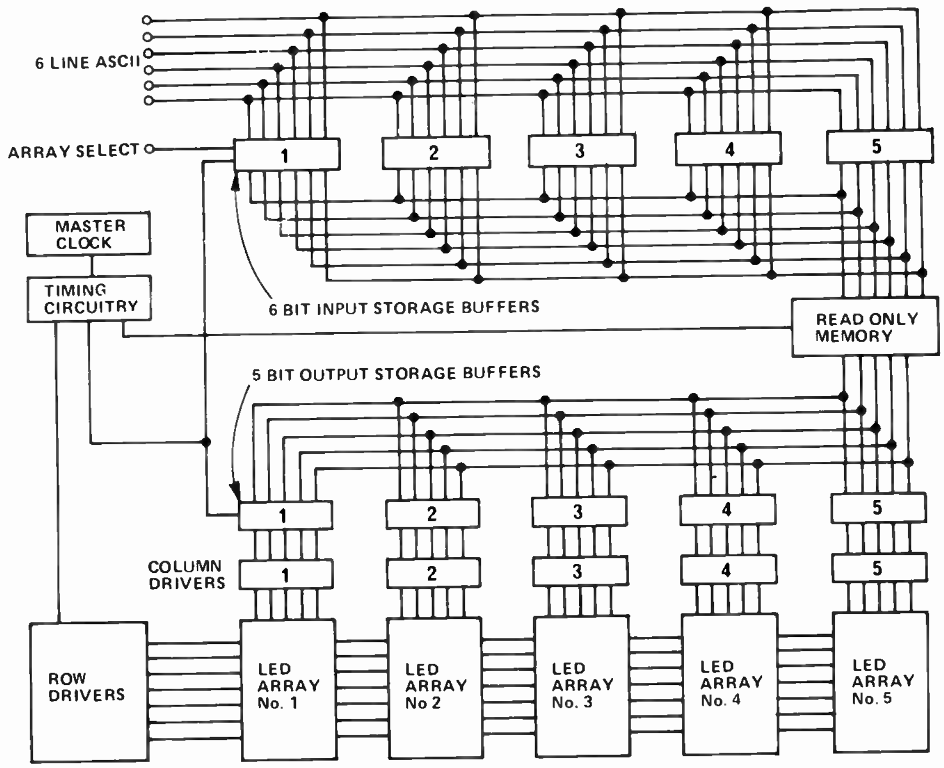


Fig. 15. Basic block diagram of a vertically strobed display using 7 x 5 dot-matrix devices.

returned to at no greater than 10 ms intervals the characters will be flicker free.

A diagrammatic illustration of how specific diodes are selected in a row is given in Fig. 14. The row switches are scanned in turn to cause a vertical scan. Simultaneous excitation of the other switch sets decides which diodes in the row are to be illuminated.

A strobing system requires a procedure to sequence the scanning action and a method of setting the selection switches that corresponds to the characters needed. The whole is controlled by a clock and timing generator. Storage buffers are also required to store the sequentially generated information. The task of creating the appropriate character timing codes is performed by a read-only-memory ROM. Clearly this method adds up to a complex system . . . really beyond this course's purpose. A schematic block diagram of a vertically-strobed five-digit LED display is given in Fig. 15. Although of apparently great complication this method is less expensive to employ than direct actuation through fixed gates. (Considerably more detail is to be found in the suggested reading list).

Another scanning method scans the matrix as a raster — across a row, one by one, and then to the next row. Strobing obtains its advantages by time-sharing common elements in a time-multiplexed manner.

MULTIPLEXING

When the input data to be displayed appears in serial form or when large numbers of displays (over four digits) are involved, multiplexing (selection of complete digits sequentially) becomes advantageous for driving seven segment and one-of-ten displays. The basic multiplexing system requires the main system units shown in Fig. 16. An upper limit to the number of digits is around 12 and higher for LEDs. There are disadvantages; namely, a higher voltage is required in the display to achieve the same brightness (LEDs are not so critical as other forms of display); the scan frequency must be at least 100 Hz to prevent flicker; transients must be carefully decoupled; and a clock failure (which stops the scan) may produce partial display failure because of excessive dissipation brought about by the increased voltage applied. (It is usual to include a failsafe protection circuit).

Again, the complexity appears great but in practice the multiplexed system

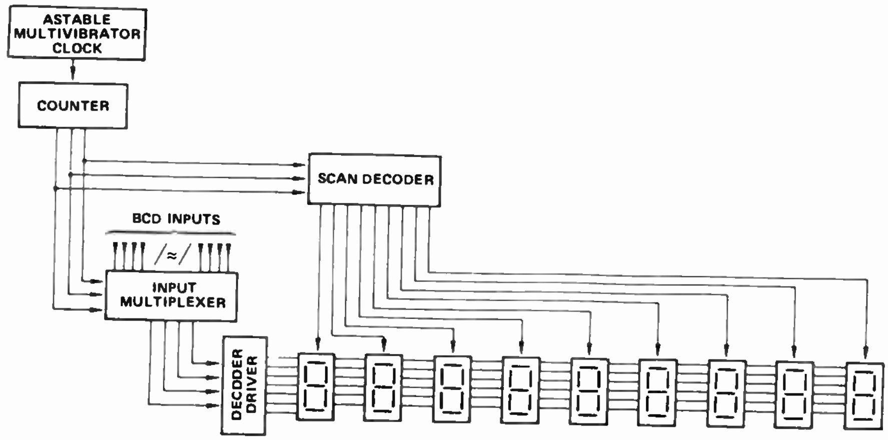


Fig. 16 Schematic of a multiplexing system for use with a multidigit, 7-segment display.

is simpler to build. For example, a multiplexed, seven-segment display, with storage for eight digits, involves around 10 dual-in-line packages and a few discretes which compares with about 16 ICs for a non-time shared system.

To further reduce the connections that must be made upon assembly, manufacturers offer multi-digit displays in which the anodes and cathodes of the LEDs are internally connected ready for multiplexed operation.

OTHER CONVERTERS

Apart from digital-code converters, other converters are required in instrumentation; for example, when interfacing different systems of logic, eg TTL to CMOS, it is necessary to alter the dc levels of signals so that the output of one system provides the logic levels required by that following. This may require amplification or attenuation or shifting of a level. However specific ICs are marketed to suit various interfacing requirements. Other converters are needed for sending digital signals through standard transmission lines in communication links, for receiving signals from lines, for increasing the logic level differences to increase noise immunity (again for transmission), and units that drive peripheral devices such as relays and indicators. Signal inversion may also be necessary — we have already dealt with the inverter block earlier in the course.

Another class of converter is needed for converting digital signals to analogue voltages (and currents) into digital form (A to D). Such converters will be dealt with in the next section.

THE CHOICE OF DISPLAY

Choosing a display can be quite a task because many options exist. During the early 1970s it was not uncommon for digital displays to be specified for applications where their analogue equivalents would have been more suitable. Now though, common sense is beginning to prevail. Analogue displays are gradually regaining ground as it becomes clear that they are more suitable for trend and other dynamic observations. Nevertheless, many of today's analogue displays use digital techniques internally.

Rotating pointers, bar-graphs and similar analogue visual effects are now being developed for use in the automotive industry. Large-scale production prototype systems are already undergoing trials in cars. From this area of development it is logical to expect these new forms of analogue display to find their way into other applications. Now that most of the development has been completed the costs should be low. The consequence is that they will be introduced very rapidly into general use.

Manufacturers are seeking and finding ways to produce displays that can be made cheaply in volume. This is possible using integrated circuit and thick film techniques. Figure 1 shows a conventional instrument panel compared with a solid-state electroluminescent unit developed by Smiths Industries. Note the drastic reduction in the number of individual parts and the great reduction in thickness.

It is also possible to make meter-style indicators in which the needle rotates — but that is more costly. We can expect to see changes in car instrument panels because of cost advantages of bar-graphs.

Electromechanical devices will still have their place for some time yet. Small-sized meter movements will be phased out but as yet solid state devices are not commercially viable for the large displays such as railway platform indicators and expressway signs.

Factors of key importance relate to the appearance of the display as seen by the user, reliability, ease of servicing, and power consumption.

Of particular importance is the 'price to use'. This can greatly exceed

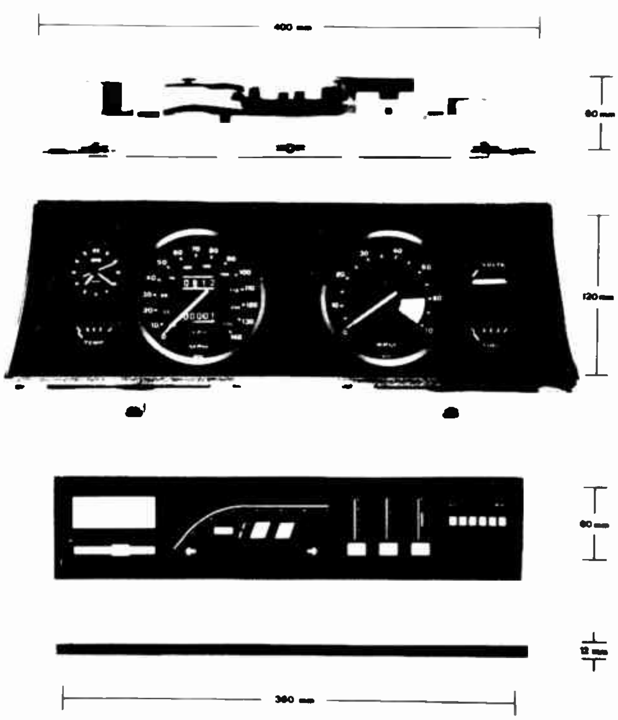


Fig. 1 (left). The conventional mechanical instrument panel (upper) contains 430 parts and is much bigger than the 35-part electroluminescent solid-state display (lower).

Fig. 2. Comparison of the display technoligies now in vogue.

		Optimum Number of Bits
Tungsten Filaments		1 - 20
Light Emitting Diodes	LED	1 - 30
Cathode Ray Tubes	CRT	10K - 250K
Gas Discharge (Plasma Panels)		30 - 5K
AC or DC Electroluminescence	DCEL	30 - 3K
Liquid Crystal Display	LCD	5 - 200
Electrochromic (liquid or solid)		5 - 200
Electrophoretic		5 - 200
Vacuum Flourescent	VAC.FL.	10 - 100

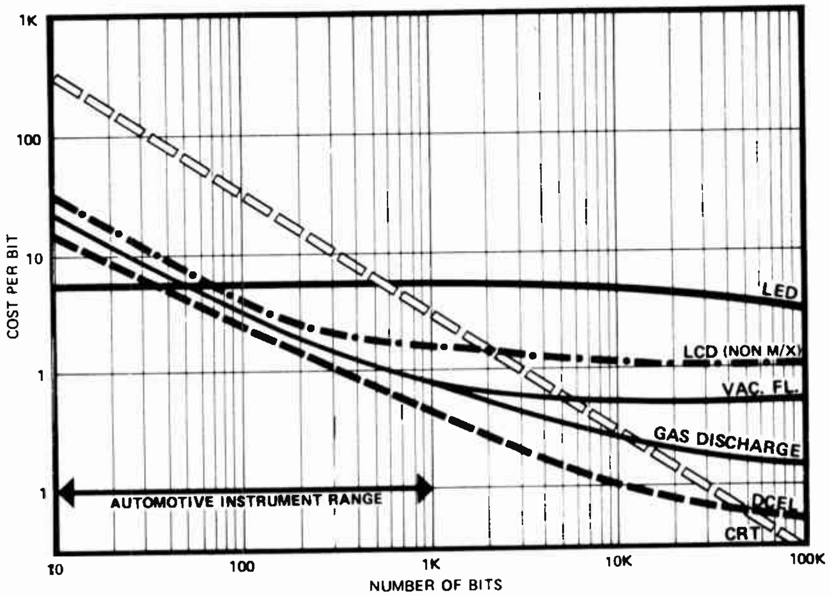


Fig. 3. Comparison of the cost per bit of the various display options.

the cost price because of the costs of power supply, mounting, wiring, and possible connectors. It is also important that at least two sources of supply are available.

Another factor to consider is the special characteristics of a display. Each has some good and some bad characteristics. For example a liquid crystal display is fine where ambient light exists but needs auxilliary illumination in low light conditions. LED's on the other hand are best seen in the dark – they need to be very bright to be seen in full sunlight.

It is also important to assess if the device is really fully developed. Many new products reach the marketplace before they have been fully tested. Today a new solid-state product can be realised and marketed in a matter of a year but it is not possible to test it for the whole of that time. The tens of thousands of hours life that may be postulated by the manufacturer is often merely conjecture. Liquid crystal displays were one example. No user wants to be part of a test programme . . . especially if he's paying for the privilege. New is not necessarily best!

The main display contenders are currently LED's, gas discharge tubes, cathode ray tubes, liquid crystals, and the fast-emerging electroluminescent panels.

The time-honoured filament lamps continue and need no further comment except to say that they are being replaced in small power displays by the more up-to-date devices.

LEDS

LED's emerged first as single element light sources of rather low brightness and in red only. Today they are available in brighter forms and of many different optical styles providing diffusing effects, wider angles of viewing and generally greater utility. Present day technology can provide 50 um square elements of which 300-600 may be integrated into a matrix. Such LEDs are available with light output sufficient for aircraft instrumentation (10^5 Lux) and can be made to full MIL specifications including operation over a temperature range of –55°C to 125°C.

They are available in colours ranging from red (the most common and cheapest) through yellow, green, orange and violet. Blue LEDs have been made but appear to lack a large enough market to enable them to be produced at a commerically attractive price.

The reliability of LED s is variously claimed to be from 10^4 to 10^9 hours. There is a suggestion (based on evidence from large-scale users) that price wars have tempted makers to reduce reliability. Reject rates as high as 20% are said to be experienced by some buyers.

LED s are fast operating: typical rise times are 10 - 50 ns. They are offered in pcb packages, in larger metal packages suitable for sealing and in more-expensive-still ceramic packages.

The ready ease with which they can be assembled into lines, circles, matrices and other graphical forms enables them to be used in analogue displays.

LED s are not necessarily the best choice for all displays. Figure 2 compares various displays on the basis of the optimum number of bits for each alternative. It can be seen that LEDs are restricted to applications where the type of display requires only a small number of bits. Electroluminescent panels (discussed below) are more suitable where the application calls for the use of many bits.

Another factor is the cost per bit to manufacture. Figure 3 compares this variable for the various types of displays. The LED does not compare well for applications requiring over 100 bits. On this basis the CRT is way ahead. As yet it is not even remotely matched by any solid-state technique. The CRT's main drawback is that it is bulky and fragile compared with most other types of display.

GAS DISCHARGE TUBES

Gas discharge tubes were the first displays that could reasonably be regarded as versatile. Many older readers will recollect the Dekatron counter tubes of the 1950's in which 'dots' moved circumferentially in a scale of ten. A later development incorporated grids placed behind one another in a single glass envelope in what was generally called the Nixie tube. Their main disadvantage for use with solid-state circuits was their need for a 170 volt supply.

With the introduction of solid-state displays it might be thought that gas discharge tubes would have been supplanted. This has not happened so far. Indeed indications are that they will be used for a considerable time yet. Their brightness and large size are still strong advantages.

Gas discharge tubes are made in many forms. These include low profile, alpha-numeric, bar-graph, special purpose graphical displays, and still in the research and development stage, are phase addressed matrix co-planar units which use thick film manufacturing techniques.

It is possible to construct gas-discharge cells so that a particular cell is set to strike and erase at different discrete voltage levels. Thus, increasing the voltage level to a line of adjacent cells will produce a bar-graph effect. Once struck, the cells latch on exhibiting a bistable storage characteristic.

Cross-bar arrangements of grids (as shown in Fig 4) enable 'dot' discharges to be established at the junction of any two selected bars. Thick-film replication methods are used to manufacture the units.

The colour of discharge tubes can be finely tuned to just about any wavelength. This is done by adding an appropriate phosphor to the cell during manufacture. White and blue remain difficult to produce.

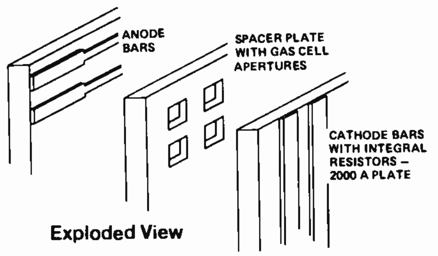


Fig. 4. Construction of dot-matrix gas-discharge display unit.

A phase-addressed technique has been developed to reduce the number of leads otherwise needed to connect all matrix positions to external circuits.

Gas discharge tubes are still, and will remain, a strong contender in the choice of a display. Figure 2 illustrates this well.

CATHODE RAY TUBES

A CRT screen of good quality and having a good linear scanning system can accommodate a display of 1000 by 1000 elements. The full range of colours is available as well as an intensity scale having perhaps 200 levels. Cost per element is very low but size and fragility go against the CRT in many applications. Eventually, as matrix manufacturing methods become more developed, the CRT's thin flat digital equivalent will become a serious rival. At present though, the CRT has no rivals for displays requiring large numbers of bits.

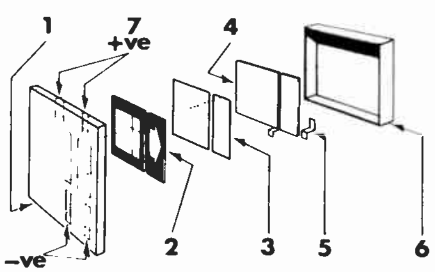
LIQUID CRYSTAL DISPLAYS

In many ways LCD's got off to a start less worthy than they deserved. Reliability was variable: many failed rapidly whilst others did very well. Failure of an individual display within a batch could vary from almost immediate through to years.

The second generation of LCD s has shown itself to be very much better if made by more controlled procedures and with better materials. Figures such as 90 000 hours to reach a 2% cumulative failure have been claimed for twisted-nematic LCD displays.

A key factor has been the realisation that a non-zero dc cell level rapidly degraded the cell. That restriction was originally controlled by the use of ac bias but now zero level dc working has been devised.

Initial commercial incentive came from watch manufacturers, but now researchers are seeking ways of building much larger panels — 150 mm square for example. Such large sizes



D.C. ELECTROLUMINESCENCE ASSEMBLY

1. FRONT GLASS
2. INSULATING MASK
3. PHOSPHOR – in contact with 7 through apertures in 2
4. ALUMINIUM REAR ELECTRODE
5. CONDUCTIVE STRIP – joining 4 to 8
6. HERMETICALLY SEALED COVER
7. TRANSPARENT CONDUCTIVE PATTERN

Fig. 5. Schematic of a dcel electroluminescent display panel.

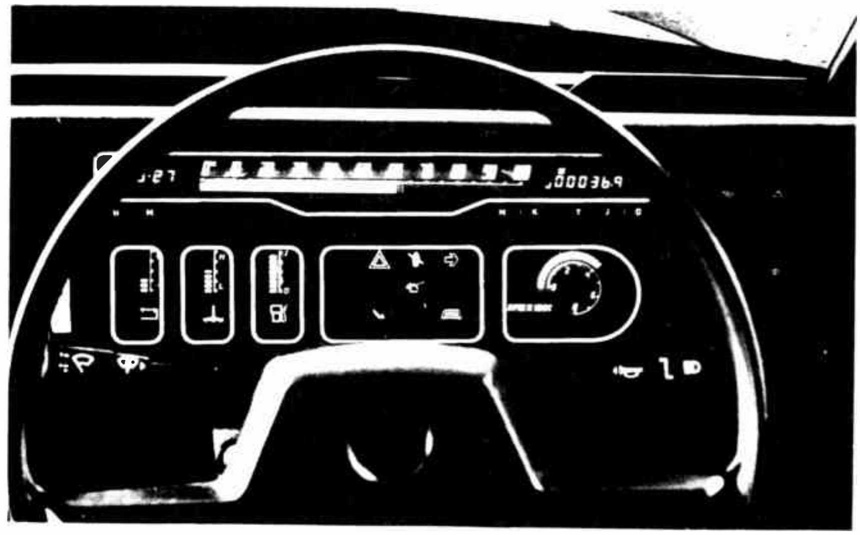


Fig. 6. On-vehicle electroluminescent display dash panel (Smiths Industries).

pose manufacturing problems for the glass enclosing the LCD material must be flat to within a mere 10 um.

Manufacturing methods are constantly being improved. The glass front seal has been greatly changed . . . that was a cause of many premature failures. Purer LCD material and improved stability with temperature and humidity have also improved.

Matrix units are being investigated but, as with all such units, connecting problems remain. Some LCD s currently being released have shift registers integrated onto the display. This trend may become common practice, for the user does not wish to connect any more than the minimum of leads from the drive circuits to the display.

In general LCD manufacturers suggest that their products are best suited for applications requiring a portable display. The current LCD s are certainly much better than the first generation and their low power consumption gives them a firm place in the display range.

ELECTROLUMINESCENT DISPLAY PANELS

Electroluminescent devices are basically just a layer of special paint between two pieces of glass.

Two basic groups exist . . . ac working (called acel) and dc working (called dcel). Each uses zinc sulphide, manganese-doped phosphors which radiate a yellow-orange light at 585 nm wavelength. The ac cells operate in a capacitive mode, the dc units in a resistive mode. Figure 5 shows the schematic of a dcel unit.

Manufacturing processes are mainly vacuum deposition using photolithographic procedures for masking. This method offers great prospects for the future. The British Post Office for example is considering 1250 character displays for phone call costing. Smiths Industries have vehicle instrument panels in pilot scale production. (Figure 6 shows a recent panel of this type).

As always, addressing the display is a problem. Multiplexing methods have been used to reduce lead counts from 257 (for a 256 unit) down to only 32. The displays can be used in a continuous mode or they can be pulsed. Pulse durations of around 0.5% duty cycle are typical using 5 — 15 microsecond pulses.

Around 120 volts is needed to drive the display: present day units require 50 mW per character. A prototype unit using CMOS circuitry consumes only two watts for a 480 character display.

This information was compiled from lectures delivered at an Institution of Electrical Engineers day meeting held in London in January 1978. No full Proceedings were published but the five speakers would be able to provide further information if contacted. Details can be obtained from the Conference Secretary, IEE, Savoy Place, London. Smiths Industries kindly provided most of the illustrations used here.

REFERENCES

"Digital Display Systems" referred to in the previous part deals with blanking and multiplexing.

"Solid State Alphanumeric Display Decoder/Driver Circuitry" — Hewlett Packard application note 931 gives details of scanning methods used with 7 x 5 displays.

"Mullard TTL Integrated Circuits — Applications" includes a chapter devoted to various kinds of code converter gating layouts.

27 D-A and A-D conversion

IN ORDER to control or modify the physical world around us we must first measure what is happening. The measurement data is almost always in analogue form, as is the actuation required for control. Between measurement and control some kind of electronic system is needed to amplify and shape the data.

We have seen that electronic systems may be of either analogue or digital form and it would seem best to use an analogue system between inputs and outputs that are both of analogue form. But not necessarily so — analogue systems are plagued with problems such as noise, dynamic range limitations, accuracy and linearity. Digital systems, as well as offering improved performance in the above areas, offer more economical processing of data, the ability to store data as long as needed, and more readable displays of data held within the system.

Thus there is much to be said for converting primary analogue signals into equivalent digital forms that are processed and stored etc until conversion back to analogue form becomes a necessity. Electronic sub-systems that perform these conversions are called Digital-to-Analogue Converters (DAC's or D/A converters) and Analogue-to-Digital Converters (A/D converters). (See Fig. 1)

We will see that these are quite complicated systems in themselves — their design a skilled task. Nevertheless, many such sub systems are now marketed as single, largish circuit blocks that are wired into the total system in the same way as other complicated system building-blocks we have already encountered. It is, however, important to understand the basic techniques used if not so much the refinement of actual practice.

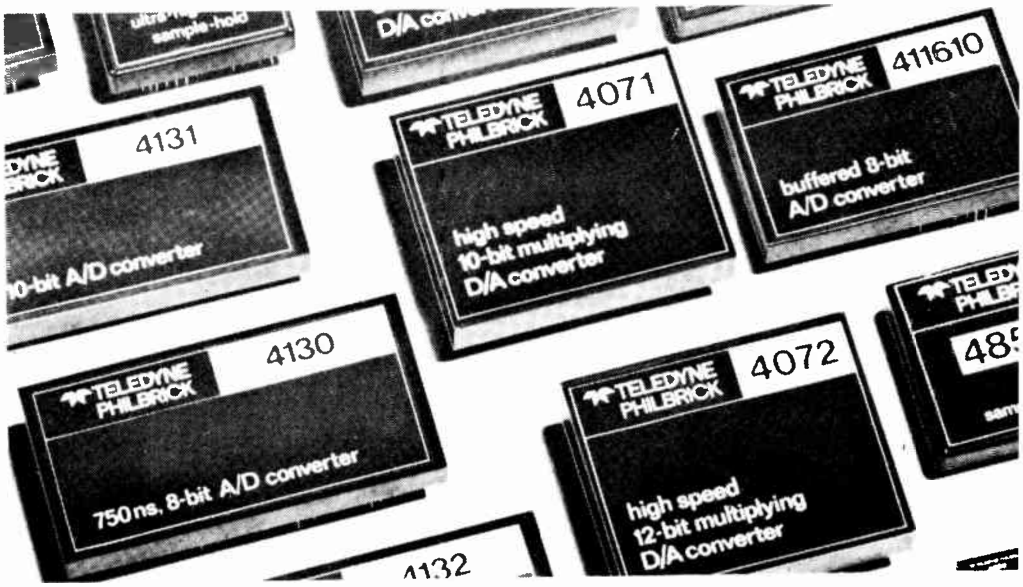


Fig. 1. Typical A/D converters.

By use of certain input combinations DAC's can also be used as multiplier/dividers of two signals and as summing/subtraction units. It is also relevant in this part to deal with multiplexers in a little more detail and with the so-called sample-and-hold circuit. These are often used in conjunction with DAC's.

Figure 2 illustrates the basic requirements of a typical data-acquisition system in which a number of physical variables are measured and processed to provide digital signals for storage. It uses multiplexer, sample-and-hold, and A/D converter sub-systems to form the whole.

Also pertinent, because similar techniques are involved, is the method for converting an analogue voltage to a signal of proportional frequency (which is a form of digital signal) — the Voltage-to-Frequency or VF converter.

The uses for A/D and D/A converters are limitless. Their application is ever-increasing as the unit cost falls to undreamed-of prices. Extreme complication using digital techniques often costs far less than simpler but less accurate analogue alternatives. Hence D/A and A/D converters will be found in digital panel meters, digital multimeters and data acquisition systems. They are also found in industrial plant; in process control of chemical and other manufacturing plant; in telemetry systems and other data transmission applications; in the interfaces (units matching the output signal requirements of one system with the input requirements of another) found between sensors and computing units; between stages of hybrid computers; and the like. Although

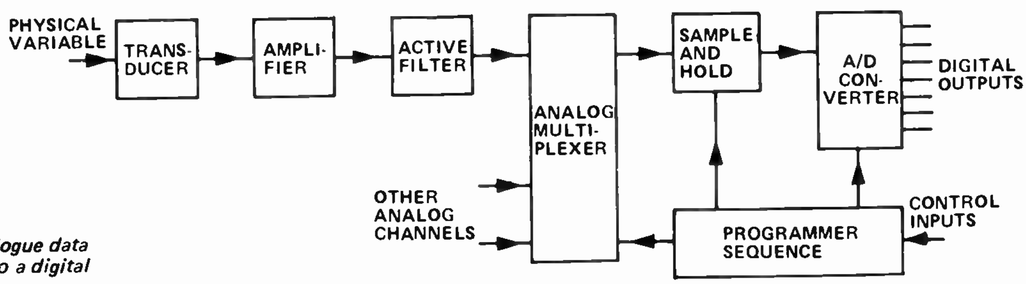
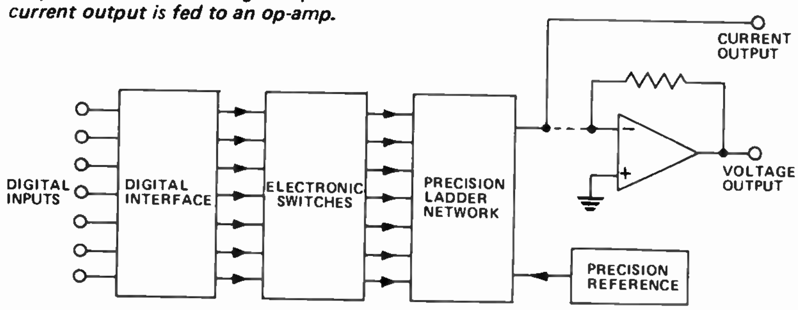


Fig. 2. Typical layout of an analogue data acquisition system interfacing to a digital output.

Fig.3. Schematic of a digital to analogue converter which provides a current or voltage output. To obtain a voltage output the current output is fed to an op-amp.



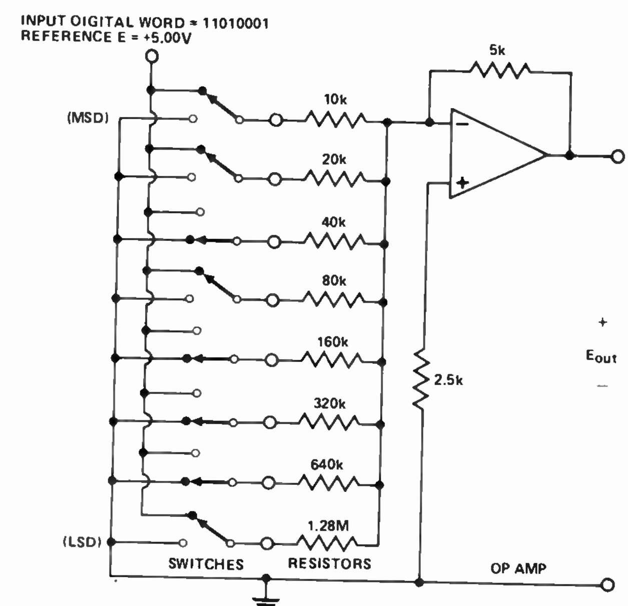


Fig.4. Switched resistor network of 8-bit, binary weighted D/A converter. 11010001 is being converted.

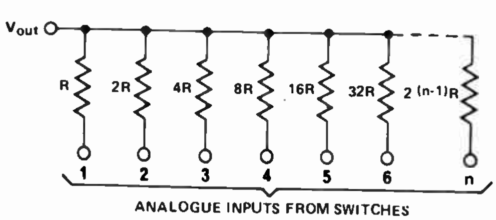


Fig.5. Basic binary-weighted resistance sequence.

highly accurate DAC's may cost as much as $2000 the great universal demand has resulted in lower performance units being available for as little as $10.

DIGITAL-TO-ANALOGUE CONVERSION

We begin with Digital-to-Analogue converters because they are the simplest in concept and use only one basic technique.

You will remember that each digit position of a binary number (held in a register, counter or other form of storage) has a weighting factor, eg 1:2:4:8 etc. An analogue signal equivalent to the binary number can be obtained by using each digital digit position to switch an amount of current (proportional to the position weighting) to a common summing junction. This system concept is shown in Fig.3. When voltage output is needed the currents feed an op-amp. The detail of a precision ladder network is shown in Fig.4 – it is set to convert the input digital number 11010001 by appropriate positioning of the digit position switches. When at 0 input the inputs to the summing op-amp are held to ground; when at 1 to an appropriate stabilized voltage.

The simplest form of ladder is used in the circuit of Fig.4. It arises from the use of a binary weighted resistor sequence shown in Fig.5. The actual values of resistors are selected to obtain adequately sized lowest and largest values, for at either end the op-amp loses accuracy due to imperfections of resistance ratios. It can be seen from the circuit of Fig.4 that resistors, even in a smallish capacity 8 bit converter, can extend to extreme values. The least significant bit must be clearly resolved when its switch operates, implying that all other resistors must have precision of absolute value and constancy with time that rises very rapidly with the number of bits required.

In practice this simple form of ladder is not used beyond about 8 bits conversion due to the cost of the precision resistors required.

The disadvantages of the simple ladder method are mostly overcome by the use of the R-2R ladder network shown in Fig.6. The through leg of the chain is permanently grounded, each spur is switched as needed to a reference stabilised voltage level. The features of this method are that only two values of resistors are needed (an easier practical problem) and that the absolute range seen by the op-amp varies much less than the above method for a similar bit capacity – it presents a virtually constant impedance regardless of the binary code sequence switched in. With the R-2R ladder it is routine to provide 12 bit conversion.

It is probably obvious that other forms of digital-coding conversion can be handled by the use of appropriate resistor weightings. For example it is often necessary to chart-plot the output of a digital instrument. Thus a BCD to analogue converter is required for such applications. Figure 7 shows the weighting sequences for the simple and the R-2R ladder DAC's needed to convert BCD inputs to an analogue output.

Resolution and accuracy – DAC's rarely go beyond 12 binary bits (or 3 digit BCD) because the output analogue signal for greater bit-ranges must be of high stability. A 16 binary bit (or a 4 digit BCD) unit could provide ± 0.005% full scale linearity and accuracy, a performance requirement that is best avoided where possible because of the high cost of the DAC. By contrast 8 bit DAC's can be obtained with accuracies ranging from ± 0.2% full scale to ± 0.01% full scale. It is important, however, to realise that whereas analogue resolution, see Fig.8, is a function of the number of bits that are equivalent to full scale, the accuracy and linearity of DAC's depend upon the tolerances and stability of resistors used in the conversion networks, for these decide the value of the slope and straightness of the slope — Fig.s 9a, 9b. It is, therefore, possible to have an highly accurate converter that has quite coarse resolution — in which case the resultant analogue output signal will consist of very large step changes. This step form of signal defect is called quantum or quantization noise. In practice resolution and accuracy are tied together keeping quantum noise to an acceptable level.

Conversion and settling times — As both D/A and A/D conversion are dynamic processes, a finite amount of time is required for each conversion point to reach its final value. In DAC's the switching and settling times of the op-amp largely dictate the time for a bit change to finally appear as a steady-state analogue signal level. Early DAC systems using mechanical switches were slow indeed — today output settling times range from a slow 25μs for very-low power consumption units to ultra-fast 25ns units.

Temperature coefficient — Each subsystem of a DAC has a temperature coefficient; resistances alter with temperature and the op-amp characteristics deviate. Both the overall conversion gain and the dc zero will be affected.

Gain will be affected due to the temperature coefficient of resistors which is typically from 50 to 100 PPM/°C (100 parts per million, PPM, is equivalent to 0.001% change per degree Celsius). The main op-amp characteristic which affects performance is offset-voltage drift — typically 30 microvolts per degree Celsius.

For each particular type of DAC it is necessary to consult the makers' specification sheets, for no general rules apply for these parameters.

DAC's are available as either current or voltage output systems. Typical outputs deliver around 3 mA and 10-20 V swings.

Further explanation of terms used is to be found in the articles listed in the further reading section.

Glitches — Certain digital input states, whilst in a transient state, can cause the output to produce noticeable transients to the smooth, stepwise analogue — signal progression. These are known as glitches; examples

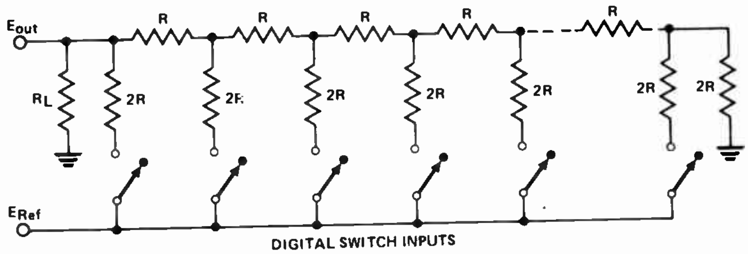


Fig.6. The R-2R ladder network requires only two values of resistance and presents constant impedance to the output.

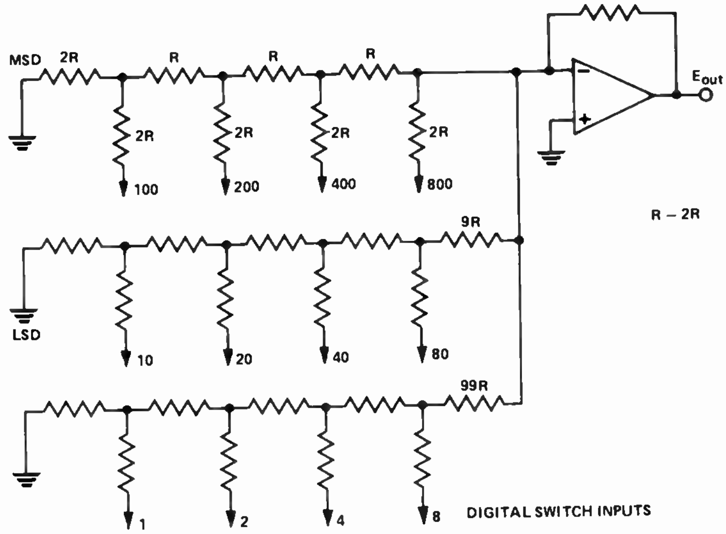


Fig. 7. Resistance layout for simple BCD-weighted DAC.

are shown in Fig.10. If the output is only to be observed after the system has settled, these matter little. In dynamic use of DAC's however, they may well excite unwanted behaviour in the system they are driving.

Deglitching in DAC's is not feasible with a low-pass filter on the output, for glitches vary widely in nature. The best solution is to use adequately fast and matched switching coupled with special deglitching, (sample-and-hold) circuits, that hold the output fixed during unwanted switching — transient conditions. Glitching states are, however, known states and are quite unlike random noise which defies prediction.

Integrated circuit current sources — As the DAC principle finds a variety of uses, manufacturers offer an integrated circuit which provides an output current, the magnitude of which is controlled by a four bit, binary-code input. The IC, as shown in Fig. 11, has

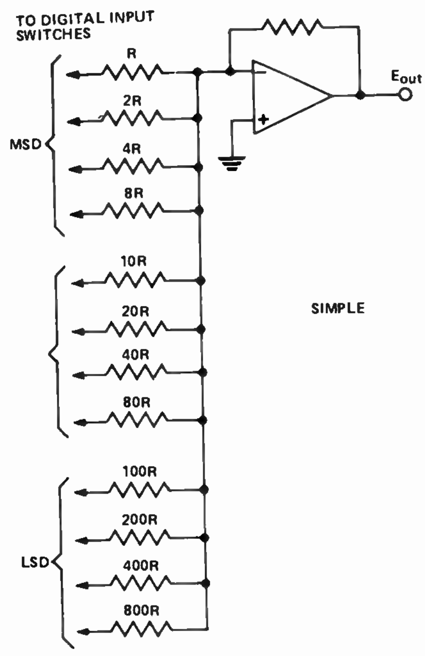


Fig. 7. Resistance layout of R2R, BCD ladder D.A.C.

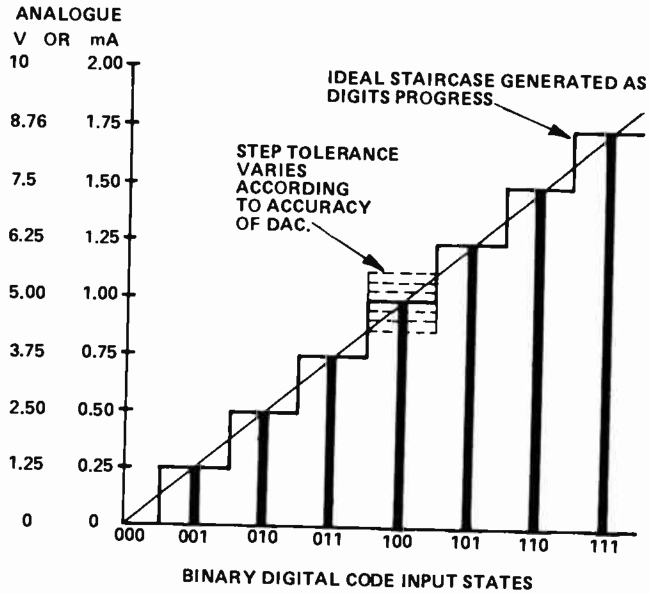


Fig.8. Output/input characteristics of typical 3-bit DAC having full-scale current of 2 mA or 10 V output.

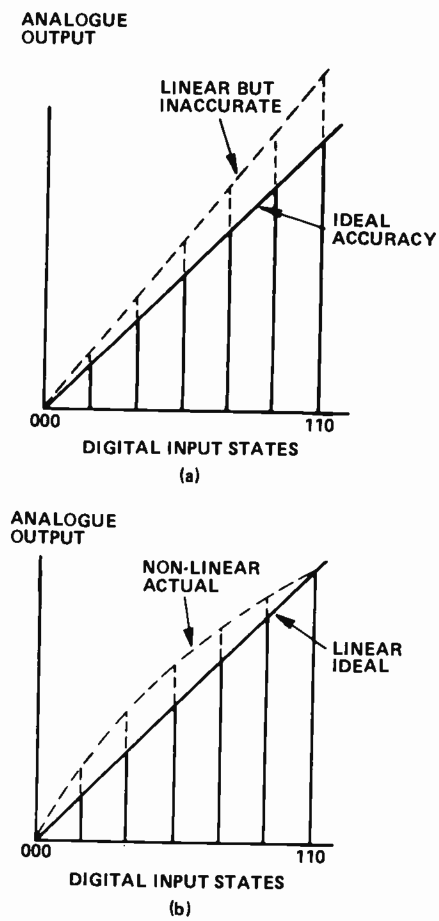


Fig.9. Output/input definitions for DAC's
(a) deviation of accuracy.
(b) deviation of linearity.

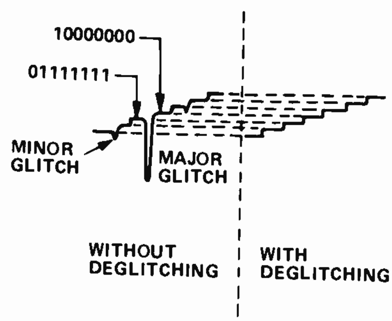


Fig. 10. Glitches occur in basic DAC's due to certain transient dynamic error conditions. They are eliminated by deglitching circuitry.

enough precision to be used in 12-bit D/A and A/D conversion. Figure 12 illustrates how two of these ICs are combined to produce an 8-bit DAC.

Digitally controlled sources – The DAC's described above are usually concerned with signal processing as opposed to analogue power control. If larger output powers are needed digitally-controlled power supplies can be used – also referred to as digitally programmed supplies. They are available with digital control of current or voltage outputs. Instruments in the Hewlett-Packard range, for example, can provide up to 125 watt maximum demand whilst the output is controlled by binary or BCD inputs with a programming time of around 350μs.

ARITHMETIC OPERATIONS WITH DAC's

Multiplication – The resistor network of a DAC has two inputs – the reference supply (which is fixed in normal D/A conversion) and the switch inputs representing digital numbers. If the reference is allowed to vary as an input variable, see Fig.13, the output of the DAC is the multiplicand of the two signal inputs. The reference may also be an ac signal and division and attenuation are also possible. The advantages of this method are the high precision and speed available.

Addition – If the difference or sum of two digital signals is needed as an analogue output, two DAC's may be combined, as sum or difference, into the output op-amp, as shown in Fig.14.

ANALOGUE TO DIGITAL CONVERSION

Conversion from analogue to digital code can be obtained by many alternative techniques, each alternative having many variations. Basically, methods group into open-loop and feedback-loop systems. In each group some four to eight ways are in common usage. Here we look at a few of the most popular techniques, beginning with open-loop methods.

Analogue-to-frequency – The analogue voltage is converted, on a continuous basis, into a signal of proportional frequency by the use of an appropriately accurate V-to-F converter. (Voltage controlled oscillators are used – they must be adequately linear). This signal, see Fig.15, is gated into a digital counter using fixed times of gate aperture. The counter accumulates a digital number equal to the average analogue level over the gate period. The counter output is released upon demand when gate periods expire. This method suffices for low accuracy analogue inputs but becomes expensive when precision, wide dynamic range is required.

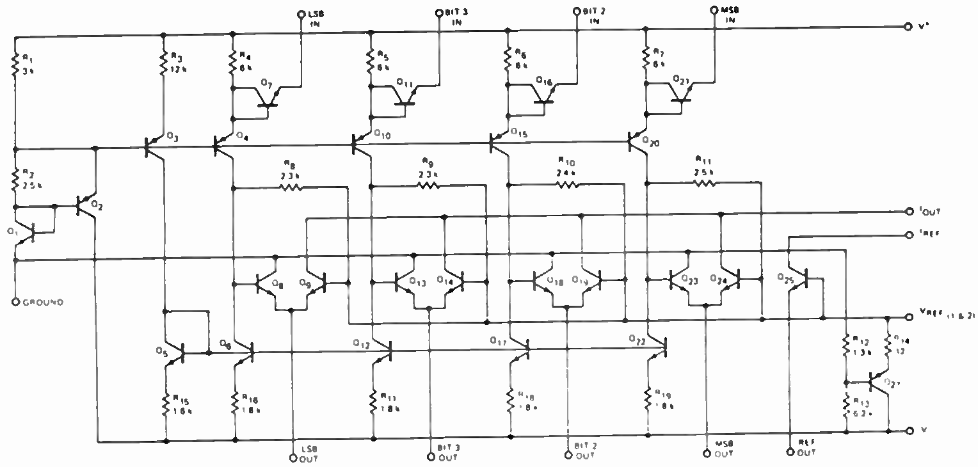
A variation of this is to reverse the philosophy and vary the pulse width of accurately generated pulses of constant frequency. The variable pulse width is then converted to a digital count proportional to the pulse width.

Simultaneous, parallel or flash conversion – The input analogue signal is presented to a stack of comparators (Schmidtt trigger action) each set to trigger at increasing binary-weighted signal levels. The set of comparator outputs are then decoded to provide the required binary output form. Decoding is needed because at any instant all comparators set to below the signal level are in the one state, all above in the other. The method is given in Fig.16. Although extremely fast – quantization time is the speed of a single comparator – the method has the serious disadvantage that large bit ranges require many comparators and numerous decoding gates.

Closed-loop methods – are more popular and there are about six main alternatives The methods known as integrating, successive approximation and servo-DAC are most generally applied.

Dual-ramp integration – The

Fig.11. Schematic of dual-in-line 4-bit current source (Fairchild μA9650).



analogue voltage is first converted to a time period which in turn is converted into a binary number by a timer/counting system. Referring to Fig.17a, conversion begins when the switch connects the analogue-signal input to the integrator which commences to ramp up. At the same time the counter begins, from zero, to count the clock pulses. When a predetermined number of pulses (1000 is convenient) appear in the counter the integrator is electronically switched over to the reference. At this point the capacitor has then charged linearly from the input, rising as a ramp to a voltage level decided by the average input-signal value as shown in Fig.17b.

As the switch changes to the reference position the counter is reset to zero and begins counting again. The reference, chosen to be of opposite polarity to the input signal, now causes the charged capacitor of the integrator to ramp back downward at a constant slope. When the integrator output reaches the zero threshold the counter is stopped and its contents displayed. The count displayed is the ratio of downward ramp counts to

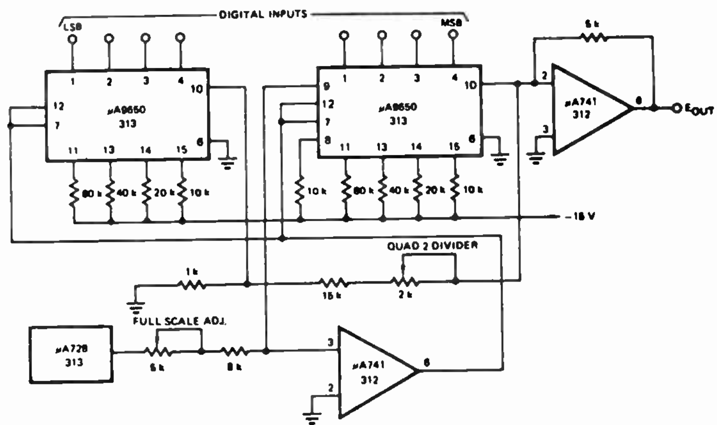


Fig. 12. Applying 4-bit current source IC's to form an 8-bit DAC.

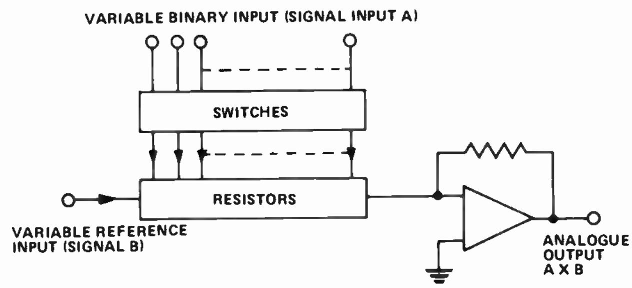


Fig. 13. Using a DAC to multiply an analogue signal by a digital signal.

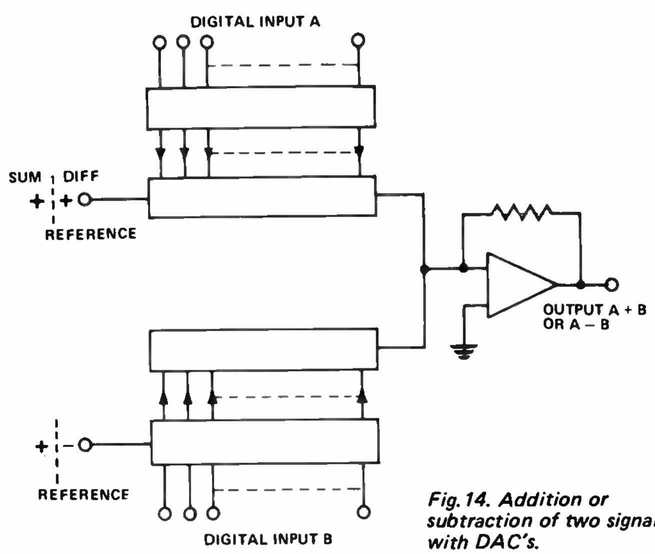


Fig. 14. Addition or subtraction of two signals with DAC's.

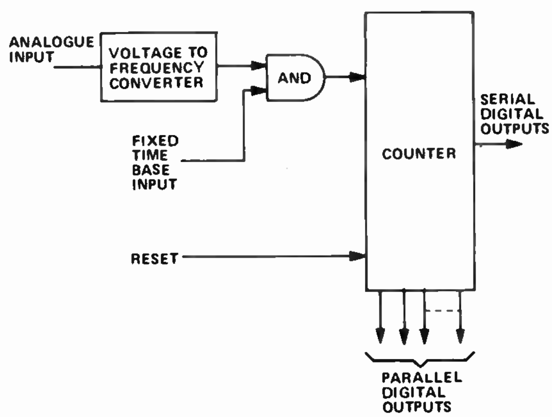


Fig. 15. Voltage-to-frequency method of A/D conversion.

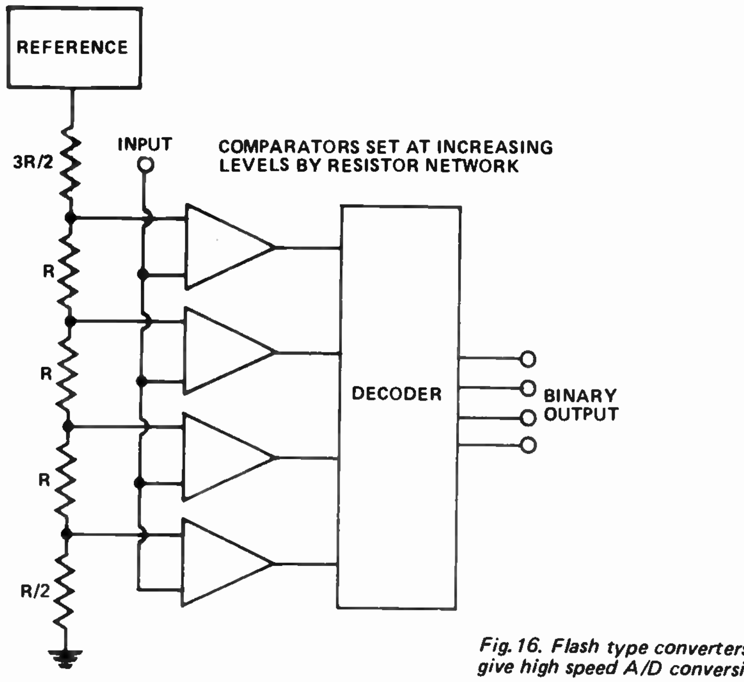


Fig. 16. Flash type converters give high speed A/D conversion.

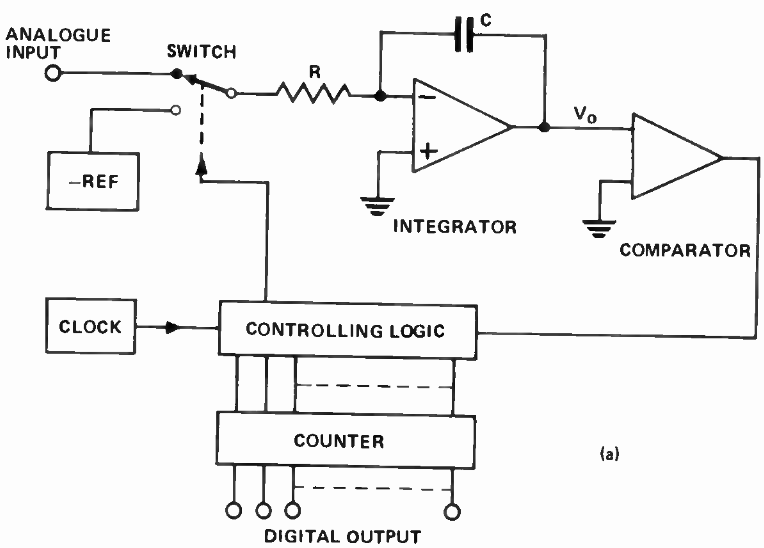
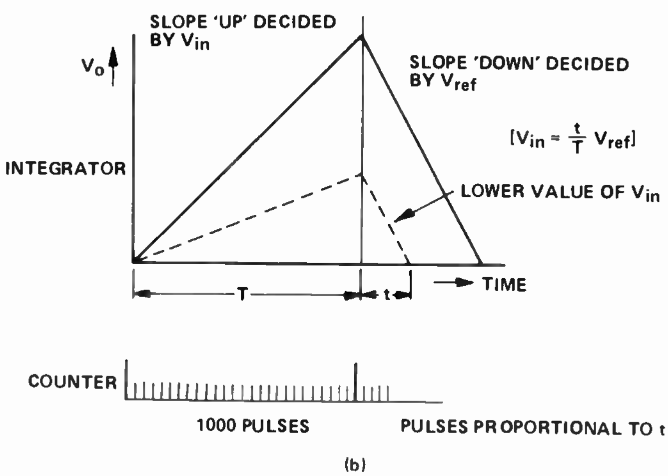


Fig. 17. A/D conversion using dual-ramp integration. (a) schematic (b) timing diagram.



upward ramp counts which, when a 1000 upward limit is used, gives a direct reading of input voltage if the reference voltage is appropriately chosen.

A simpler form, using only one ramp, is also used but it lacks the features of the dual-ramp method in which the absolute value of the capacitor and the clock frequency are of no significance provided they are stable for the duration of the conversion period. The dual ramp method does, however, require a relatively long conversion period but this is an advantage in one respect — the value measured is more accurate. This is due to the fact that when noise is integrated over an extended time period it tends to zero.

A more sophisticated triple-ramp method provides increased speed and accuracy for a moderate increase in cost and complexity. In essence two reference signals are provided, one acting as a 'coarse', the other as a 'fine' ramping control. The 'coarse' ramp rapidly converts the bulk of the input signal level leaving the 'fine' ramp to add the extra resolution. Figure 18 shows the schematic and timing diagram of a triple-ramp A/D converter.

Successive approximation — Due to its high resolution and fast conversion speed successive approximation is the most widely used method. A schematic diagram is given in Fig.19. Conversion progresses step-wise with the precisely generated DAC output being compared against the unknown analogue input. The first comparison is made with the most significant digit of the DAC, which gives 1/2 full scale, being compared against the unknown. If it is smaller, the bit is retained as a '0', if larger it is set to '1', thus the MSD value of the programmer is found. The next digit, working towards the least significant end one digit at a time, is then tested for the same criteria being set accordingly. The process is repeated until all programmer digits are set to '0' or '1'. The value in the programmer is then transferred to the register for outputing in parallel or serial form. Conversion time is not decided by the value of input as in ramp methods, duration of conversion being the number of bits times a fixed digit test interval, which can be as fast as 100ns. By comparison, from one maker's options, successive approximation instruments offer conversion times which range from 1-60μs compared with 2.5-6.0ms for integrating converters. Accuracy clearly depends upon that of the DAC which forms part of the comparison system.

Servo-DAC method — Fig 20 shows this system. When conversion begins, a counter is gated and commences to

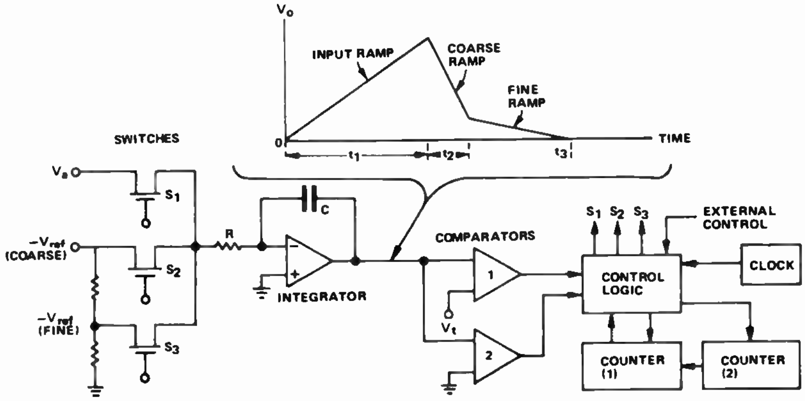


Fig. 18. Triple-ramp A/D converters offer faster speed for slightly more complexity of integrating method.

count upward. Its digital output is converted back to analogue form by a DAC. The output of the DAC is compared against the unknown input voltage. When the two analogue voltages are equal, the comparator inhibits the counter. At that time the value in the counter is a digital representation of the input – with 1:1 correspondence; or other ratios depending upon the summing resistances used. It is a simple low-cost method providing reasonable accuracy but operates at a slower speed than offered by successive approximation designs.

Non-linear conversion – Each bit of the above methods represents an equal quantum error. Thus one quantum error in full-scale is considerably less inaccuracy than in say a tenth or hundredth of full-scale. The smaller the reading, the greater the relative error of quantisation. When range-changing is not practicable a non-linear digital method can be used to compress the large scale in order to reduce the percentage of reading error The method is explained in Motorola Application Note AN-471.

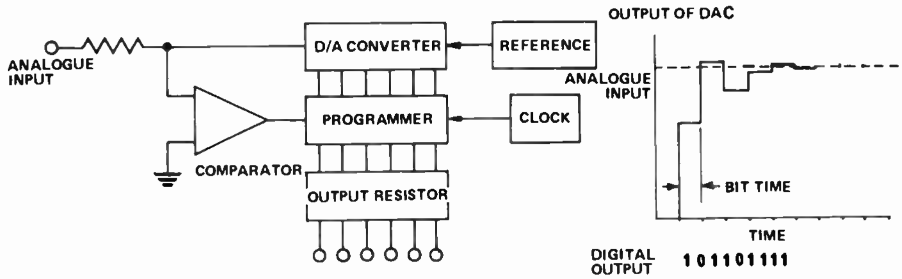


Fig. 19. In the successive approximation A/D converters a DAC is used to convert the incremented digital stage output back to analogue form for comparision with the input.

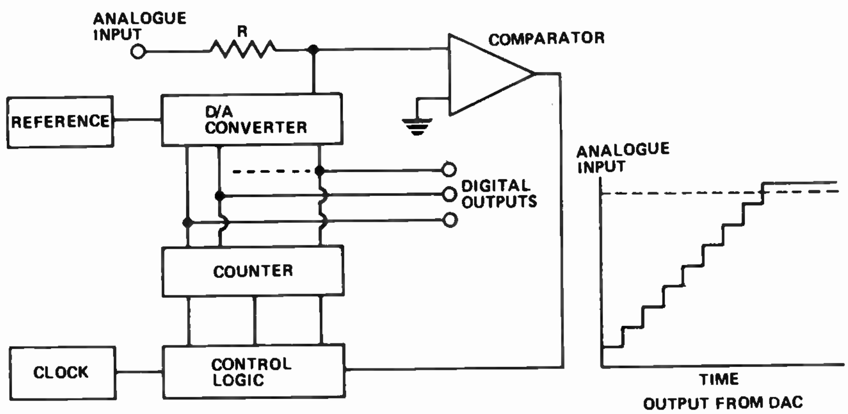


Fig.20. Schematic of servo A/D converter system.

SAMPLE-AND-HOLD UNITS

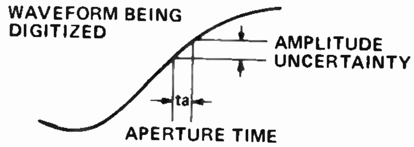
A digital signal provided by an A/D converter represents some measure of the analogue level seen in a certain gating period – the so-called aperture time. Aperture time, bit resolution and maximum signal frequency are strictly interrelated. Figure 21 is a chart enabling this characteristic to be found. For example, we may need to digitize a 10 kHz sinusoid (as the highest frequency to be preserved in a complex waveform) to a resolution of 12 bits. The chart shows we must have an aperture time of no greater than 42 ns. Thus we see extremely fast converters are needed for direct conversion of moderately high-frequency signals at high resolution.

A sample-and-hold circuit circumvents this difficulty by taking a rapid narrow-aperture sample of a signal and holding it in a simple analogue store long enough for the converter to act with a much wider aperture time. Figure 2 shows such a system. To preserve the highest signal frequency of a complex signal we only need to sample at twice (or higher) the signal frequency (Shannon's sampling theorem). This is a considerably slower rate than needed for direct A/D conversion. If sampled too slowly, not only will higher frequency information be lost but an effect, called aliasing, will occur by which a lower frequency is generated that may not exist in the original signal.

As mentioned above sample-and-hold circuits are also used in DAC's to remove glitches.

Basically a sample-and-hold comprises a capacitor with which to store an analogue voltage level, and a switch to charge the capacitor to that value in a way that can be rapidly and effectively isolated from the source. In practice low leakage FET switches are used in conjunction with IC op-amp integrators. It is also important to buffer the output of the sample-and-hold to reduce the loading which would otherwise decay the stored level.

Many circuit variations exist, the one shown in Fig.22 – a closed loop configuration – gives good linearity and accuracy. When extremely long storage times, or negligible decay with time is needed, the voltage on the capacitor can be transferred via an A/D converter into a digital storage register and back again into analogue form via a DAC as shown schematically in Fig.23. This naturally increases the cost considerably. More detailed information is available in "Analog-digital conversion handbook" TH6, by Analog Devices.



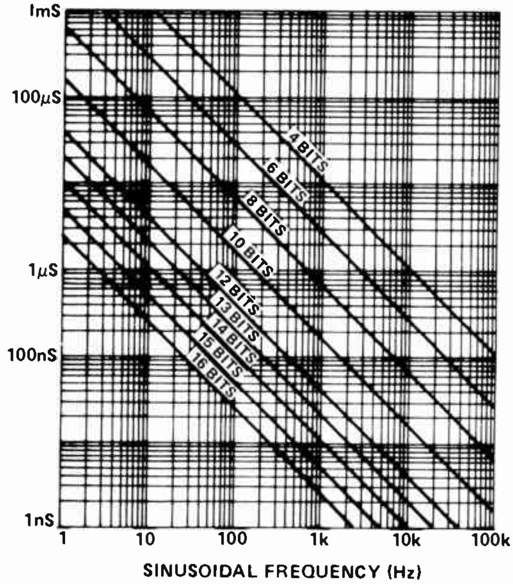


Fig.21. Graph showing relationship between aperture time required, resolution of signal conversion and frequency of sinusoidal signal undergoing conversion.

MULTIPLEXERS

The task of the multiplexer, shown in Fig.2, is to sequentially connect a multiplicity of compatible inputs to a single output line. In the case given in Fig.2 it feeds a sample-and-hold, which stores the signal for A/D conversion.

A multiplexer consists, therefore, of as many switches as there are input channels to be combined. In practice these must possess adequate speed and very low on-to-off switch resistance ratio. Solid-state multiplexers mostly use MOSFET switching devices feeding a buffer stage (a voltage follower configuration which has extremely high, 10^9 ohms, input impedance). Fig.24 shows such an arrangement. Fig.25 is the internal schematic of an 8 channel unit which has an ON resistance of 300 ohms and OFF resistance of 100 megohms with 0.03nA leakage. The settling time to 0.01% of value is 1μ sec and it can be sequenced to 500 kHz. Channel selection is made via the appropriate digital-code input.

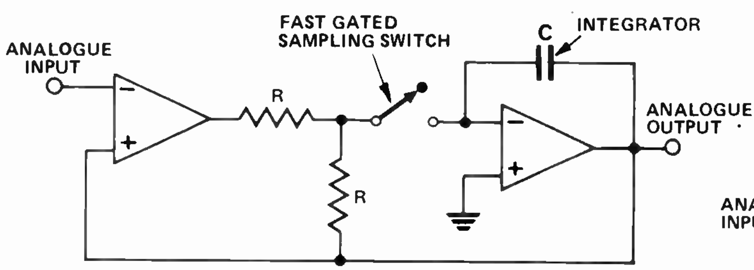


Fig.22. One form of stable and accurate sample-and-hold circuit.

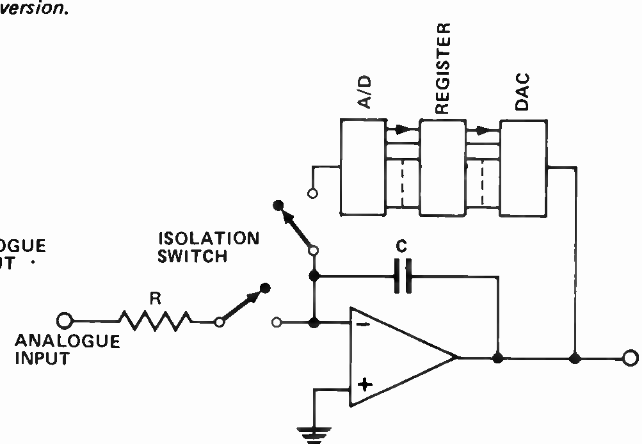


Fig.23. Infinite hold is obtained by transferring the integrated value into a digital store using A/D and D/A conversion.

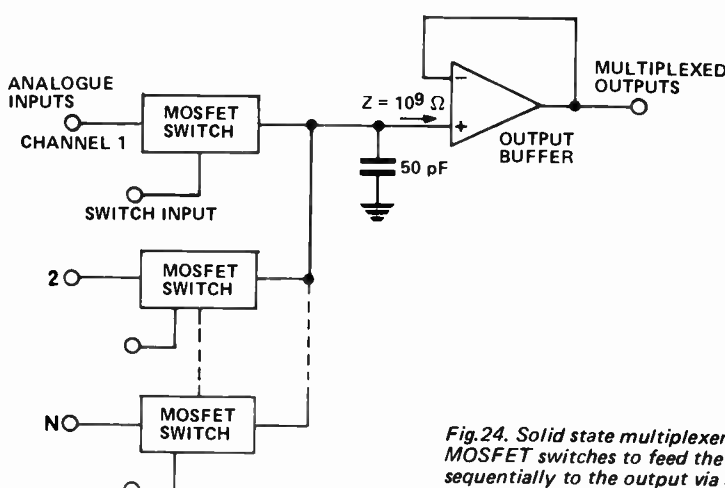


Fig.24. Solid state multiplexers use MOSFET switches to feed the inputs sequentially to the output via a buffer stage.

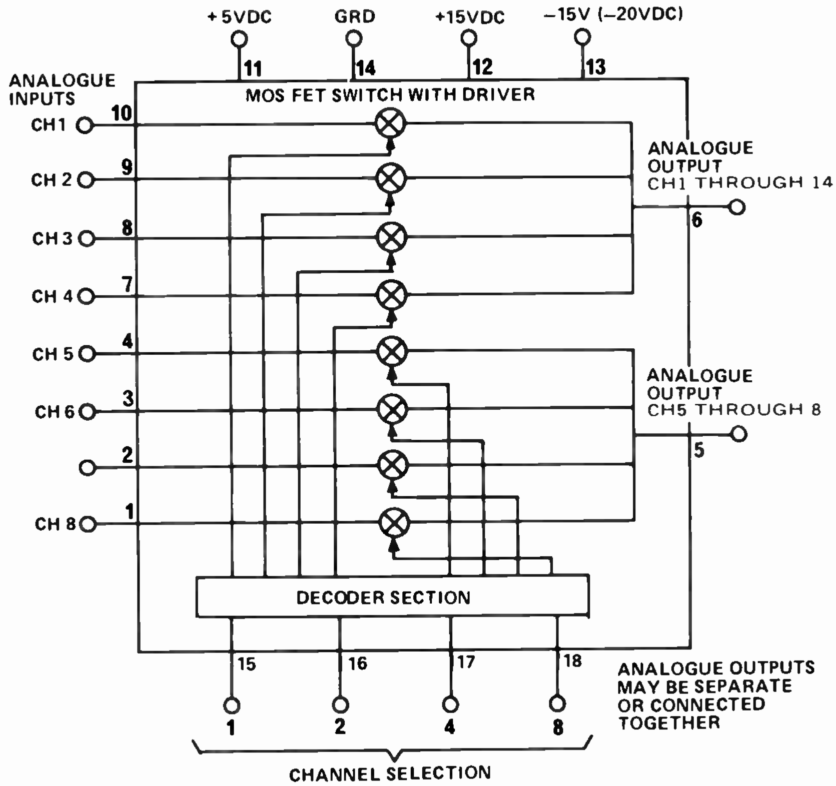


Fig.25. Internal schematic of Datel 8 channel, MM-8 multiplexer.

REFERENCES

A comprehensive discussion of the topics, and a long bibliography, of this part is to be found in "Analog-digital conversion handbook", D.H. Sheingold, Analog Devices, U.S.A. 1972. Less extensive but nevertheless very useful articles are — "Engineering product handbook", Datel Systems, CAT-T99405, 1974, U.S.A.

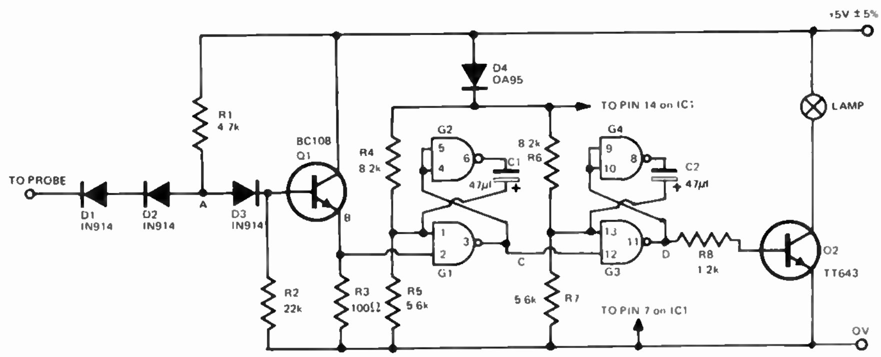
"Analog-to-digital conversion techniques", E. Renschler, Motorola Semiconductor Products Inc., AN-471, 1969, U.S.A. "Product Guide", Analog Devices, 1975, U.S.A. ●

Measurements on digital circuits... the logic probe

In analogue circuits, performance is checked or faults found, using, at the very least, a multi-meter. It is used to check the voltage levels at various places as they can provide much information about the operation. An oscilloscope is very much more useful but is not always available. The simple analogue reading meter can provide a good start.

In digital circuits, voltage levels contain far less information about circuit condition as information is contained as one of just two voltage levels. For digital work, another simple tool exists. It is called the 'logic probe'. These are usually self-contained units — they may require an external voltage supply — and are used to check the logic state of a circuit at a chosen point. As an example consider the circuit given here (from ETI project 113).

Connection of the probe input to the circuit point will cause one of five circuit conditions to operate the in-built lamp in a different mode. It is suited for use with logic of RTL, DTL and TTL systems. It is not necessarily useable on CMOS logic as that can use as wide a variety of levels as the designer selects.



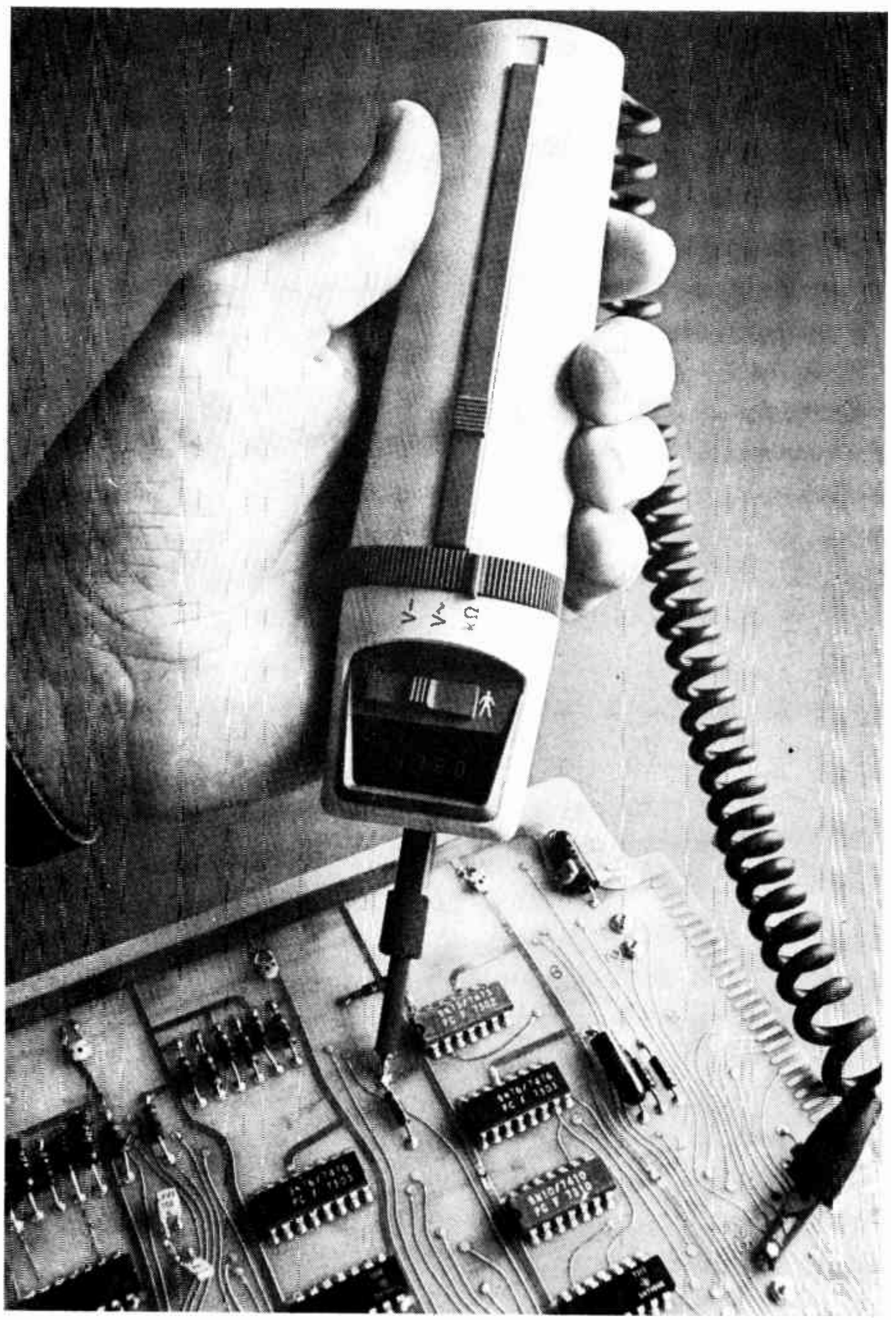
IC is any 7400 series.

SPECIFICATIONS

Supply voltage	5 Volts dc ± 5%
Input voltage (high input)	> 2.4 Volts
Input voltage (low input)	< 0.8 Volts
Indication — steady positive	lamp on
Indication — steady ground	lamp off
Indication — fast positive pulse	lamp flashes on
Indication — fast negative pulse	lamp flashes off
Indication — pulse train	lamp glows less brightly
Minimum detectable pulse width	50 nanoseconds
Extended indication of lamp	100 milliseconds

Specifications of supply and operation for logic probe.

28 Digital instruments and test equipment



This compact hand-held digital multimeter is produced by Hewlett-Packard.

IN THE previous two sections we discussed the basic building blocks of digital systems. We are now in a position to study how these blocks are assembled into specific types of general-purpose instruments and test equipment.

Digital instruments may be defined as those in which the major proportion of the circuitry is digital rather than analogue. The circuitry of such instruments is seldom all digital, as few of the natural processes we require to measure are in digital form (a notable exception is counting — a digital procedure). Hence the digital instruments that are used to measure real-world variables usually have an analogue-to-digital converter at the input. However there are many modern instruments designed to 'see' into logic circuitry — such instruments (having direct digital inputs) are truly digital.

Many test procedures could be implemented using solely analogue circuit techniques. However, there now exists a definite trend to replace analogue techniques with much more complex, but cheaper, digital equivalents.

Digital measuring equipments range in size and complexity from the simple panel meter for measuring voltage, current or resistance, Fig. 1, through medium complexity, portable and highly-flexible digital instrument systems, Fig. 2, to large automatic testing plants, Fig. 3, which operate on the commands of in-built computers.

Today, many testing instruments (even quite small units) possess self-testing facilities, in-built diagnostic ability and other advanced capabilities such as the automatic readjustment of the circuit under test to bring it within quoted specifications.

SOME HISTORY OF DIGITAL INSTRUMENT DEVELOPMENT

We have seen in Part 21 that most analogue signals can be converted to a digital equivalent. This concept first found serious economic application in computational systems, as was briefly explained in Part 22.

In the early days of digital systems even simple equipment had to be built using large numbers of thermionic valves and electro-mechanical relays. In the early 1940's the power requirements and the sheer bulk of such systems severely restricted the application of digital techniques to computers. Hence the early uses of digital systems in testing and evaluation first appeared in applications where portability was not required.

The need for an adequately-fast training simulator for aircraft pilots led to the development of the 'Whirlwind 1' computer by MIT in 1946. This evolved from work previously done on the use of digital techniques in the fire control of guns.

Digital voltmeters first appeared in the mid fifties and at least two companies, Non-Linear Systems and Schlumberger, lay claim to being first in the field. In the period 1955 to 1960 new skills, developed in the manufacture of computers, were used to build smaller cheaper and more effective instruments thus commencing the swing from analogue to digital instrumention. The introduction of transistors enabled large-scale digital process control to be realised — Texaco Refinery at Port Arthur, 1959, was the first. They also allowed portable digital instrumentation to be made.

In the early 1960's integrated-circuits were conceived. The cost and space savings gained with their use removed any doubt that digital techniques were not competitive with traditional analogue methods.

THE DIGITAL MULTI-FUNCTION METER

We have already dealt with the basic sub-systems used to form most digital instruments. They usually comprise various assemblages of digital displays, A/D and D/A converters, counters, registers, gates and sources of precise frequency signals.

A digital multi-meter, for instance, consists of analogue input, function and range selection stages which feed an A/D converter; this drives the digital readout form of display. They are marketed in a wide range of forms — from the neat compact fixed purpose units like that of Fig. 4 to systems such as shown in Fig. 2, in which modules are interchangeable to suit the very wide range of possibilities that the common power supply and digital readout allow. Functions offered (by choice of appropriate modules) may include more than the traditional multi-meter measurements. Frequency measurement, timing, counting, totalizing, and ratio measurement may all be offered. Readout of any physical variable is possible provided a sensor exists that may be correctly interfaced with the digital end of the measurement system.

The simplified block diagram of the very compact H.P. 970A a hand-held digital multimeter, is given in Fig. 5. The circuits are of analogue form until the comparator stage after which

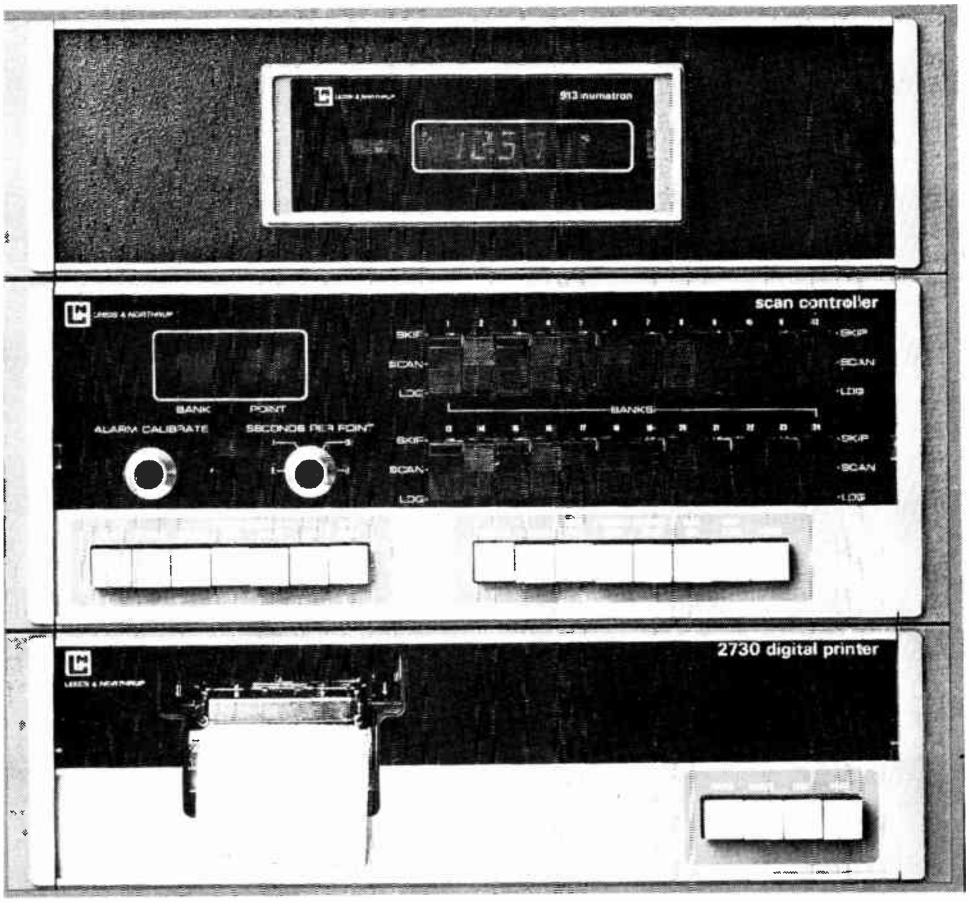


Fig. 1. Digital panel meters are extensively used in equipment, often replacing the moving needle indicator meter.

Fig. 2. The H.P. 3470 series measuring system consists of modules that clip together to form the digital multimeter instrument system required.

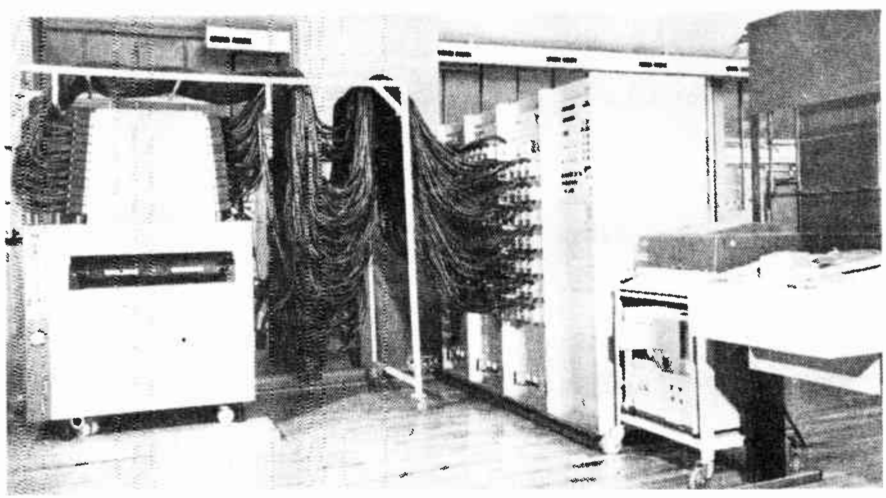


Fig. 3. Automatic Wiring Tester VD30, by Siemens — it handles 12,000 connections.

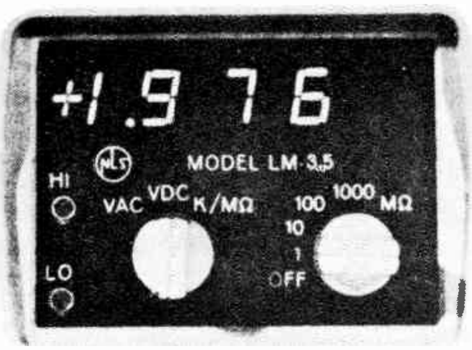
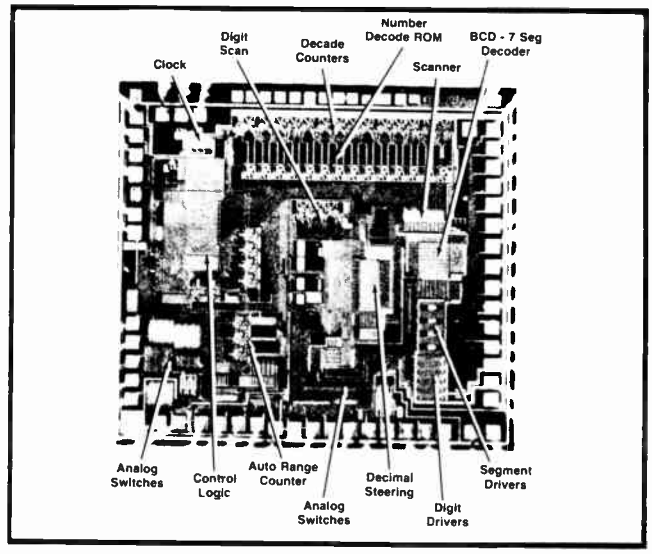


Fig. 4. This digital multimeter, by Non-Linear Systems, fits in the palm of the hand.

Fig. 6. Digital instrument circuits are often manufactured as custom-built chips. (a) Digital chip of HP 970A is just 3.9 by 4.3 mm in size.



digital signals are used. This unit uses the dual-slope integration method to convert from analogue to digital. Most of the circuitry is manufactured on just two custom-made monolithic IC chips — Fig. 6a shows an assembly with 40 flip-flops, 19 MOSFET switches and some 3500 bits of ROM (read only memory); Fig 6b shows the chip which carries the bulk of the linear circuits used.

Multimeter measurements of alternating current or voltage are only meaningful if the waveform is sinusoidal. If the waveform is complex it is necessary to use an oscilloscope to gain knowledge of the ratio of peak to average or rms as measured by the multimeter. There are instruments available which incorporate the multimeter and CRO functions within the one case. Such an instrument is the Tektronix type 213 unit as illustrated in Fig. 7. This instrument displays either the waveform or the scan-generated multimeter measurement as required.

A modular system approach is more expensive but gives versatility and continued flexibility to adapt to changing needs by not-so-expensive additions. Figure 8 shows one maker's modular approach — it retains the aesthetic shape of a complete portable set regardless of the number of modules needed at one time. The wasted space and extra weight penalty of the slide-in modular package is thereby avoided.

COUNTER–BASED DIGITAL INSTRUMENTS

Combining a display with a suitably gated counter and timing system provides the ability to count events; totalize and indicate elapsed time; determine frequency and period of periodic waveforms or provide a time-clock. In these options little analogue circuitry is involved, the unit either generating its own digital signals (a clock) or operating on input signals that are already in digital form.

We have dealt with the internal operation of counters in Parts 24 and 25: here we expand their use by studying the various modes of operation possible with a basic counter.

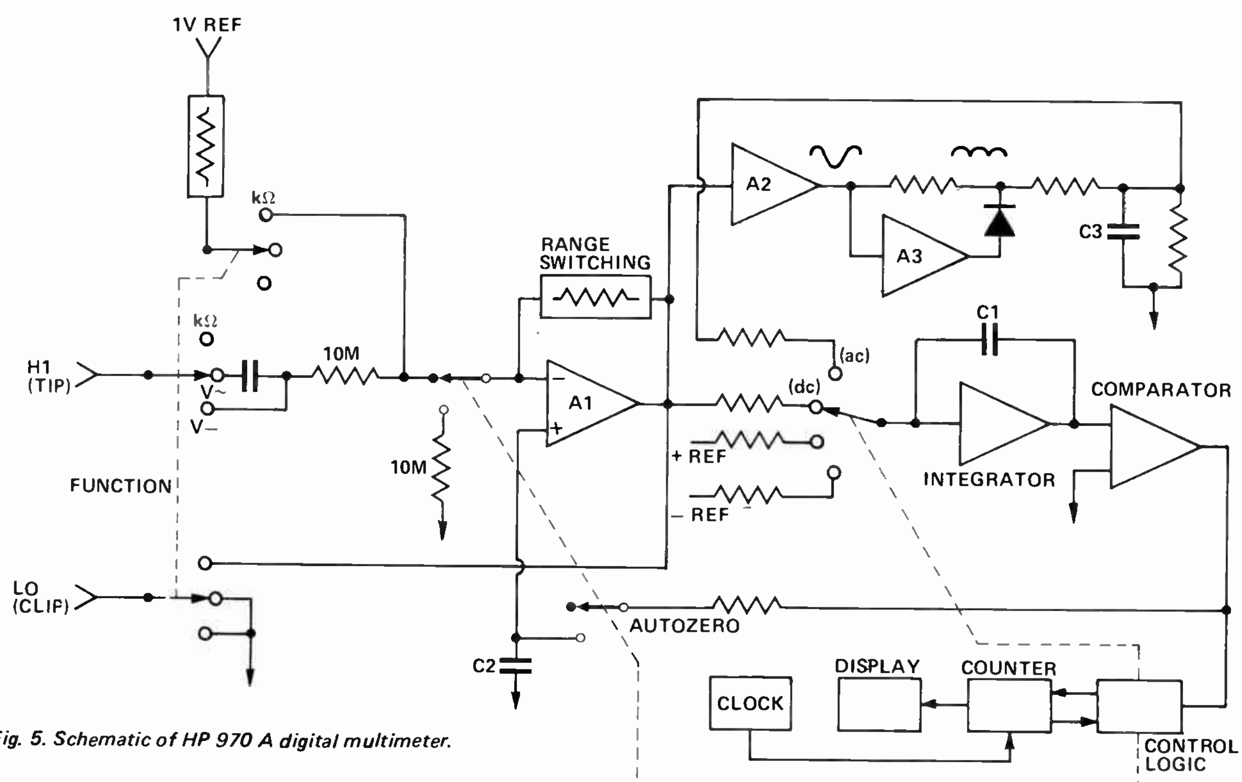


Fig. 5. Schematic of HP 970 A digital multimeter.

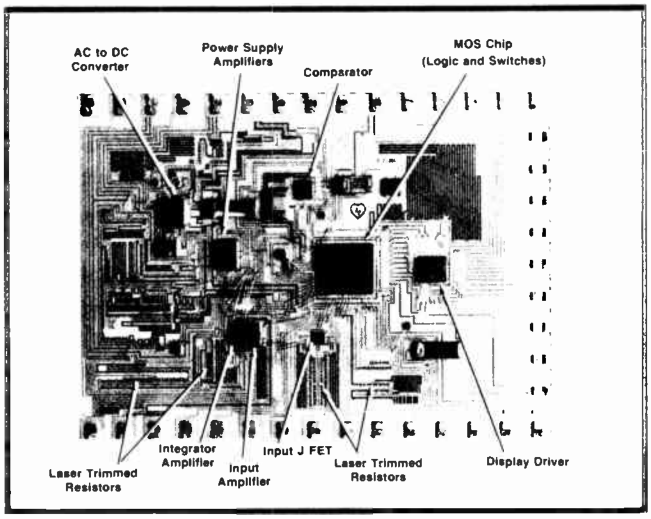


Fig. 6(b). Thin-film hybrid of HP 970A carries much of the analogue circuit on its 28 x 38 mm substrate.

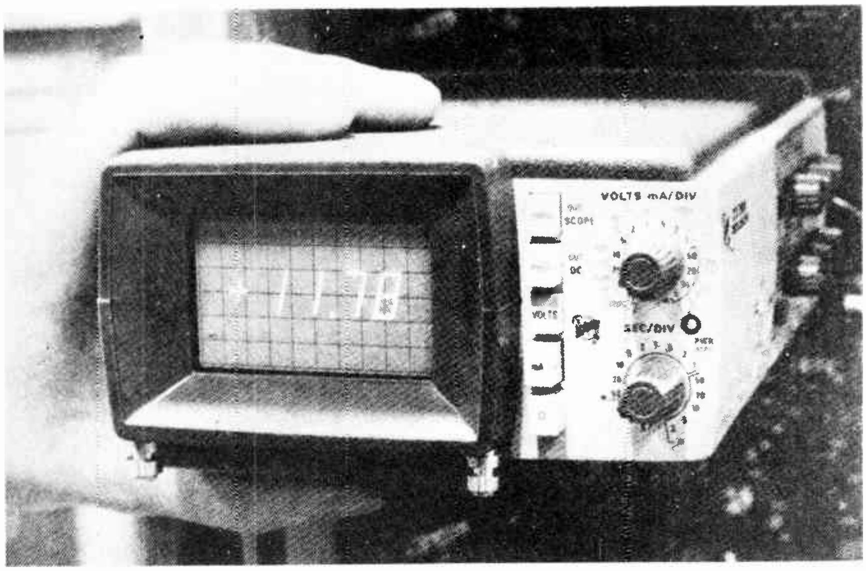
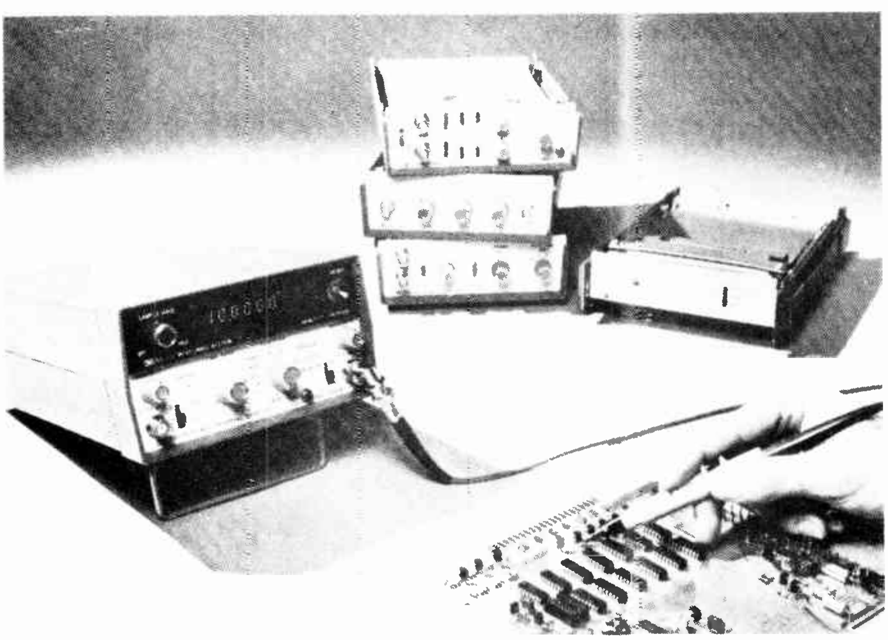


Fig. 7. The digital multimeter function is combined with a C.R.O. facility in the Tektronix 213 unit.

Fig. 8. Compact HP5300 series of modules allows many counting/timing measurements to be made.



TOTALIZING AND BATCHING

This is the simplest use of a counter. Events to be counted (for example packages on a conveyor belt may intercept an optical link, thus causing an electrical pulse to be generated each time the beam is broken). These pulses enter the counter, (Fig. 9) and are shaped into clearly recognisable counting pulses. Whether or not an event is counted is decided by the condition of the input gate which can be opened or closed on electronic command. In many applications the gate is quite simple, but its design can be a major problem when very fast signals are to be totalized to high-accuracy.

A batching counter goes a little further in that it counts to a predetermined value. When this is reached it provides an output command to the process being batched (for example, tins being counted into cartons) which causes some change in the process. At the same time, if the process is repetitive, the counter is reset to the starting value ready to count the next batch. It is sometimes more convenient to count downward from the number required, operating the batch command at the zero value. More complicated batching systems may have a stored program that sets each batch sequence to varying count values.

TIME INTERVAL MEASUREMENT

It is possible to measure the time-interval between two events by feeding pulses of known time separation into the input of the counter, as shown in Fig. 10, where a clock drives the counter. An example is the timing of a race. The gate is opened at the starting signal and stopped at the end. Counts accumulated represent the time interval. Obviously there is an advantage in choosing a pulse repetition rate that suits the units of time being used. The choice of clock frequency therefore depends on the resolution needed. For example, to measure one second with a resolution of 1 in 10^6 a pulse frequency of 1 MHz is needed to gain 1 μs discrimination.

In some applications a common gate – control input is suitable – where the on and off event reproduces the same situation such as in period measurement of a sine-wave signal. Often, however, two separate input channels are needed so that each can have the specific pulse conditioning needed by different signal sources. Race timing, for example, might initiate the count on the sound of the starter gun and stop it on the signal from a pressure-pad sensor.

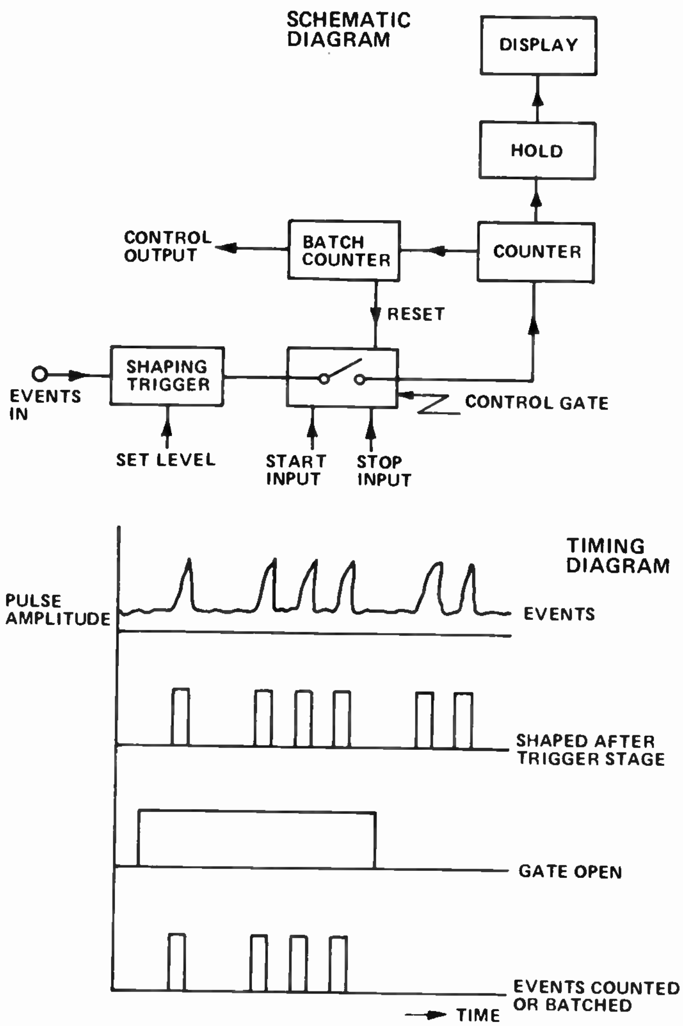


Fig. 9. Schematic diagram of totalizing and batching use of counter with display.

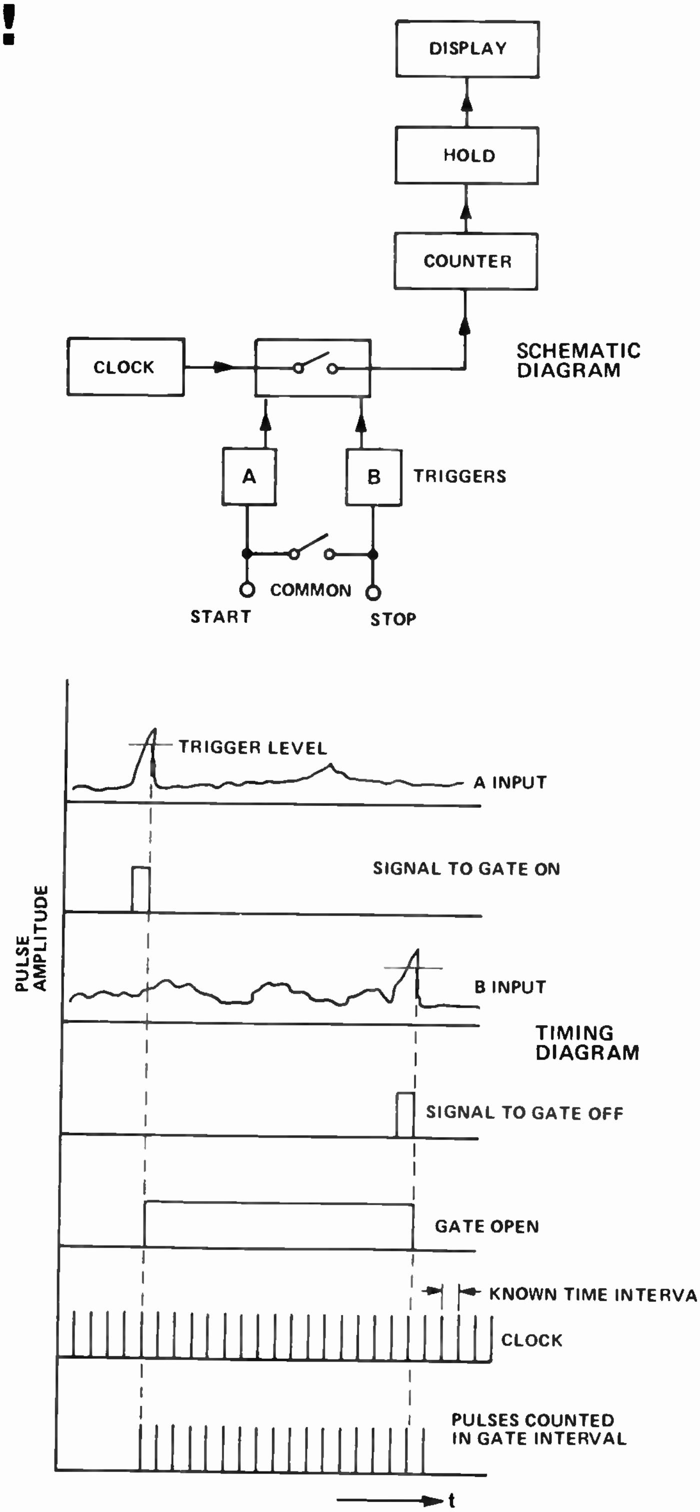


Fig. 10. Addition of a clock to the totalizer yields time-interval measurement with this configuration.

Because many triggering signals are ill-defined in time, most counter/timers have input stages that trigger at preset adjustable levels – Schmidtt triggers or comparators are used. This enables the operator to discriminate the events to be counted from relatively noisy backgrounds that have a lower peak value – see later. Another reason for selective-level triggering is to allow the counter to operate at different points on a waveform – a sinewave input can be used to produce pulses of varying widths in this way. The counter output may also be used to trigger an event so as to provide automatic timing sequences, this being similar in priniciple to batch counting.

PERIOD OF REPETITIVE SIGNAL

The time interval measurement arrangement also enables the period of a wave-form to be measured. The most basic procedure is to gate the clock into the counter for the interval between the same trigger level of successive waveforms. The precision can be greatly improved by extending the gate-open time to 10, 100, 1000, or 10 000 periods, dividing the count by the appropriate divisor (which means a mere shift of the decimal point if decimal multiples are used). This is referred to as multiple-period measurement, it gives greater precision but at the expense of greatly increased time for each measurement.

PULSE WIDTH

A special case of period measurement occurs when the width of a pulse is to be determined. If the pulse had a perfect square response profile the on and off gating points would always give an accurate answer because the triggering transitions would occur precisely on the rise and fall of the pulse. Trigger-level would not affect the interval measured. Practical pulses, however, will not be perfect, the edges having definite rise and fall times. In this case the trigger-level becomes critical in width determination, as is depicted in Fig. 11. Counters usually provide a slope selection control. This decides whether the trigger operates on the positive or negative slope, that is, a or b slopes respectively (Fig. 11).

PHASE DIFFERENCE

Two identical waveforms can be regarded as two separate inputs for the start-stop inputs. If both trigger at the same point on each waveform (preferably at the zero-crossing to gain maximum precision) the ratio of the time-interval between the two crossings to the period of the waveform is the phase shift in terms of a fraction of one cycle.

FREQUENCY MEASUREMENT

The frequency of a repetitive signal is defined as the number of cycles (events) per unit time. A digital frequency meter, therefore, can be made with controlled gate-on period of precisely known time interval. In other words, it is the same combination as the interval timer (i.e. oscillator counter with display, gate control and precision) but with the difference that the clock now controls the gate (not the counter) through a precision divider that scales the basic clock frequency down to obtain the gate duration needed. Gate periods range from 1 μs to 1 s. A simplified schematic is given in Fig. 12. The input signal is often sinusoidal; the input stage shapes this into a square wave to enhance individual cycle detection by the counter.

NOISE ERROR REDUCTION

The above descriptions give the basic operating modes of the various counter/time/frequency-meter combinations. In practice a number of refinements may be incorporated to obtain better practical performance.

Noise can be reduced by incorporating a fixed amount of backlash in the trigger circuit; this produces what is called the trigger window. On the way up the trigger level is at a higher level than on the way down, as shown in Fig. 13. Provided the noise added to the signal has an amplitude smaller than the window width, the counter will only trigger once on the way up and once on the way down. This method works well for high-frequency measurements where the noise is usually a small percent of the signal-plus-noise signal amplitude.

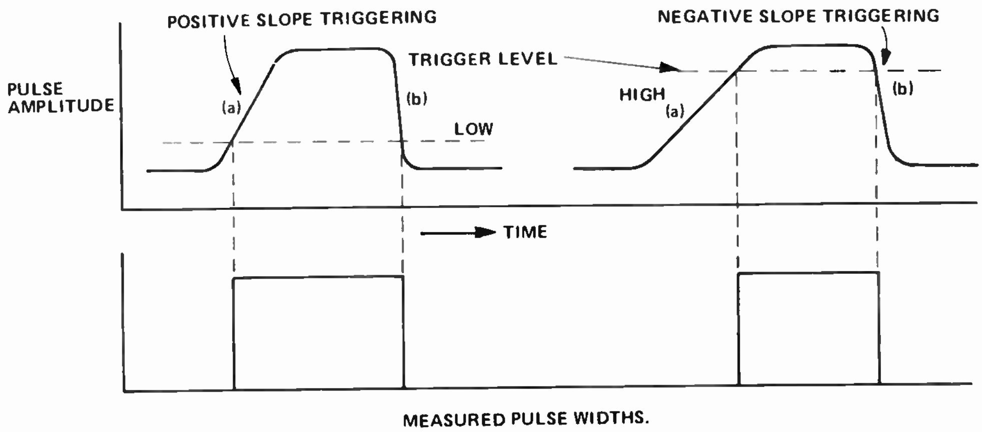


Fig. 11. Trigger level must be considered in pulse width measurement to obtain the parameter required.

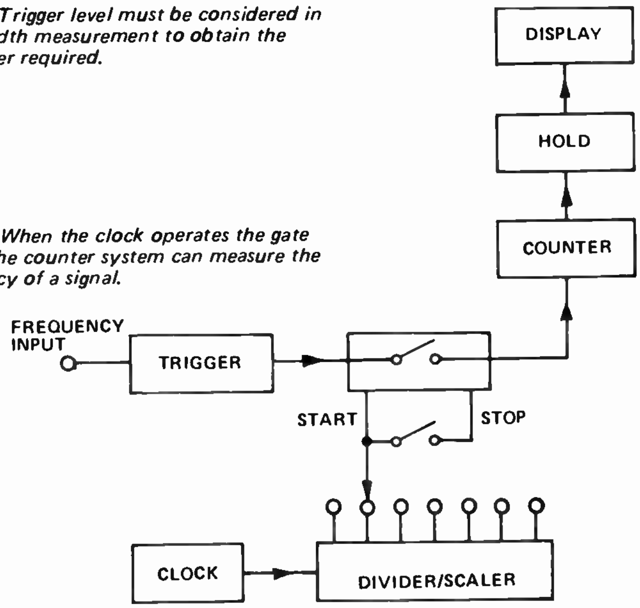
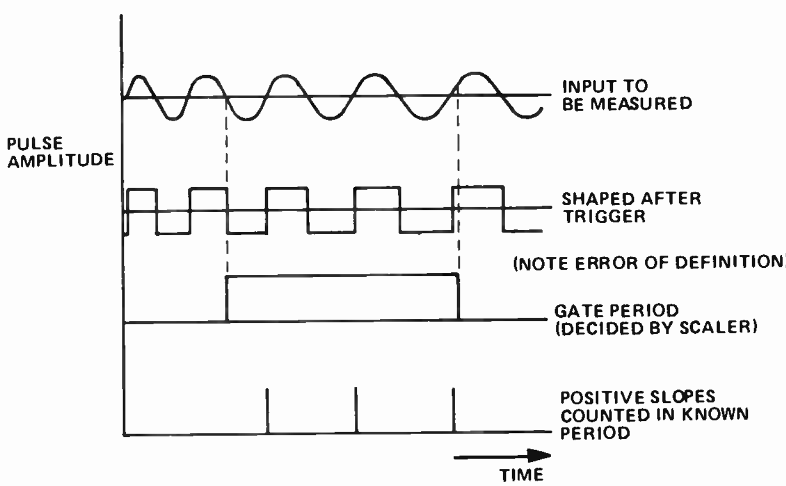


Fig. 12. When the clock operates the gate length the counter system can measure the frequency of a signal.



Low frequency measurements can often involve interference sources that produce rapid spike transients. One simple method of reducing this is to use filters. Advanced designs contain filter systems that reject all frequencies higher than that being tested, the appropriate filter being automatically selected by the counter itself after it has made a determination of the frequency of the signal.

A recent approach to the noise problem is to set up a time-window (as opposed to the trigger height window) that, once the counter gate is on, inhibits the off-state chance until after

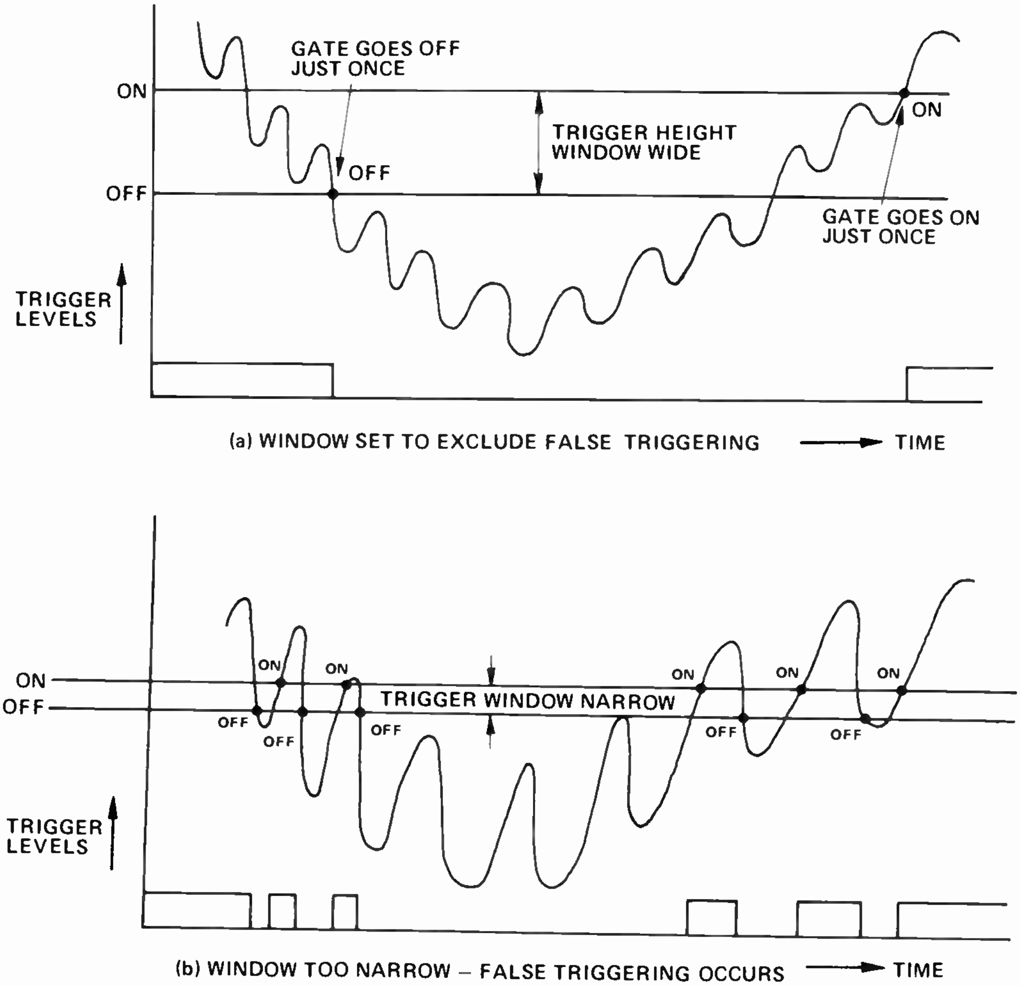


Fig. 13. Noise reduction in counting by use of a trigger height window.

a time just shorter than the expected interval. This is known as trigger masking. It is very useful in eliminating contact-bounce retriggers.

Before using a counter/timer on an unknown waveform it is, where feasible, good practice to study the waveshape on an oscilloscope in order to decide the best strategy for trigger-level and height-window width settings. Figure 14 illustrates the differences between window level settings on various waveshapes.

As the readout is in digital form it is necessary to hold the display at the determined value for a period long enough to allow the value to be read. Some units incorporate a control that gives the operator a choice of hold time.

The following part of this series will continue with this general discussion of digital instruments, covering physical variable transducer systems, the various kinds of analysers and correlators, waveform generators and computer-controlled testing systems.

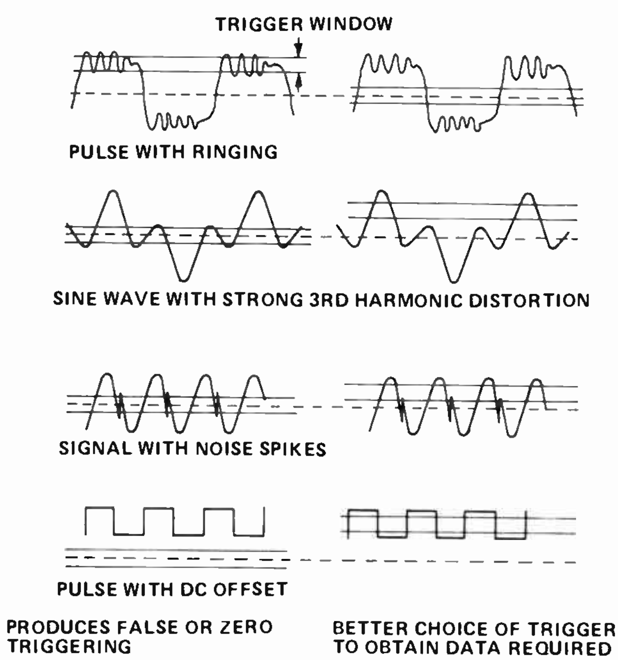


Fig. 14. Trigger height for the window must be chosen to suit the wave shape.

REFERENCES

Varies issues of Hewlett Packard Journal contain many detailed articles on a wide range of digital instruments.

"Digital Instrument Course – Pt. 2", A. J. Bouwens, Philips Industries Holdings, Sydney, is useful to have. This booklet describes the practical use of counter/timer instruments in many varied applications.

29 More about digital instrumentation

VERY FEW VARIABLES to be measured by electronic means provide a digital signal directly: this is because the real-world is predominantly analogue by nature. Consequently, most so-called digital measurement systems involve a number of stages to make the signal compatible with the digital circuits of a system.

The most straight-forward 'digital' measurement method (at least at present), is to employ a suitable analogue sensor that provides a voltage (or current) output, related to the variable being measured. This signal then feeds an A/D converter to obtain a digital equivalent.

The low cost of digital calculation circuitry now enables linearization of sensor processes at moderate cost. Figure 1 gives an example of the digital linearization used in a thermocouple thermometer unit.

By referring to Fig. 2 we see that the linearization process, in the dual-slope digital voltmeter section of the system, is achieved by changing the ramp slope at a number of points. To do this a rate multiplier is used to multiply the clock frequency by a variable number, N/256, where N may be any number between 1 and 256. By this means 256 different ramp slopes may be generated. The slope in use is tracked by a segment counter which, in turn, causes a read-only-memory (ROM) to set up the correct digital-readout code.

Some sensing principles lend themselves to a more direct digital signal approach. For example, in the Moire-fringe displacement sensor, a grid of fine lines (called a grating) formed on glass is attached to the moving (or fixed) member of the machine whose movement is to be monitored. This is shown in Fig. 3. The other member carries a small index grating set to produce Moire-fringes which move as the two grids pass relative to each other. Movement of these fringes is monitored by photocells which provide a number of electronic pulses proportional to the magnitude of the displacement. These pulses can be counted directly with a reversible direction counter, thus allowing both directions of movement to be followed. Such a unit is known in the

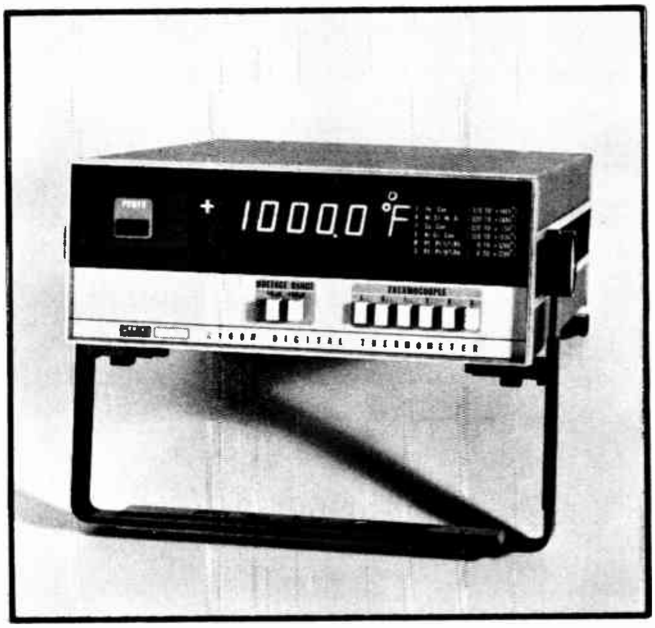
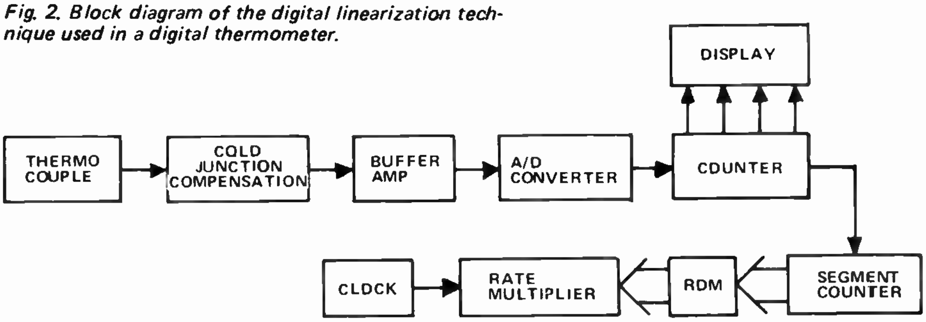


Fig. 1. This Fluke 2100 series digital thermometer incorporates digital linearization. It is suitable for use with any one of six common thermocouples.

Fig. 2. Block diagram of the digital linearization technique used in a digital thermometer.



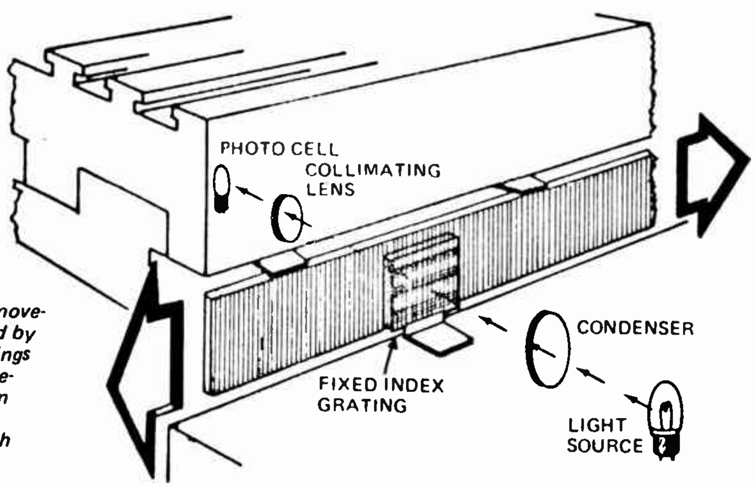


Fig. 3. Accurate measurement of movement is performed by using optical gratings to produce "moire-fringes" which can be counted as the gratings move with respect to each other.

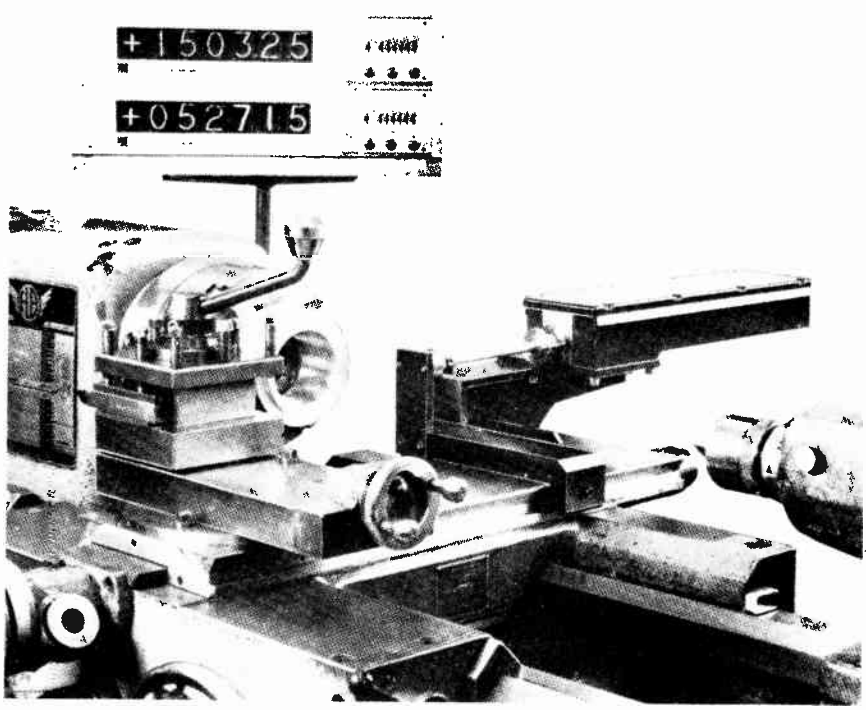


Fig. 4. This lathe is fitted with a digital readout of displacement. The self-contained Heidenhain linear sensor is to be seen high at the back of the cross slide.

metal working industry as digital readout (DRO) – a typical system is shown in Fig. 4. A somewhat similar length measuring system is the laser interferometer – this also provides fringes that can be counted to provide a measure of absolute displacement. Clearly in such cases the digital instrument must not lose or gain stray counts or else the wrong value is indicated. Such systems are called incrementals.

There is an alternative method which uses digitally encoded discs similar to those shown in Fig. 5. This is an absolute method which is not subject to pulse loss or gain, or to power failure errors which occur in the previous system if not fitted with a special non-volatile memory. The discs of such a system are read optically as though they were registers or other forms of digital store, each position having a different digital code as read across a radial line.

In some forms of digital pulse transducer it is the rate of pulse production that represents the variable, not the absolute number of pulses. An example of this sensor is the turbine flow meter used to measure liquid or gas flow. Figure 6 shows such a flowmeter where a small turbine rotates inside a pipe at a speed related to the flow rate. Rotation may be converted into pulses using optical, magnetic, capacitive or, in earlier designs, mechanical sensing. This form of sensor provides a variable frequency output which can be converted by a counter/timer system into a direct readout of flow rate.

Digital transducers are somewhat similar. They provide a signal which varies in frequency as the variable being measured changes. The sensor of such a transducer is made such that it alters a parameter of a frequency generating circuit. For example the quartz-crystal thermometer shown in Fig. 7 operates in this manner. In this unit temperature causes the resonant frequency of a crystal, mounted in the end of the probe, to change in a predictable fashion.

It is interesting to note that many natural physiological sensors operate on the pulse-rate system – neurons (the digital nerve sensors on the end of the nervous system) trigger with pulse repetition rates that rise in accordance with the intensity of the actuating signal (heat, cold etc).

Considerable effort has been expended – especially in the Eastern European countries in the late 60's – to produce reliable low-cost industrial sensors that provide a digital form of output. These have not, however, been accepted to the extent hoped. The current low-cost of extremely powerful digital circuits, however, is likely soon to produce a trend toward sensing devices having digital output.

ANALYSERS

Analysis is the general process used to break down an unknown by methods which separate and distinguish basic elements of seemingly complex arrangements, the elements so derived being satisfactorily understood basic quantities.

Synthesis is the alternative approach wherein a system is built up from known elements to produce the complex case.

Analysis may be regarded as being

Fig. 5. The absolute digital code disc acts as a store of the angular value of displacement. It is interrogated as though it were a register whose value is read optically across a radial line.

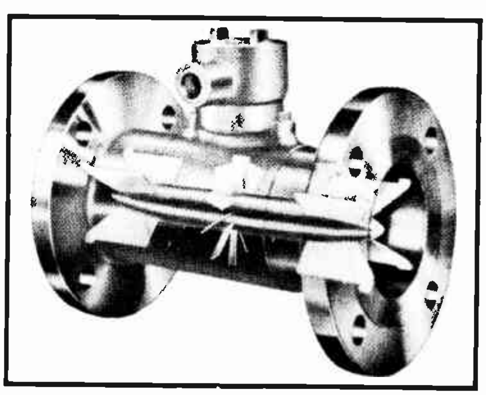


Fig. 6. Turbine flow meters like this one provide an output in the form of a pulse train with a pulse repetition rate proportional to flow rate.

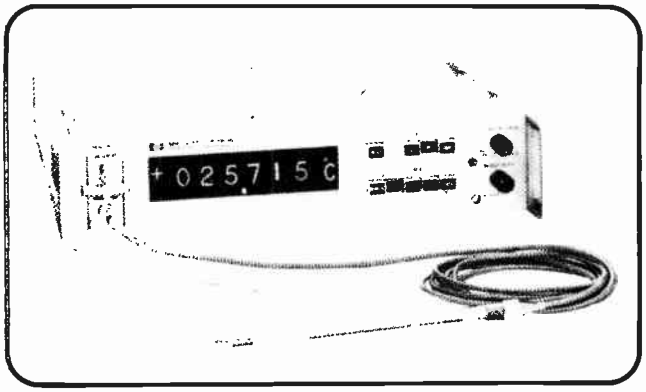


Fig. 7. Quartz-crystal thermometer from Hewlett Packard.

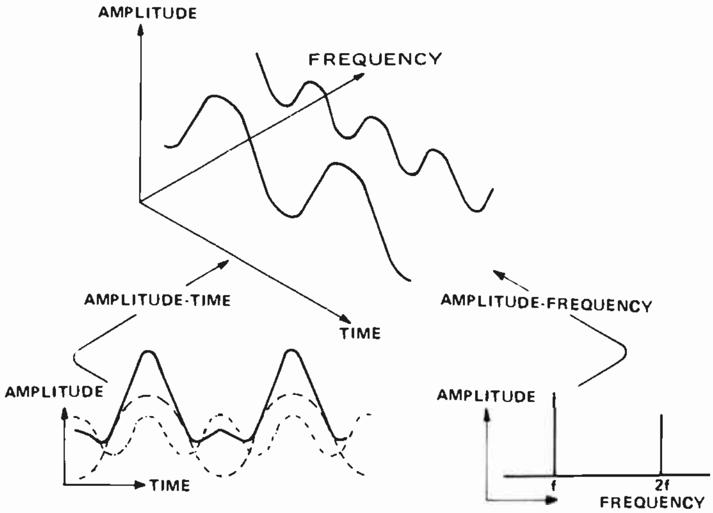
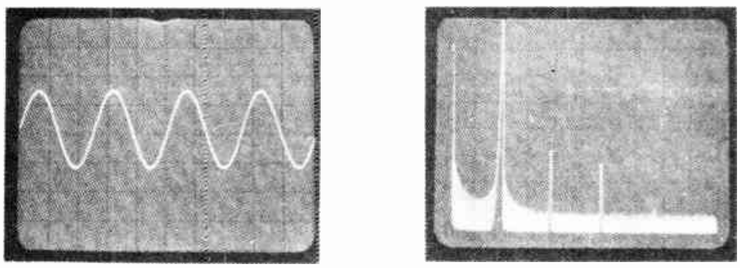
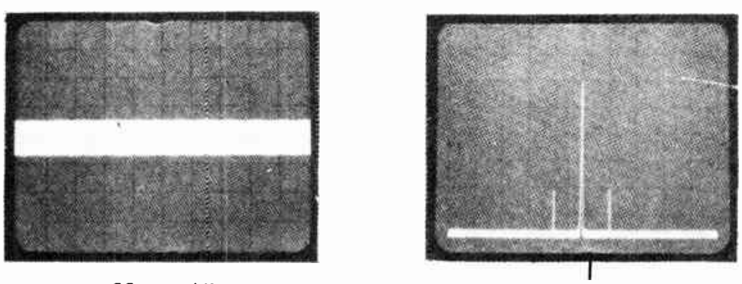


Fig. 8. Second harmonic distortion is not always easily seen on an amplitude versus time display. In the amplitude versus frequency display the second harmonic distortion and its amplitude are clearly seen.



2 µsec/div *0* *1 MHz*

Fig. 9. Frequency domain displays are often better than the time-domain method as these HP displays illustrate: (a) In the time domain (left) the signal looks pure but the spectrum analyser shows that it has significant distortion.



20 msec/div *15 MHz* *200 kHz/div*

Fig. 9b. A 2% amplitude modulation is barely discernible on time domain plot (left). The frequency domain plot clearly shows the frequencies present and their amplitude.

required when the behaviour of an existing system needs to be *studied.* Synthesis is used when a system is to be *devised.* There are of course many instances when both approaches are used to yield a solution.

Various types of electronic analysers are used in electronics. We will look here at spectrum analysers, logic state analysers and pulse-height analysers as these types are commonly met in modern circuit work. Each of these operates on an existing electrical signal breaking it down into frequency content, logic-state content and height of pulses, respectively.

SPECTRUM ANALYSERS

Signals in the time-domain, that is those displayed as amplitude versus time graphs, can also be displayed in terms of their amplitude-versus-frequency and phase-versus-frequency characteristics. (This was discussed in Part 4 where an example wave-form — a square wave — was broken up into its harmonics). The relationship between time, amplitude and frequency are seen by studying the three forms (shown in Fig. 8) of a fundamental sinewave having a large degree of second harmonic added in. Signals displayed as amplitude (or phase) versus frequency are said to be in the frequency domain. This kind of plot shows the frequency spectra of the signal, hence the name spectrum analysers.

The role of spectrum analysers is to display the signal content in its frequency domain form. There exists many instances where this form of display is better than a time-domain representation. Typical examples are where a fundamental has distortion (Fig. 9a) or where low levels of modulation or noise exist (Fig. 9b). Neither of these conditions could be satisfactorily detected, let alone measured, by a time-domain test.

Basic spectrum analysers use analogue circuitry and therefore do not qualify properly for inclusion in a discussion on digital instruments. However, as we will see later, the current trend is to include digital techniques in such instruments. Advanced analysis equipments, for example, often use a built-in digital computer.

There are two alternative forms of spectrum display. First, the repetitive signal can be studied over an extended time period by scanning across the expected frequency-range with narrow band-pass filters. A speedier, but more expensive method, works in a real-time mode thus preserving the time-dependency between signals. These are known as swept-tuned and real-time spectrum analysers respectively.

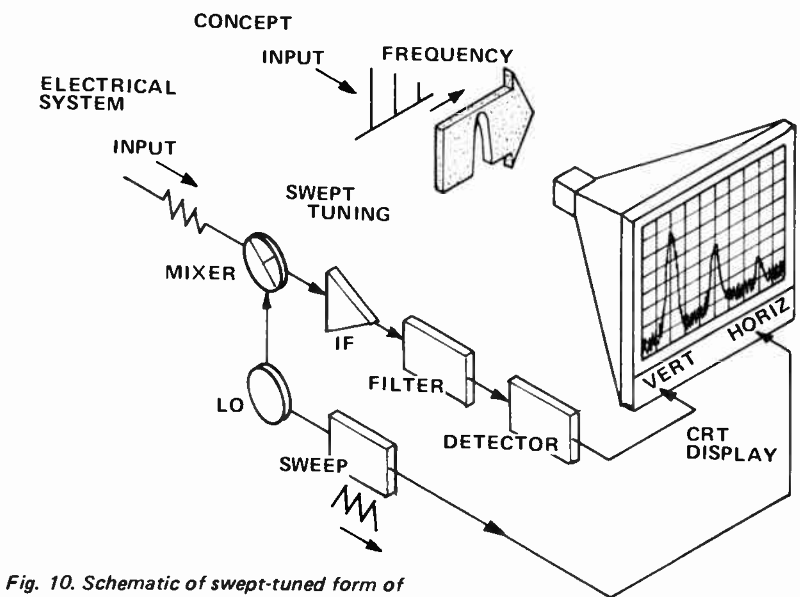


Fig. 10. Schematic of swept-tuned form of spectrum analyser.

Swept-tuned systems – Basically the task is to establish the amplitude (and sometimes phase) of the signal at each frequency in turn. Many practical difficulties exist because the absolutely narrow band filter does not exist and even if it did, it would take an enormous time to sweep it across the full bandwidth of the signal. Practical filters also have finite bandwidth and roll-offs. The bandwidth of the filter may also need to change if the requirement is for a filter bandwidth that is always a given proportion of the signal frequency as it sweeps the range.

Most difficulties are overcome by mixing the signal with a swept local oscillator and then detecting the output and using it to drive the Y plates of an oscilloscope. The sweep signal drives the X plates. Figure 10 depicts this arrangement.

Real-time systems – These use a stack of band-pass filters and detectors each connected to the signal simultaneously and with each having staggered centre frequencies. This is shown schematically in Fig. 11. The scan generator multiplexes the individual channels in order to produce a continuous spectrum on the oscilloscope screen.

It is clear that this method is much more expensive because many filters are needed. It does, however, enable a detailed analysis of once-only transient signals which could not be analysed with the swept-tuned arrangement of a spectrum analyser.

A range of spectrum analysers is available for the study of signals from 5 Hz to 50 GHz. Different instruments (or the use of different plug-ins with the same display unit) are needed because units typically cover only 4 to 5 decades, that is, say 5 Hz-50 kHz, 10 kHz-300 MHz and so on. The range is, however, ever widening. Wide range, however, is not always the virtue needed for spectral resolution is related to width of display screen.

Fourier Analysers – A third method of providing a frequency analysis is based on direct mathematical calculation using the Fourier transform technique to convert a time-domain signal into its frequency-domain equivalent. Such systems are extremely expensive compared with the above analysers, but provide a vastly greater capability.

They can also handle signals at the very-low-frequency end – dc to 100 kHz is typical. Their operation is quite different from the above in that the signal is fed as data values into the analyser unit via keyboard or paper tape from another computer or mass-storage system. It can also be fed in as an analogue signal from, for example, magnetic tape. The heart of the Fourier analyser is a micro-programmable computer system which can be set to compute using various programmes such as the so-called Fast-Fourier method of analysis. The same unit may also be able to carry out correlations between signals, plus many other processing techniques.

Digital circuitry in spectrum analysers – Digital circuits are being added by manufacturers to enhance the performance of analysers. Advantages claimed include operating ease and better placement of controls. Digital storage of the display signals can be used to enhance the display brightness and to allow a spectrum to be 'held' for comparison against a second spectrum obtained later. Digital

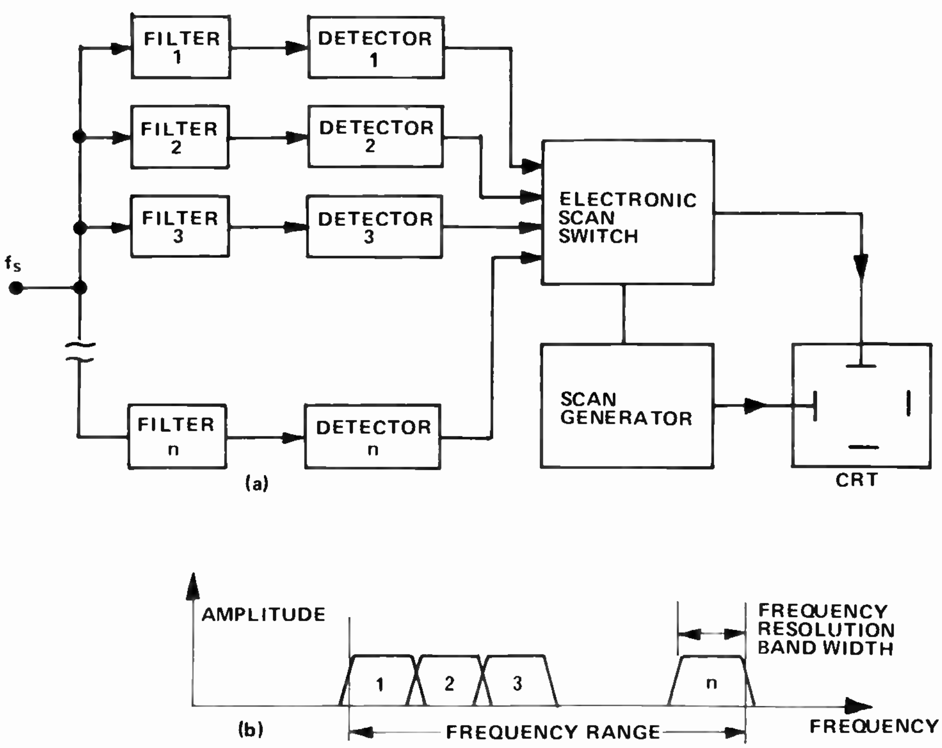


Fig. 11. Block diagram of real-time spectrum analyser based on stacked filters:
(a) schematic
(b) frequency response showing individual filter windows.

included; this reduces the noise thereby enhancing the signal/noise ratio on the display — as is illustrated in Fig. 12. Character generation (using digital methods) has been incorporated to display the relevant graph-axes factors — as shown in Fig. 12. The same unit also uses a photo-optical absolute-digital code disk to replace the mechanical switch usually used in a range control-knob.

Spectrum analysers are invaluable and are finding increasing use. Successful use is, however, a matter of experience and frequency-domain techniques are not dealt with as extensively as time-domain ones in training programmes. More details are available in the reading list — we can only provide the most elementary introduction here.

Logic-State Analysers — We check the operation of analogue circuits by measuring signal levels and frequency spectra at various points in the circuit. Digital circuits are different in that they contain the signal information in the form of multi-digit 'words' made up of two-state bits. To check operation, therefore, we must ascertain simultaneous logic-states at various points in the circuitry. The simplest analyser for this work is a probe which indicates logic hi or lo state at a selected point; coloured lights are used as indicators. A store function can be built-in to the probe to catch a short transition that would not otherwise be seen in the lamp display. It must also have connections suitable for PC board digital circuitry — see Fig. 13.

The single probe can be used to analyse the state of a circuit by moving from point to point in turn. To speed-up the analysing process a more extensive facility to use would be one that simultaneously shows the logic states of multiple points in the so-called Data-domain. The

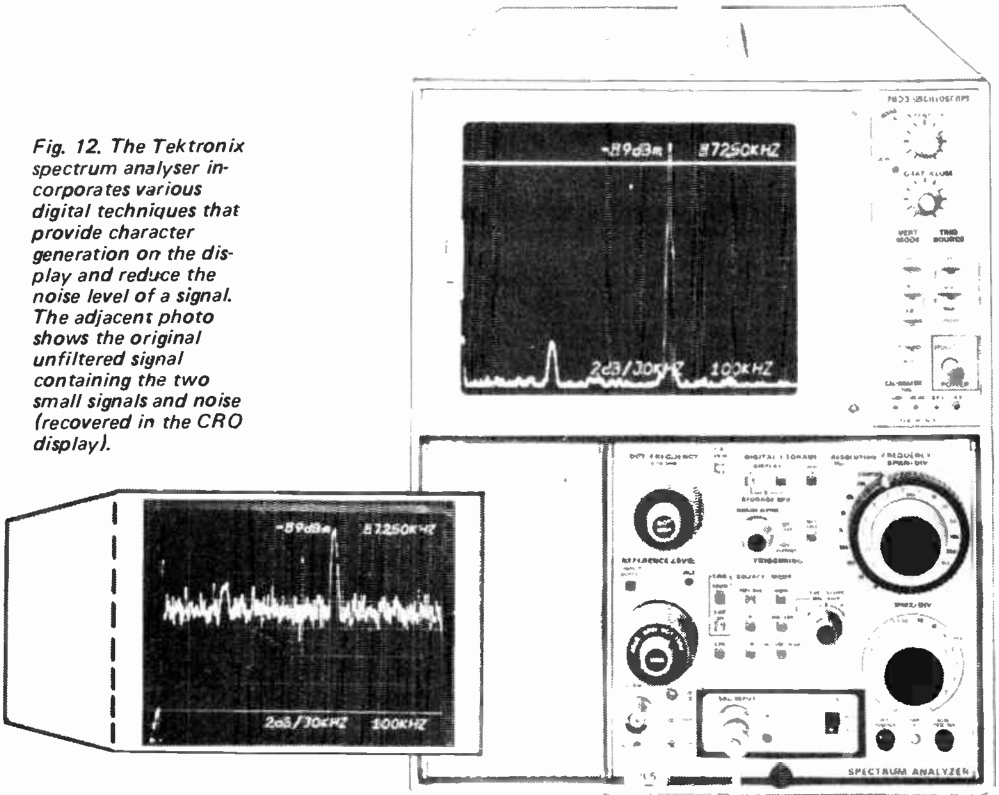


Fig. 12. The Tektronix spectrum analyser incorporates various digital techniques that provide character generation on the display and reduce the noise level of a signal. The adjacent photo shows the original unfiltered signal containing the two small signals and noise (recovered in the CRO display).

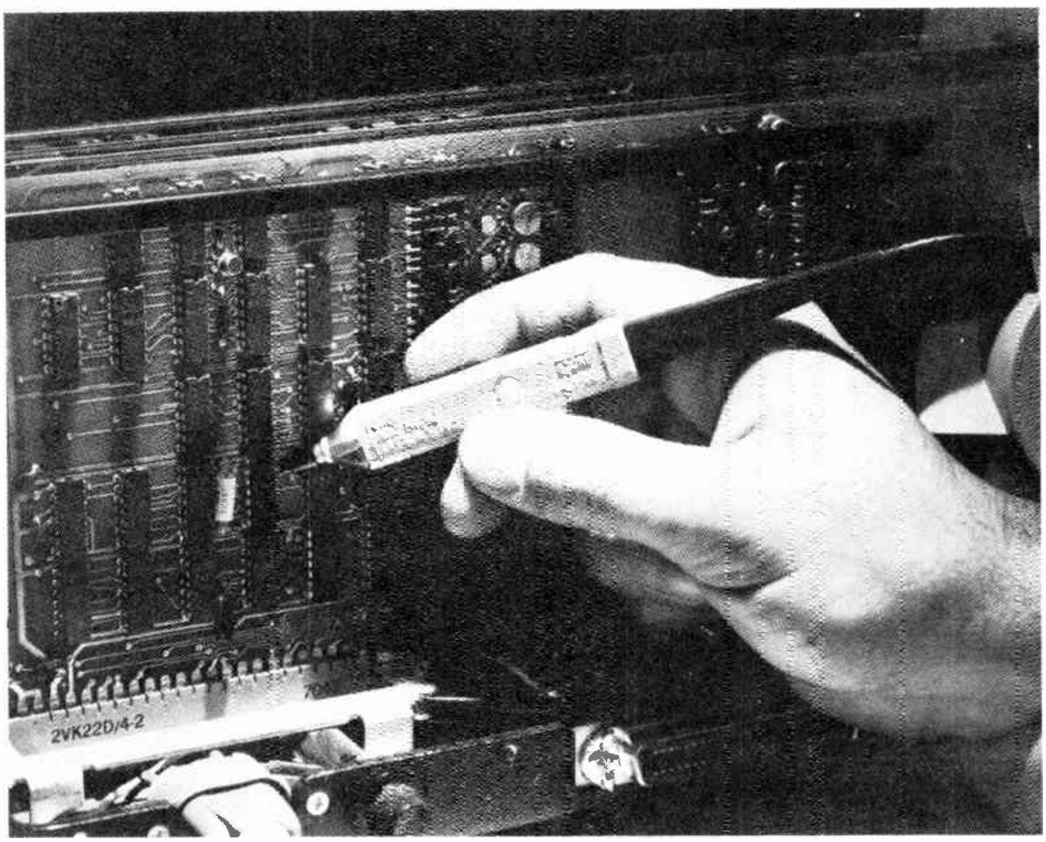


Fig. 13. Simple logic probe in action.

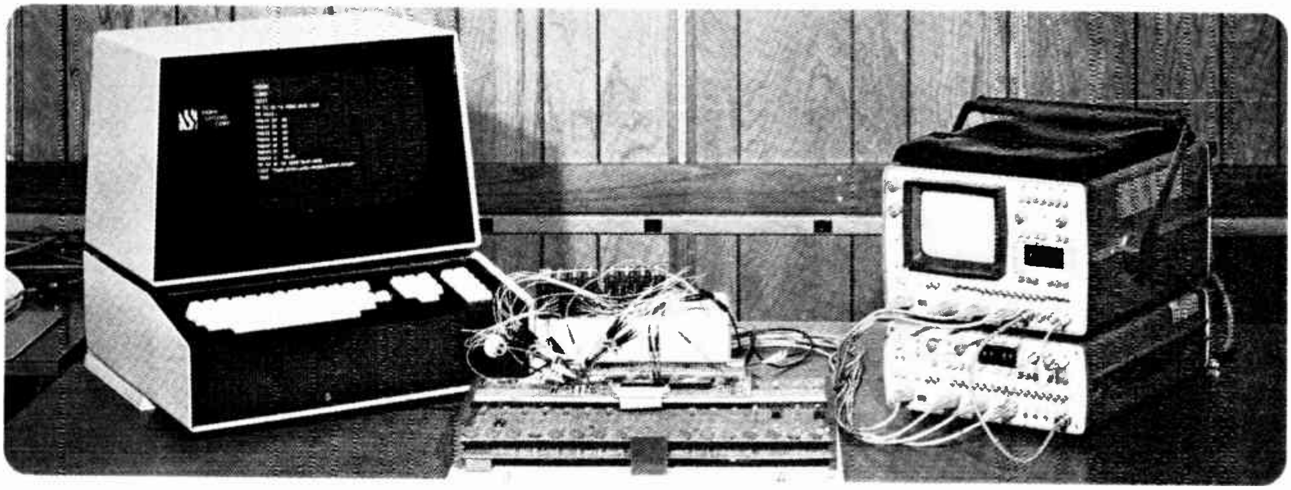


Fig. 14. Hewlett Packard logic-state analyser as set up to test a printed-circuit board card.

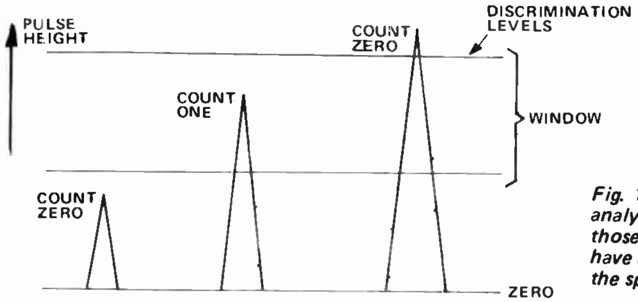


Fig. 15. Pulse-height analysers count only those pulses which have amplitudes within the specified window.

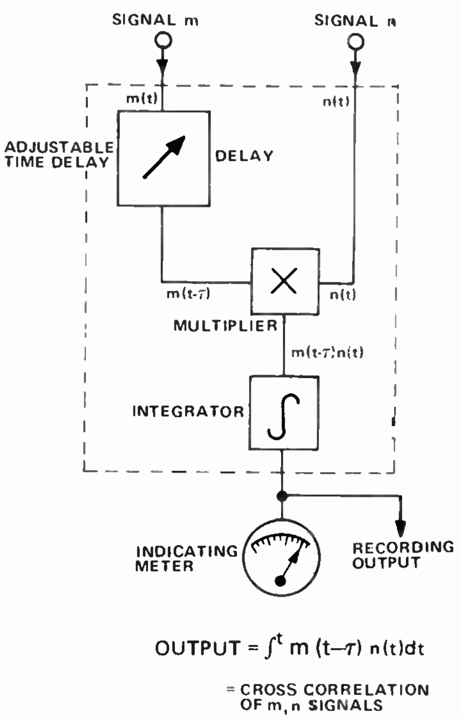


Fig. 16. Block diagram of the mathematical function performed by a correlator unit.

Hewlett-Packard system, for example (shown in Fig.14), displays over 500 points as a matrix of 0's or 1's on a CRO screen. These instruments are used to debug, test or trouble-shoot complex digital circuits. Only large laboratories, however, would be able to justify the cost of such advanced logic analysers.

Pulse-height analysers (Discriminators) – Measurement processes involving ionising radiation and sometimes light-intensity levels rely on pulse counting, the pulses appear as rapid electrical currents produced from a photo-multiplier or ionisation detector. The relative amplitude of a pulse often distinguishes it from pulses from other sources. For example, different radio-active isotopes produce pulses of different energy, enabling an assay of radio-active mineral to be made by a study of occurence of pulses of different height. Photons arising from the various noise sources in photo-multipliers have different energy from those generated at the photo-cathode. This is a true detection process: noise can be reduced by discrimination of pulse heights.

Pulse height analysers use carefully selected trigger levels to accept only those pulses (for counting) that arise from the particular source of interest. Pulses above the trigger window, or below are rejected (not counted) as demonstrated in Fig. 15.

CORRELATORS

Correlation as a physical process can be visualized by considering two identical optical patterns formed on film transparencies. When the two are exactly overlaid, the maximum transmission of light occurs. At any other position the light transmission falls off to a minimum at greatest misalignment. The mathematical process modelling this is that whereby the two patterns (represented as formulae) are multiplied together and the multiplicand signal is averaged as depicted in Fig. 16. This process is repeated using a slightly different spatial (or timewise) phase-shift between signals each time the sums are performed. The same idea of correlation applies for time-domain as well as space-domain signals. When two identical 'patterns' are compared in this way it is called auto-correlation. When two patterns that are not identical are compared it is cross-correlation. Correlation brings out any similarity as a peak in the correlator output when the greatest degree of correlation is achieved.

Early correlators used entirely analogue methods to carry out the signal processing. Today, correlators mainly use digital techniques in the manner shown schematically in Fig. 17. The two inputs are converted from analogue to digital form in the quantizers. One channel is progressively delayed by storage in a shift register. Multiplication and

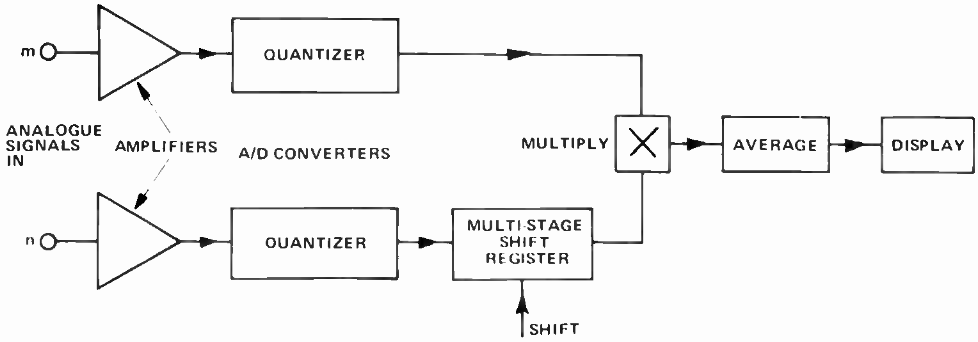


Fig. 17. The simplified block diagram of an auto correlator.

averaging are also achieved by digital methods.

Correlation finds use in the detection of periodic signals buried in noise; in establishing the degree of coherence between random signals; in establishing transmission time and source of transmission of a signal; in identifying the characteristics of a complex system (the input is perturbed with a noise-like signal which is cross-correlated with the output to give the transfer functions of the system). It is a particularly powerful tool and we should see more of its use if the cost of digital correlators continues to fall.

SYNTHESIZERS

These are a special kind of signal generator in that the signal output is formed by addition of a number of sources or by manipulation of a single, stable-reference frequency. A music synthesizer provides a whole range of musical sounds by combining many different tones into a single output. Although synthesizers work upon basic analogue signals the trend is to combine or modify the signals using digital control.

The advantages offered are (in the variable frequency generator kind of synthesizer) that a very stable reference oscillator has its frequency translated to (literally) billions of other values (the HP 8660 gives 10 kHz-2600 MHz) whilst retaining high stability. By pressing digital-key inputs, any chosen frequency value is generated. It is also possible to control the output via a programmable BCD digital input. Programming enables an enormous range of signals to be synthesized, a typical requirement being as part of an automatic test procedure. Figure 18 shows the philosophy of the HP 3330 series of automatic synthesizers with a typical programme card marked up for a frequency sweep routine.

Digitally-controlled power sources may be used to synthesize varying voltage (or current) levels over a test period at the commands of a mini computer in the same way as the above unit synthesizes frequencies.

Frequency and voltage synthesizers are often combined in the hybrid-computer (digital and analogue combined) in order to generate synthesized signals which are needed

a.

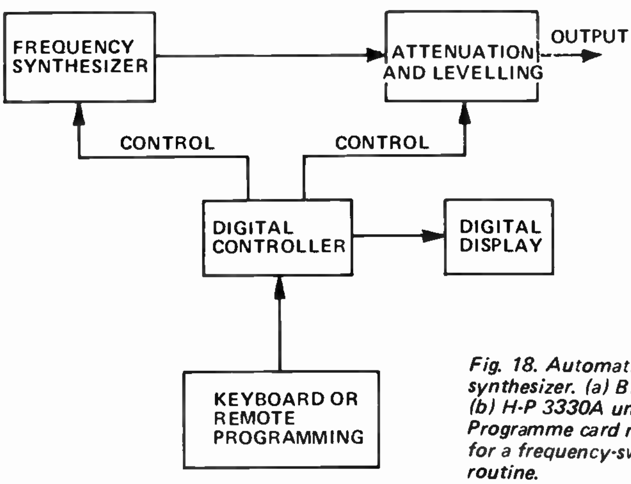
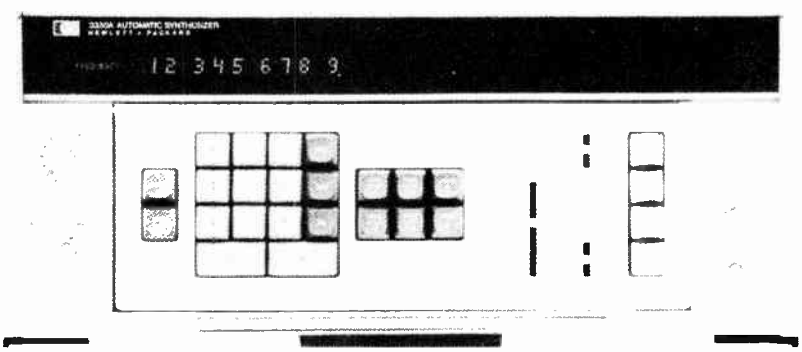
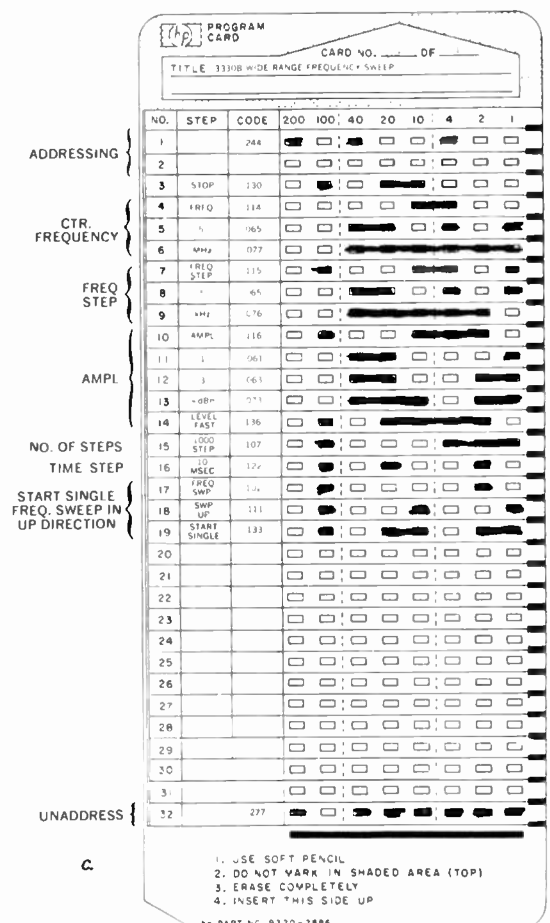


Fig. 18. Automatic frequency synthesizer. (a) Block diagram. (b) H-P 3330A unit. (c) Programme card marked up for a frequency-sweep routine.

b.





c.

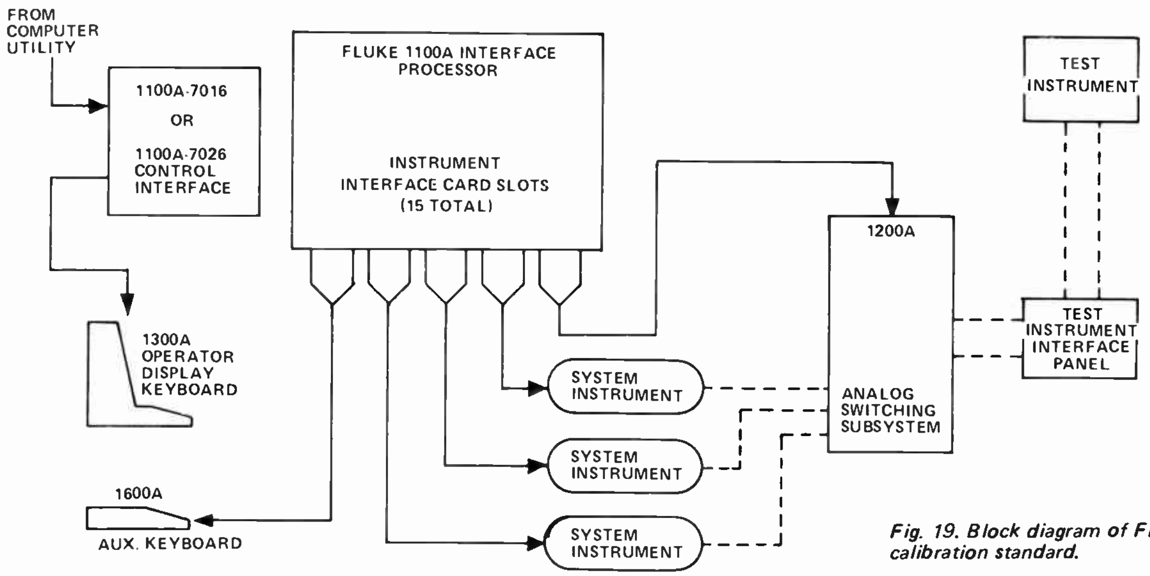


Fig. 19. Block diagram of Fluke calibration standard.

to derive a simulation of a complex system, such as a missile in flight.

COMPUTER CONTROLLED TEST SYSTEMS

With the enormous increase in complexity of routine complex processes (such as aircraft instrumentation and controls, refineries, automatic and large-volume manufacture of electronic systems) came the need to improve and speed-up the testing procedures needed to check out the thousands of different parameters involved. Computer controlled testing is far more reliable than human operator testing and is extremely fast. It can be economic even for the testing of small volume electronic equipment, especially where a large range of tests is involved.

The instrument or process to be tested is interfaced to the main test console which usually incorporates a wide range of facilities that are chosen with flexibility of operation in mind. Figure 19 shows an automatic system used to calibrate a test instrument. The test programme must be devised by a highly-trained professional designer, but once developed and programmed the testing can be performed by a less trained person.

It is not possible here to deal in depth with automatic testing as the range of requirements and equipment available are both great. The overall concept and scope of an automatic test system is shown in Fig. 20. Suffice to say very complicated automatic testing systems are in routine use in a wide variety of manufacturing and maintenance situation.

REFERENCES

Digital Sensors – these are variously described in the many general books on sensors but rarely as digital concepts, specifically. Recent releases are:

"Transducers in measurement and control" – P.H. Sydenham, UNE Publishing Unit, Armidale, 1975 (Available ETI).

"Transducers in Industrial Measurement" – P.H. Mansfield, Butterworths, London, 1973.

"Analysers – detail is available in the data sheets and journals of companies such as Marconi Instruments, Hewlett-Packard, Tektronix, Honeywell, Spectral-Dynamics, Bruel and Kjaer and others.

"Counting Photons", ETI, November, 1974 and "Electronics in Medicine – Pt. 2" ETI, August, 1975, dealt with pulse-height analysis.

Hewlett-Packard offer a series of varied booklets on Spectrum Analysers, as well as video tapes. Application Note 150 provides basic understanding.

Synthesizers of the music kind are discussed in "Electronics in Music", F. C. Judd, Neville Spearman, London, 1972. A complete project of a music synthesizer was covered in articles in ETI beginning October, 1973. Part 1 contains reference to digital control of the synthesis process.

Manufacturers' data sheets are a source of reference to both laboratory synthesizers and programmable dc power supplies (which are also referred to as programmable power – DACs).

Auto Testing Systems have been described in a number of ETI articles. "Computer-interfaced instrumentation in the Development Laboratory", April 1972 discussed the potential. ●

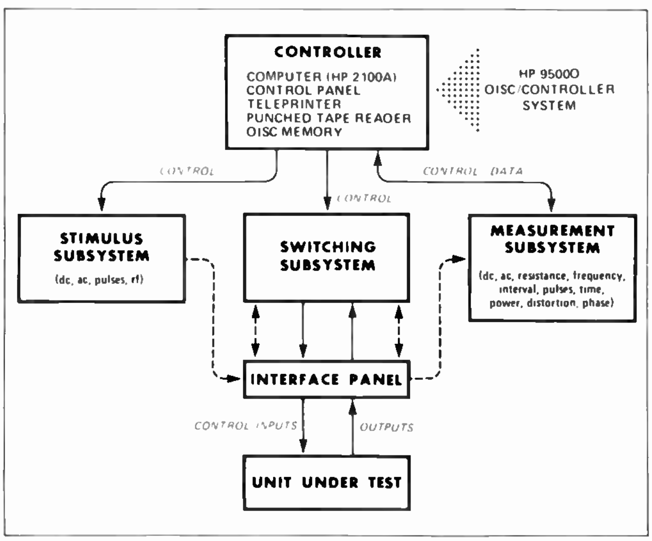


Fig. 20. Overall concept of HP9500 series of automatic test systems.

Notes

30 Digital computers

THROUGHOUT THIS COURSE we have been steadily building up sufficient information to enable discussion of computing machine operations. What follows is necessarily an introduction only — computers are now extremely sophisticated in design and the manufacturing methods very specialised. It is, however, quite important that the operation of computers be understood by electronic craftsmen at a general systems level. This, and the next part, will introduce the philosophies, the hardware and the operation of digital computers from a technical rather than user-only viewpoint. (Analogue computers were briefly mentioned in part 12 — they are still valuable in some applications but in general, machine computing is now mainly done digitally).

Already we have introduced the systems approach of understanding analogue and digital electronic systems (Parts 1, 21); how information can be conveyed in binary code form and how different channels can be handled simultaneously on a common transmission line (Part 5); how square-wave clock signals are generated was covered in parts 17 and 18. The history of the development of logical operation by electrical switches and how logic gates operate to perform simple arithmetic was covered in Part 22 — other basic digital functions being dealt with in Part 23. Storage of digital numbers in solid-state counters was the subject of Part 24 and the conversion of signals from code-to-code and from analogue-to-digital, and vice versa, were discussed in Parts 26 and 27, respectively. How computers become involved in testing was briefly mentioned in Part 29. These are each important concepts, worth revising at this stage.

With this much learned there is little else that need be studied about components (stores of various kinds and IC developments in computers are discussed later) in order to understand how digital computers operate. Our emphasis must now be on the design arrangement of the computing machine as a whole and how the user can make it work.

WHAT IS A COMPUTER?

Regardless of whether a computer is digital or analogue in operation its role is to perform various kinds of mathematical operations. The analogue machine cannot perform logic operations: (unless cojoined with a digital computer, in which case it is known as a hybrid computer — as shown in Fig. 1) its use is generally restricted to what are called linear mathematical problems in which signals vary continuously and information is transferred as levels not as digital codes. Analogue computers can be very good at such operations, often better than a digital computer of similar cost. The digital machine, on the other hand, (a general purpose installation is shown in Fig. 2) can perform almost any kind of mathematical manipulation, however special techniques are often needed to solve analogue problems. Analogue type signals must be sampled and each sample converted into a digital equivalent before they can be processed in digital machines: this is where the digital machine in certain applications may be less efficient than the analogue alternative.

As well as performing arithmetical operations (called scientific computing) the digital machine can be instructed to process or sort discrete data in digitally encoded form (called

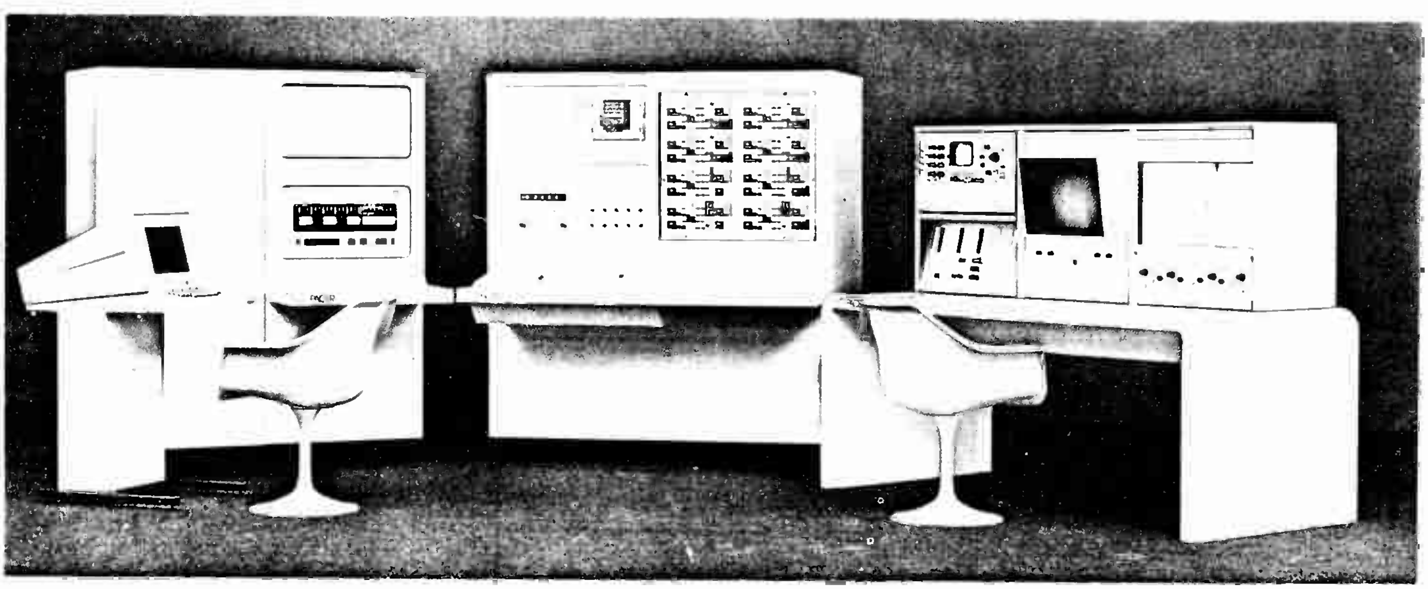


Fig. 1. Here a digital computer and an analogue computer are combined — the result is known as a hybrid computer.

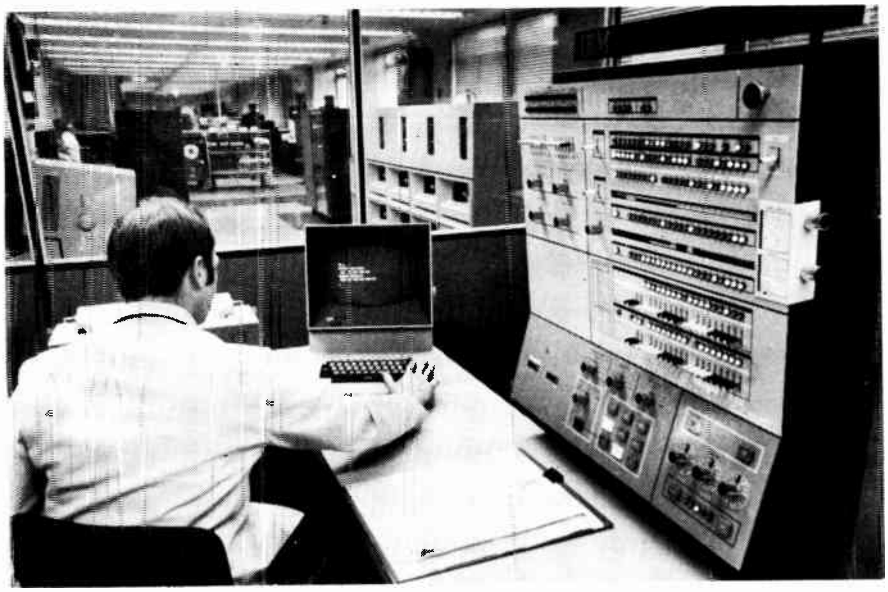


Fig. 2. General purpose digital computer, this centre is used by Lloyds Bank in Britain.

data processing or DP, for short). Typical computer data processing operations are the sorting of numerical data — for example to see how many people have heights of various chosen values, or the booking of airline seats. Mixed working, where scientific calculation and data-processing are both involved, occurs for example, in costing out a building estimate, raising a stock value for a business, or producing pay-slips.

Digital computers may also calculate tables by automatically incrementing the input data between preset limits. For example the computer could be asked to generate and print the sines of all angles between 10° and 90° at 1° intervals.

We pause now to note that we call such machines computers *not* calculators. The term calculator has traditionally been used to describe machines which perform a fixed set of mathematical calculations. The term computer on the other hand, is reserved for those machines which may be reconfigured by a set of programme instructions to perform any particular task. However such distinction between the roles of calculator and computer is becoming increasingly difficult to make. Some computers are now dedicated to performing calculator like tasks and some calculators are now so flexible that they can be programmed to perform a variety of tasks.

In the 1950s, when powerful electronic computers were emerging, the popular concept was of a machine that would soon have thinking powers of its own — and its own will and imagination — as depicted in Fig. 3. Although we must concede such is probably possible one day — no one has yet gained an inkling into how this extra facility could be realised. Computers are merely machine slaves that, if working internally as the designer thinks and intends, will perform as commanded. The operator informs the machine of its job via the programme presented to it. Where the computer has valuable merit is in its ability to perform calculations and process numerical data at rates vastly greater than a human mind, with rarely an error, and for hours on end if need be. It is a tool and no more. To say the computer accidently sent the $1,000,000 bill to Bill Blogs is entirely incorrect. The programmer or the machine did not perform as hoped through one or the other being defective in the instructions given or the way they were obeyed.

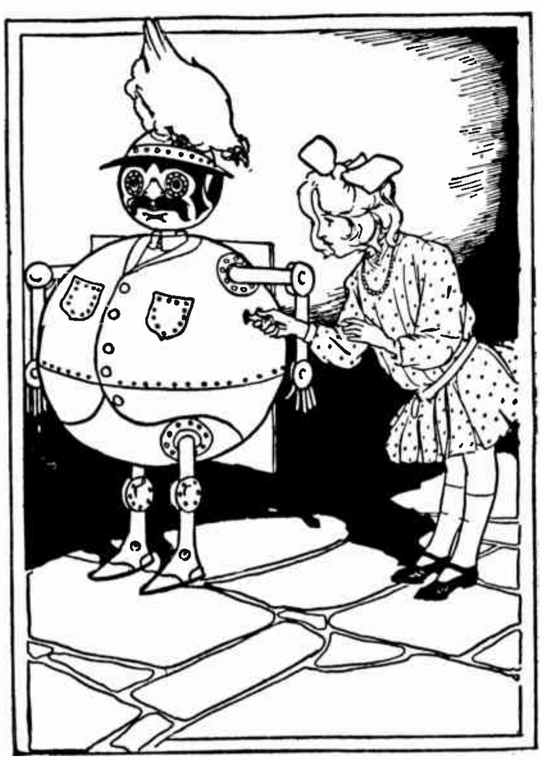


Fig. 3. As yet, computers can only do what they are programmed to do.

As well as computers that operate only when the operator gives instructions there is also the dedicated machine that, once set internally to compute or process in a predetermined way, becomes part of a process. It helps control by working at the same rate as signals are generated in the process — real time working. Process-control computers, as these are called, operate on data and perform calculations as part of many feedback loops in, say, a chemical plant. Figure 4 shows this use in a diagrammatic form. Other names variously used to describe this use are in-line, on-line, direct-digital-control (DDC) or just plain computer control. Wherever automation of extensive complex process is necessary a computer will usually be found — waste-water treatment plants, paper manufacture, natural gas and electricity distribution networks, satellite control and power-station plant operation are but a few of thousands of in-line applications. Computers are far more useful in this task than human operators — see Fig. 5.

On-line operation (although not generally agreed upon) is a term probably best reserved for cases where each of many input terminals connected to a central computer can gain access to the unit when it becomes available. This is also known as time-sharing and is used where the signal processing rate need not match the process. The computers used in

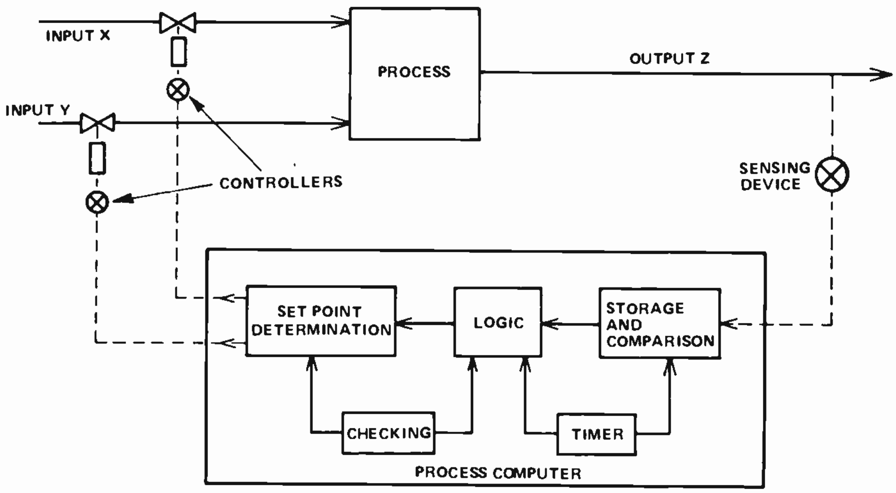


Fig.4. Process control computers are electronic data processing machines, dedicated to a specific task.

banking in Britain operate in a time-sharing mode – bank branches, as shown in Fig. 6, can gain access to the central-account records – a short wait may be necessary. When the computer works on diverse problems at the will of the operator and is not used for any dedicated purpose it is said to be off-line.

Originally electronic computers were huge – several rooms filled with racks of valve electronic circuits. In the mid-sixties manufacturing techniques and designs were such that a new style of less versatile but compact computer was marketed – the so-called minicomputer. Figure 7 shows but one kind of mini-computer system employed to control a process by providing instructions as needed. (It is not used in closed-loop as this process does not feed data back to the computer).

We do not use the word "generation" in connection with the minicomputer because that term is used in computer jargon in two distinct ways. It may describe the hardware used – first generation computers use thermionic valves and ordinary cable wiring, such as shown in Fig. 8, second generation machines use discrete transistor circuits on printed-circuit boards, third generation machines use integrated-circuitry and the most recent, about to emerge, fourth generation computers use large-scale-integration LSI manufacturing methods – see Fig. 9. A fifth generation computer is yet to emerge as an accepted concept. The other use of "generation" is in describing the system interconnections – the philosophy of system hardware interconnection and style, and capacity of the store involved.

	Machine	Man
Speed	Much superior	Lag 1 sec.
Power	Consistent at any level	1500W for about 10 sec, 350 W for a few minutes, 150 W for continuous work over a day.
Consistency	Ideal for routine, repetition, precision	Not reliable – should be monitored by machine.
Complex activities	Multi-channel	Single channel.
Memory	Best for literal reproduction and short-term storage	Large store multiple access. Better for principles and strategies.
Reasoning	Good deductive	Good inductive.
Computation	Fast, accurate – poor at error correction	Slow, subject to error Good at error correction.
Input sensitivity	Some outside human senses, e.g. radioactivity	Wide range (10^{12}) and variety of stimuli dealt with by one unit, e.g. eye deals with relative location, movement and colour.
	Insensitive to extraneous	Affected by heat, cold, noise and vibration.
	Poor for pattern detection	Good at pattern detection. Can detect signals in high noise levels.
Overload reliability	Sudden breakdown	Graceful degradation.
Intelligence	None	Can deal with unpredicted and unpredictable. Can anticipate.
Manipulative abilities	Specific	Great versatility.

Fig. 5. Fitt's list summarizes the relative advantages of man versus machine control.

A HISTORY OF COMPUTING MACHINES

Intertwined with the development of machine operated logic (studied at the beginning of Part 22) was the gradual increase in sophistication of computing machine systems.

Earliest devices were simple calculators based on mechanical concepts. They performed simple addition, subtraction and sometimes multiplication and division, doing this without the ability to store or hold values other than inputs and computed output.

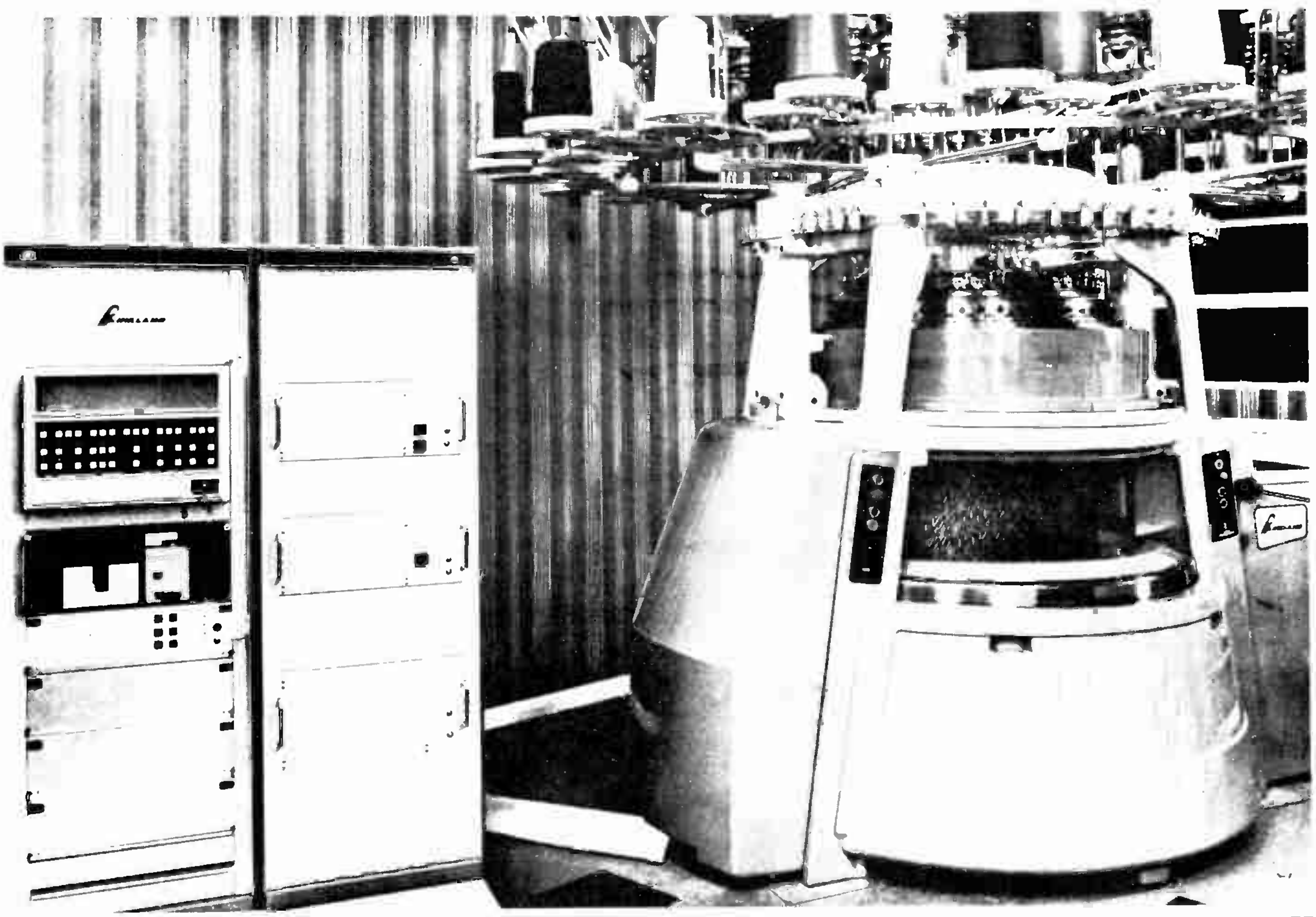
Space does not permit extensive description of this history – see the reading list for that. Figure 10 shows the style of the first calculating machine of the "modern" kind. This performed arithmetic addition and subtraction only, by mechanically manipulating interconnected counting wheels and was probably made by Pascal in 1642. In 1671 Leibniz modified the same mechanism (see Fig. 10) to obtain multiplier action, producing his own design calculator much later – in 1694. Because mechanism manufacture at that time was crude indeed – all parts were individually hand-crafted – the Leibniz machine was not reliable even though the concepts involved were sound. Improvements in mechanical manufacture had to occur before a routinely useful gear and crank calculator could be built (by de Colmar in 1820). Thus, through these and many other gradual improvements to method and manufacture, the scene was set for grander ideas.

A major advance was made by Babbage. Charles Babbage was born in Devon, England. In 1792, he became a Professor of Mathematics at Cambridge University and had a consuming passion for mechanical machines that could perform far more advanced mathematical operations than any previous apparatus. His first machine, shown in Fig. 11, was devised to solve differential equations by calculating differences. This was his "Difference Engine" of about 1812. In 1833 he conceived a second, quite different general-purpose engine — the so-called "Great Calculating Engine". In principle, it could do any mathematical operation by following instructions programmed into it by the operators. It could also make decisions, on what to do next, that were based on its just calculated results.

Babbage used punched-cards for input information (a reasonably logical choice in view of the many repetitive industrial processes using this control medium at that time), a memory (which he called "the store"), a number processing section (called the mill), a means of transferring results to and from the store, and automatic output (as cast type ready to print). It was a grand machine having ability to store 1000 fifty-digit numbers in its store. It even had overflow indication.

Fig. 6. In time-shared operation a central computer is made available to terminals. This map shows the links of bank branches to two central computer centres via concentrators.

Fig. 7. Mini-computers come in all shapes and sizes. On the left, in the console, is the H.P. 2000 that controls the pattern being knitted on the Kirkland knitting machine.



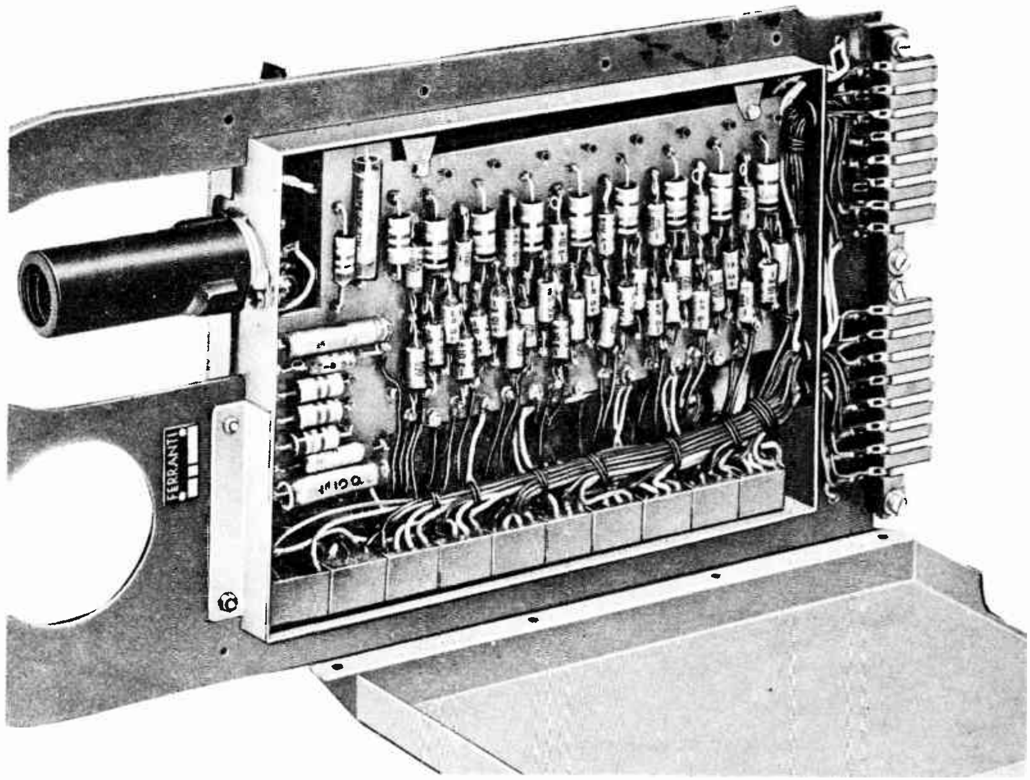


Fig. 8. Compact electronic computer systems become reality when valves were replaced by solid-state components. This single plug-in unit, from a Pegasus computer of the 50s, would today have its entire function made on a pinhead in LSI technology.

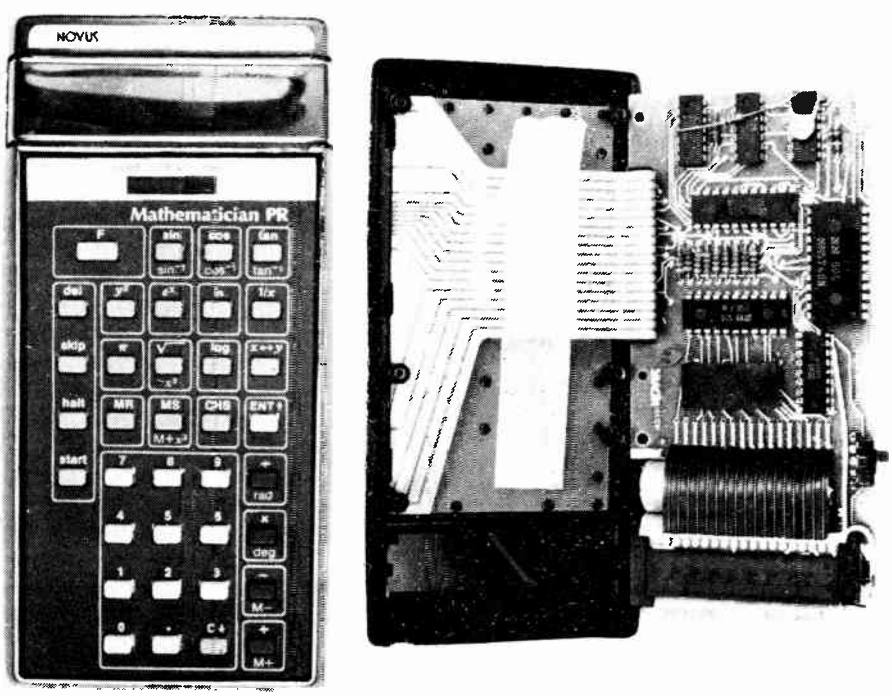


Fig. 9. Today's computers use LSI techniques in which thousands of transistors and diodes are contained within a single chip.

The intended power supply was steam. Sadly, Babbage's engines were not proven in practice in his time; those built were either not completed or proved too unreliable. Manufacturing methods were still incapable of maintaining the tolerances needed – it was a classical example of a concept waiting for the requisite technology.

However from that time on calculators rapidly became more sophisticated. Keyboard entry (Fig.12) of data, instead of the need to turn wheels, was introduced, (but the mechanism was still handcranked) number length was limited and speed was very slow (by today's standards). Around 1910 electric-motor drives were incorporated to perform the numerous mechanical rotations needed to transfer the carry-over value through all decades.

Complicated mathematical equation solving in the 19th and very early 20th century was performed on other kinds of special purpose mechanical calculating devices. The planimeter, which determines area under a curve, was devised in 1814, the mechanical ball-and-disk integrator was devised in 1876 (by Lord Kelvin's brother). With these and other basic mechanical-function solving ideas, Lord Kelvin and others put systems together that carried out specialised calculations. Kelvin produced a tidal

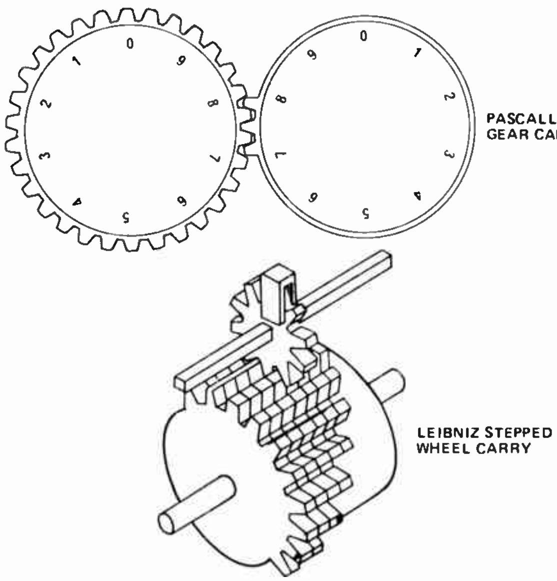


Fig. 10. Pascal's calculator of 1642 used stripped-gear toothed-wheels to produce a carry to the next decade: The Leibniz machine made use of the stepped wheel.

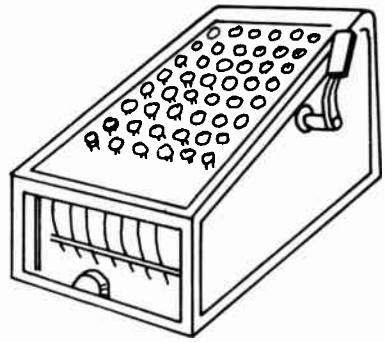


Fig. 12. Keyboard data entry was introduced around turn of the century — but speed of entry was still very slow.

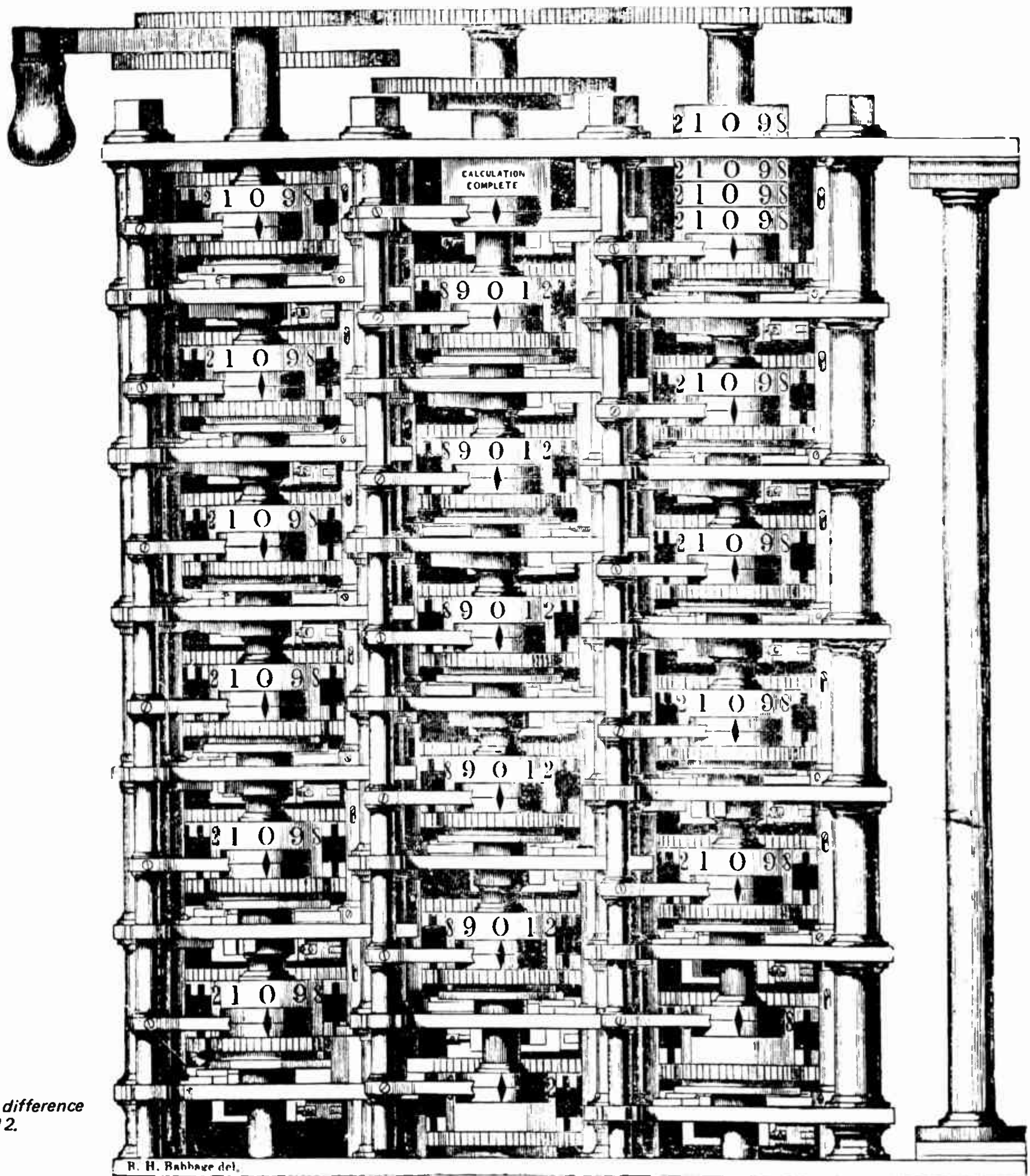


Fig. 11. Babbage's difference engine of circa 1812.

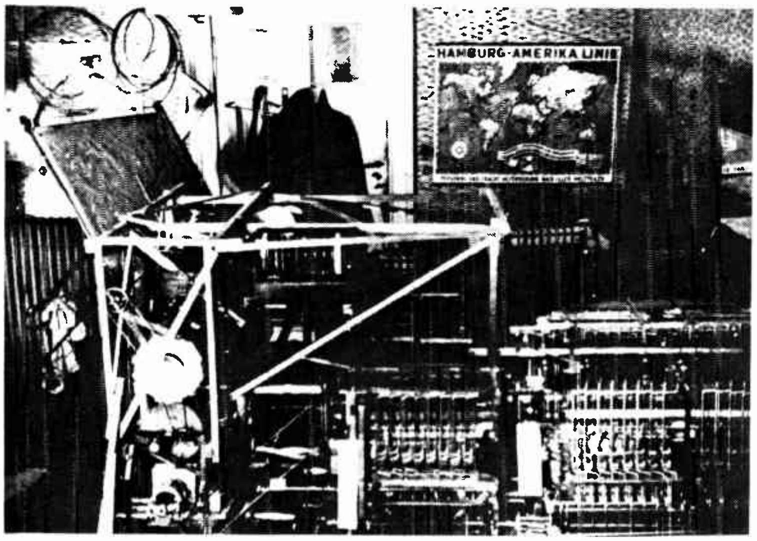


Fig. 13. This relay-switched digital calculator was built by Zuse in Germany in 1936. ((This photograph has been included because of its historical interest — unfortunately the original print is of border-line quality).

amplitude and phase predictor for sea-tide forecasting around 1874. Later in 1898 Michelson (of speed of light fame) worked with Stratton to produce a mechanical harmonic analyser.

Special-purpose mechanical calculators were still in use in the 1940s. During World War II, for instance, gun crews fed data concerning range, direction and wind strength into computers by which the correct aiming information for the gun was computed.

Today a few equipments still perform simple operations by mechanical means for in applications where electrical power is not available and the inputs not in electrical form it may be more economic to use mechanical methods.

With the advent of electronic amplification at the turn of this century electronic circuitry gradually replaced mechanical mathematical functions. This was feasible because of the superior speed of calculation, reduced manufacturing tolerances and greater reliability of electronics. The swing to electronics was intensified by the need to process an ever increasing amount of data that arises in, for example, more complex equation solving, census taking, or warfare. Hollerith devised the punched-card sorting machine to help handle the U.S. census data. This device won an 1890 competition organised by the U.S. Government.

Electric computers using the same basic system that we use today became reality around 1936 when Zuse, in Germany, built the relay-switched digital calculator (shown in Fig. 13). This machine featured automatic computing, binary arithmetic, floating decimal point and punched-tape programming. In 1937 the USA's IBM Corporation began development of a machine called the Automatic Sequence-Controlled Calculator, or, locally, just Mark I.

The trend toward total electronic working continued. ENIAC, generally recognised as the first all-electronic computer, had 18 000 valves and could operate at 500 additions per second. This was followed, after many other developments, by the first production computer — the Remington Rand UNIVAC I. It has been estimated that all computers installed in the U.S. in 1955 could do just 250 000 additions per second. Just one low-cost mini can do that today.

In 1959 a U.S. refinery installed the first process-control computer system and in 1960 a large steel corporation in U.S. was the first to use a computer to carry out inventories, handle orders and control production. Airline booking by computer began in 1964.

Integrated circuits (in the third generation machines) came into use in 1964 via the IBM 360 system and by 1970, in the U.S. alone, roughly 1 000 000 people were employed in making and using digital computers.

Single chip, fourth generation machines came to reality around 1972 with the use of LSI. Today (or at least when this was printed early in 1981) pocket scientific calculators containing over 30 000 transistors in LSI form can be purchased for less than an hour's wages. The cost of modern computers is now governed by the cost of the peripheral bits and pieces rather than the processing unit itself — the cost of the electronic components is now just a minor part of the whole.

BASIC ORGANISATION

The complete electronic data processing (EDP) system comprises *hardware* and *software.* The former pertains to the physical machinery of the computing system — that which can be seen to exist in containers and cabinets. Software is the jargon term used to cover the multitude of different programmes devised to instruct the hardware about the tasks it has to perform — these may come in punched card, punched tape, magnetic tape and disks or in written format.

The hardware of electronic data processing systems comprises the several basic functional blocks depicted diagrammatically in Fig. 14. Peripherals enable the electronic circuitry of computation to communicate with external information flows via the input and output units. The heart of the system is the central processing unit (CPU) comprising a very fast-access store (also called the high-speed memory) of digital numbers; a unit that performs simple arithmetic operations at high-speed (called the arithmetic unit), and a control unit that coordinates all units by stepping (clocking) the system on bit-by-bit by means of a clock pulse source.

A CPU can serve many different functions and all CPUs are not identical by any means. Typical tasks are to control the peripherals and the input/output information flow, perform the arithmetic in scientific work or compare data in data-processing uses where the logical capability is exploited more directly.

Data is shunted back and forth between units on the bus lines using parallel, and serial forms of binary number transfer. (A number of binary bits, when combined into a number, are described as words). Different manufacturers use different word lengths — 24 in ICL 1900, 32 in IBM 360; 36, 48 and 60 are also used. The term 'byte' will also be met and this is the designation for a short segment of the full word. For example an 8 bit segment of a 24 bit word.

Words are held in the store when not being operated upon. As well as being a binary number that is directly equatable to a decimal number, words can also represent instructions for the control unit to use, a piece of a number, a sequence of letters or any

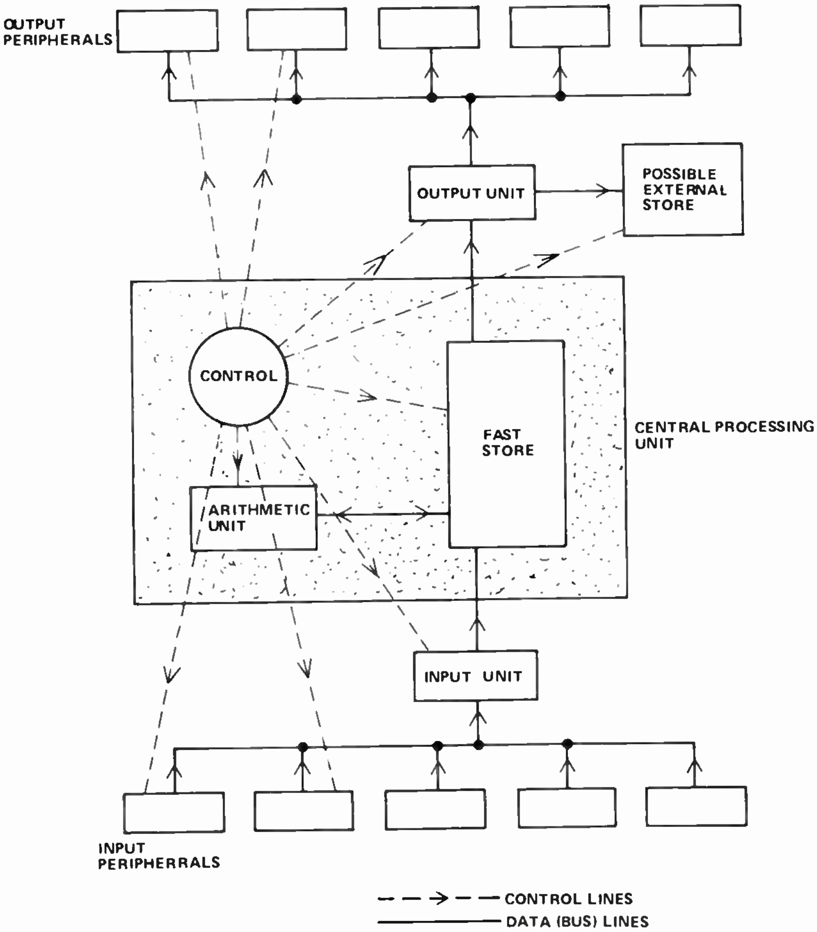


Fig. 14. Basic functional blocks of electronic data processing system.

other symbols (eg graphics) as desired.

Words usually include one extra bit called the parity bit. The parity bit is used as a continuous check that the words transferred between locations have arrived as sent without any binary position of a single word having its state altered on the way.

In order to make use of this versatile arrangement, a CPU must have a set of instructions to tell it when and where the data has been placed for optimum use of time. The programme, that is the software, performs this task at speeds much greater than human operator or the input units could. A programme is loaded into the CPU at the speed allowable by the input mechanism, during which time the computer is often employed on other tasks. Once loaded the system is started on the problem, and then runs at the maximum speed of which it is capable. To give some idea of speeds involved, a CPU internal operation will take around 1μs or less whereas a fast peripheral barely gets down to 100 000 μs per operation. The design of EDP systems is very much one of careful systems organisation to avoid wasted operating time.

We say software programmes operate with various languages. The most basic and original language is *machine language* wherein the programmer must specify exactly which bit must go to exactly what store location and so on. This is seldom used today at operator level. With time, languages have come closer to normal spoken language; this being achieved by building more and more automatic programming functions into the CPU. The closer the language to everyday expression the higher the level of the computer language. Many aids have been established to ease the skill needed by the programmer in compiling a workable programme. We are however a long way from programming by merely talking to the machine. Computers must still have their instructions in a strict written format. In compiling the exacting programme sheets the user must first establish what he wishes to do mathematically, and detail the steps required on a flow chart. Figure 15 shows a typical flow-chart for a data processing problem.

It is not possible to state, in general, how long a computer may take to provide a solution: tasks can take many hours to mere fractions of seconds, depending on the size of problem and computer power. Off-line computations may not necessarily have to be performed at high speed (except when other jobs await their turn) but in process-operations it will often be vital that a calculation is made in sufficient time to gain stable control. Remembering that all calculation required must be reduced to the basic four functions of add, subtract, divide and multiply it does not take much of a calculation to consume many milliseconds, especially when the decimal accuracy needed is high. A particularly fast computer may be essential to obtain millisecond time-constant control in computer-control work.

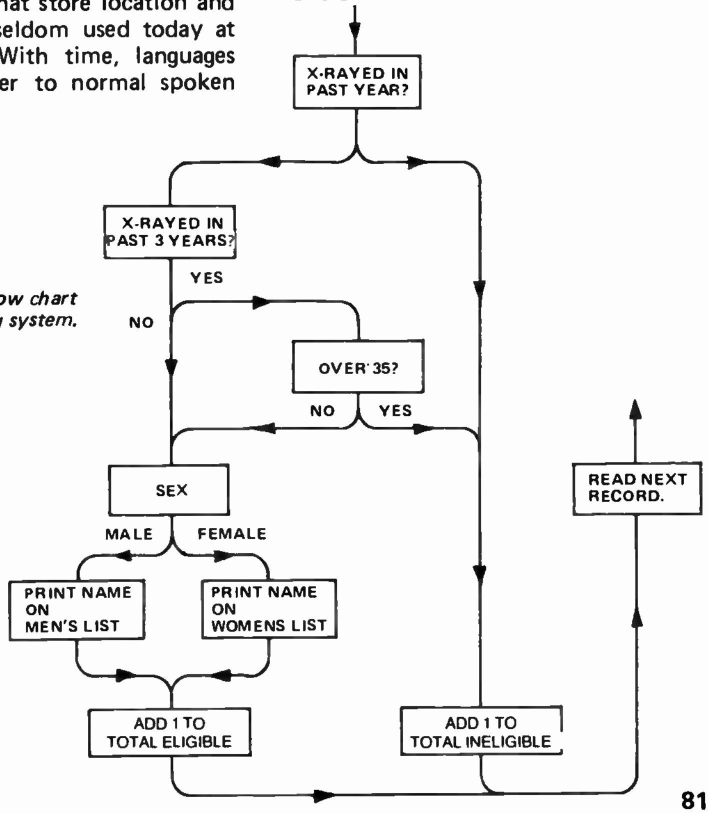


Fig. 15. Typical flow chart for data processing system.

ELECTRONICS–it's easy!

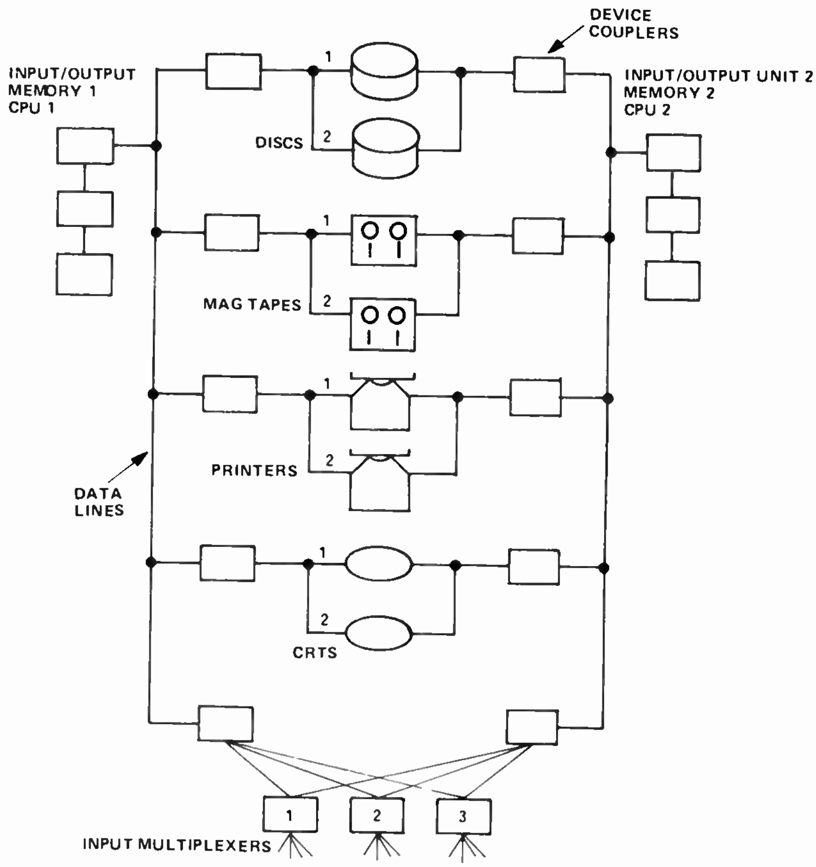


Fig. 16. Redundant circuits are incorporated so that one unit will continue to perform a vital task even if though its complementary unit fails.

ADVANCED ORGANISATION

Even the best designed digital circuits occasionally go wrong or pick up stray noise thus causing errors. A single parity check greatly enhances the chances of detecting errors, but with the development of faster machines that conduct vastly greater numbers of operations in a given time, the reliability of the systems to perform correctly without error comes into question. When reliability is a vital consideration, as for instance it is when designing computers for manned space shots, the equipment may be duplicated or triplicated – this extra equipment is called 'redundant'. An obvious way to incorporate redundancy is shown in Fig. 16 where all units are simply doubled-up and connected so that one can perform the task if the other fails. There are preferred ways to connect extra equipment, the general rule being that as many cross-connections are made as possible as demonstrated by the two

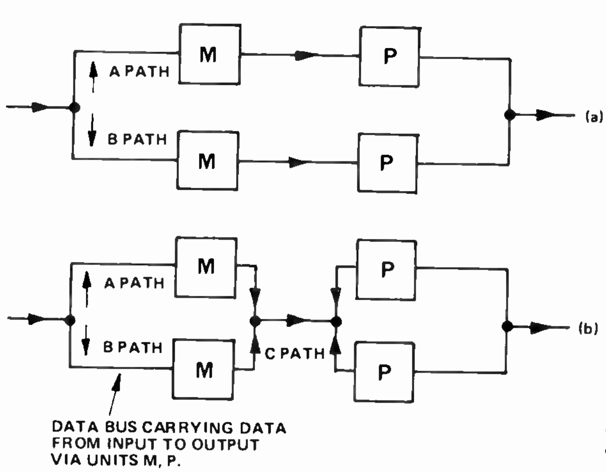


Fig. 17a. The data path may be via A or B. Fig. 17b, if an additional path 'C' is added, reliability is clearly improved.

systems shown in Fig. 17. In Fig. 17a, the data path can be via A or B which is clearly more reliable than via just A if both A and B have equal reliabilities. If an additional path C is added, as shown in Fig. 17b, we have improved the chance of data being processed by an M and P unit sequence.

The reliability of systems is measured in terms of the mean-time between failures (MTBF) and the mean-time to repair (MTTR). As a guide only, Fig. 18 shows typical values for the various kinds of units involved. (The following part gives more detail of peripherals mentioned in the figure).

Taking the idea of interconnected redundancy to the limit we have a system schematic like that given in Fig. 19. At each nodal point any one of a multiplicity of units can be brought to bear on the nodal task. If a single unit fails, the effect is not a total shutdown of the task but a slight degradation in speed and capability of the whole system. This has been called the *fail-soft* design and such systems exhibit the graceful degradation that occurs in physiological brains. The concept of total interconnection is loosely analogous with the way in which physiological brain cells are connected.

In the next part of this series we look at the peripherals used, various kinds of stores, microprocessors and the latest manufacturing techniques.

REFERENCES

Two books, already referred to in Part 22, are relevant, these are:

"A Computer Perspective", C. and R. Eames, Harvard University Press, Massachusetts, 1973. (This is a definitive work on the development of data processing equipment from 1800 to 1940).

"Electronic Computers -- Made Simple", H. Jacobowitz and L. Basford, W.H. Allen, London, 1967. (Although out of date with respect to certain aspects of hardware this provides a valuable basis for technical understanding of both analogue and digital computers. It also explains the arithmetrical operations).

"Introducing Computers", M. Laver, HMSO, London, 1973. (A version compiled for users with a little technical knowledge. It discusses programming procedures).

"Computers at work" J.O.E. Clark, Bantam Books, London 1973 (A most useful book on where computers are used).

"Electronic Computers", S.H. Hollingdale and G.C. Tootil, Penguin Book A524, Harmondsworth, 1965. (A fine layman's summary of analogue and digital computers including a lengthy chapter on what sort of jobs computers do).

Computer programming is covered in many texts and booklets. One example is:

"Elements of Computer Programming", K.P. Swallow and W.T. Price; Holt, Rinehart and Wilson, New York, 1965.

When the need to learn how to programme a computer arises it is best to seek specialised advice about reading material pertaining to the computer to be used. There are numerous models available each having its own peculiarities and each requires considerable operator training-time. Fortran, by IBM, and its dialects are commonly used programmes; an inexpensive programming primer is:

"A First Course in Fortran", E.J. Burr, Department of Continuing Education, University of New England, N.S.W. 1974 (Third edition).

ALGOL language began to emerge in 1958 as a step toward a universal computer language for scientific working. COBOL is the commercial counterpart. Relevant books are:

"Basic ALGOL", W.R. Broderick and J.P. Barker, IPC Electrical and Electronic Press, 1970.

"A Guide to COBOL Programming", D. McCracken, Wiley, New York, 1970. ●

Fig.19. The fail-safe design — see main text.

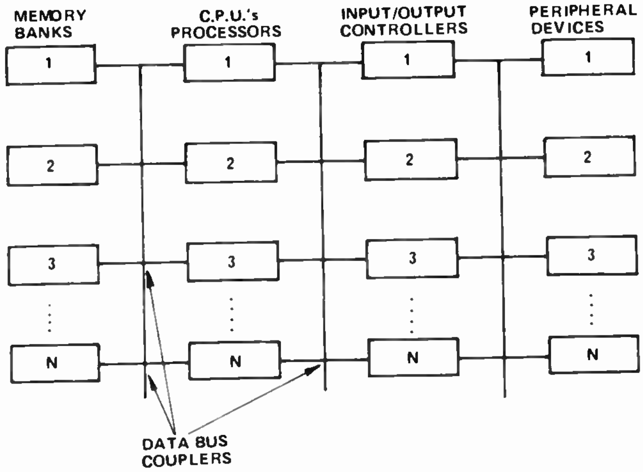


Fig.18. Typical mean time between failures — and typical mean time to repair.

	CPU	Memory	Input/output unit	Device coupler	Disk store	Mag-tape store	Line printer	C.R.T. display	Multiplexer
M.T.B.F. hrs x 10^3	6.0	15.0	6.0	12.0	15.0	3.0	1.5	4.0	10.0
M.T.T.R. hrs	3.0	3.0	3.0	2.0	15.0	2.0	1.5	2.0	2.0

31 Computer peripherals, stores and microprocessors

ALL COMPUTING SYSTEMS HAVE a Central Processing Unit, (discussed previously) and a number of pieces of external equipment associated with them. Such additional units, known as peripherals are necessary to handle the flow of information between the outside world and the Central Processing Unit (CPU).

The range of peripherals available today is extensive. Basically the design aims are to provide interfaces between the human or automatic plant user and the computing system which are the easiest to use, the cheapest to implement and which have the means to transfer data as fast as is desired.

At present — though this will undoubtedly change in the future — we are unable to communicate with the computer by the same means that we communicate with each other — that is by direct speech and vision. Peripherals, are by necessity of our technological and economic limitations still very much compromises to the ideal, except in applications where the computer interfaces to hardware plant, such as in process control, when interface problems are easier to solve as such systems communicate by the same signal formats.

Card and Tape Punches and Readers — In order to make good use of the high speed of electronic computing circuits, the input and output functions should ideally be capable of transferring the data at a comparable speed. Rarely has this ideal been realised. The throughput rate of peripherals has been speeded up enormously since the first EDP system but, similarly, the rate of computation has been increased.

Because of this shortcoming, data (in human operator use) is first prepared by hand onto a medium that can feed into the EDP system at rates far exceeding the operator's ability. It is then stored in the machine ready for access when the CPU needs it.

The earliest form of input/output medium used punched holes made in a pile of paper cards or a continuous tape. We inherited these from a 17th century weaving machine via the Hollerith census sorter. Figure 1 shows the commonly used Hollerith coded punched card. The holes are punched out in a code that represents the alphanumeric symbols shown above each row. Figure 2 is a section of punched tape: these are available with 5,6,7 and 8 hole positions across the tape width. (The smaller hole is for the timing drive sprocket). Tape readers are built to read code from a specific width tape: that is, a 5-hole tape could not be used on an 8-hole system. Tapes and cards which are to be used extensively can be made in more durable materials such as oiled paper, Mylar and aluminium-Mylar.

The holes in cards are produced by mechanical punches. These comprise a punching head by which the appropriate holes are made for each character in response to a typewriter keyboard-input. Keyboard layouts are based on the familiar office typewriter. Extra keys are added for computer applications to enable a greater range of control by the operator. Such additions vary widely.

Tape can be punched automatically whilst the teleprinter type of terminal, such as shown in Fig.3, is used as a typewriter. Where the tape is generated as part of an automatic process — as in a data logger, a smaller punch unit is used which incorporates punch drivers activated by control signals — no keyboard is needed. Such a unit is illustrated in Fig.4.

Card and tape readers consist of a transport mechanism that passes the medium across reading heads. Recognition of a code represented by holes is accomplished by mechanical fingers making direct electrical contact (in the slower readers) or by solid-state optical sensing using LED lamps and photo-diode arrays set to sense the passage of light through a hole position. Some method of synchronising the code position with the data values is essential.

Cards can be punched by an operator at rates between 250-500 per hour. They are often checked on a verifier machine that determines if the card is punched in the same way as the check operator keys the code a second time. They can, by contrast, be machine read or sorted, at 200-1000 cards per minute depending upon the complexity of the task.

Tape punching is confined to similarly slow rates of production at the operator stage of preparation. When the punch is

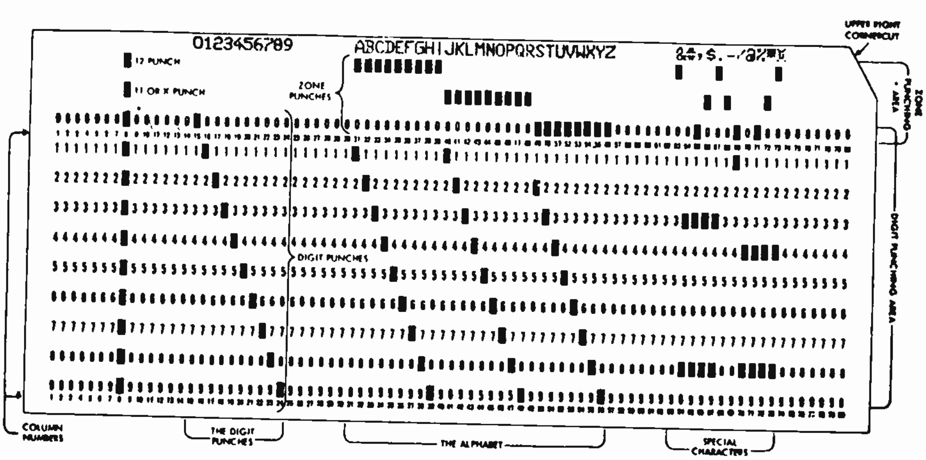


Fig. 1. Standard Hollerith Code used for punched cards.

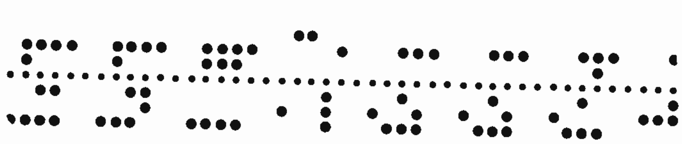


Fig. 2. Section of 8-hole punched tape. Two rows on the wider side of the central sprocket holes are not used in this data.

Fig. 3. Keyboard teletype terminals provide hard copy and perforate a paper tape (seen lower left) at the same time. The same facility usually can also print hard copy from a ready made tape.

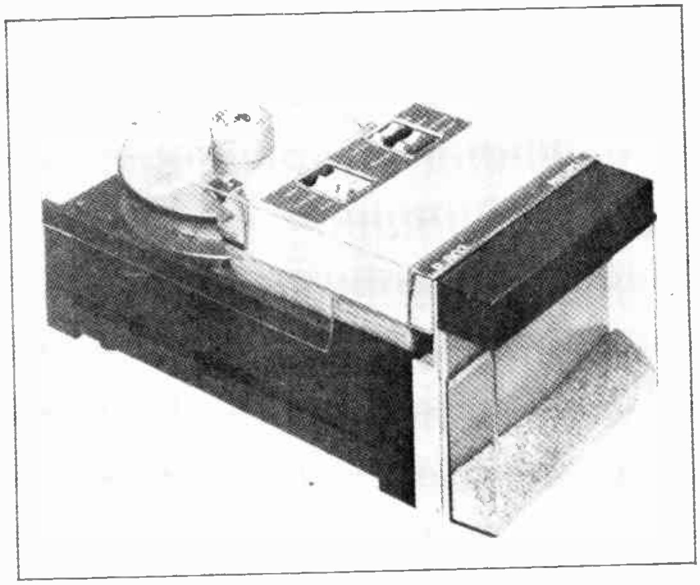


Fig. 4. This Philips P803 unit punches tape at 75 c.p.s.

machine operated, punching rates can rise to 150 characters per second. The speed at which punched tapes can be read varies from very slow, using mechanical sensing up to 600 characters per second or more with high-performance optical equipment.

A considerable amount of electronic logic and drive circuitry is needed to operate a punch unit. Figure 5 is the block diagram of a reader using brushes to sense the presence of holes. Input commands to the punch would emanate from the control unit of the EDP system.

Magnetic Tape Input/Output Units – Cards and paper tape store information about commands to the EDP system (the programme) and hold the numerical data to be manipulated. They are, therefore, a form of permanent data storage. They suffice (in the form described above) as a data store when the data quantity is not great. A recent trend, which has speeded up data transfer and reduced the bulk needed to store the programme and data, makes use of magnetic tape in cassette form.

The compact unit shown in Fig.6 can transfer data at 6000 bits per second at a density of 30 bits per millimetre of tape. (Total capacity on a cassette – five million bits). These can also be used as additional memory in the system.

Printers – Teletype units are able to provide hard copy printout but due to the slow printout resulting from letter by letter operation they are not used as the main alpha-numeric output of an extensive EDP system. They can printout at only 10 characters per second or so.

The line printer was evolved to speed up this form of output. It prints all the characters of a complete line simultaneously. Line lengths are typically 132 characters and the faster models can print lines at rates exceeding 1000 lines per minute.

Printing mechanism vary considerably, ranging from development of the fundamental typewriter method, to devices that print each character from a 5 x 7 matrix of dots. Line printers were originally bulky units. Today desktop, typewriter size units, are in common use (Fig.7).

Printers can be programmed via the EDP system to provide any format required – periodic reports, invoices, records, data lists, software record. A crude form of graphical display can also be produced using the position in a line

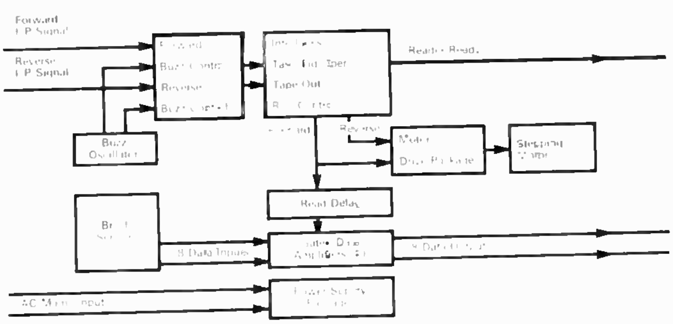


Fig. 5. Block diagram of early model Data Dynamics low-speed tape reader (30 c.p.s.).

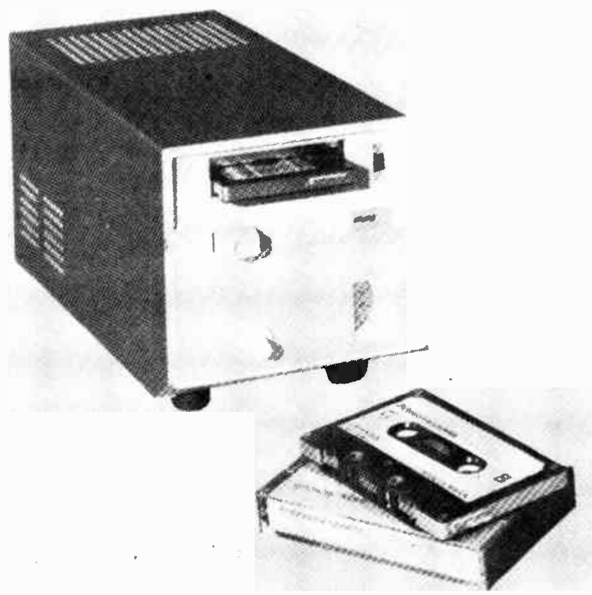


Fig. 6. Cassette form of magnetic tape is finding greater application as a standard EDP and computing caculator peripheral.

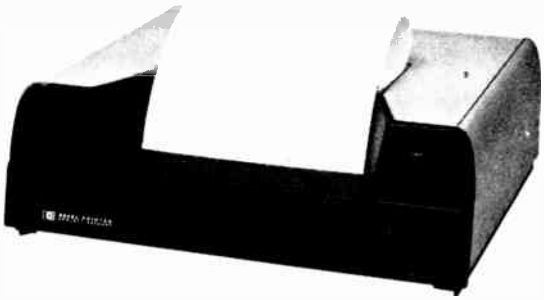


Fig. 7. This Hewlett Packard 2607A desk-top line printer provides hard copy output (with 6 copies) at 200 lines per minute.

as one ordinate and the lines as the other.

When computers are used for automatic pagination the printer can be one that produced print-type direct.

Graphic Display – Plotters – Many computational tasks ideally require a graphical display of output information not a long list of numbers. Plotters may be of x-y type or y-t type.

The x-y type of plotter is arranged so that the graph paper is held stationary and the pen is capable of being driven both vertically (y axis) and horizontally (x axis).

The y-t plotter has a roll of graph paper which is driven at a constant (and usually adjustable– speed; the pen can be driven in one axis only (y axis). Hence the y-t plotter basically plots a single variable against time. Plotters made particularly for computer operation will be provided with the interface facility that enables direct connection to the EDP system. (Normal plotters require an extensive amount of extra equipment to make them compatible).

Computer controlled plotting of x-y format has the ability to be scaled on demand and to generate alpha-numeric legends on the plot. It is an easy matter to replicate the plot – the programme is run again.

Plotters may be of the analogue drive kind (a later part discusses plotters in detail) but due to the nature of digital processing the result may still have a quantized appearance if the resolution is not sufficiently small. Alternatively the axes may be driven with stepping motors – such machines are called incremental plotters.

Flat-bed style of x-y plotters are available which can handle paper of all sizes – from a few centimetres square to size of a wall. A medium-size computer controlled flat-bed plotter is shown in Fig.8.

Line drawing rates are limited by inherent electro-mechanical response to around 0.4 m/s in small plotters. The very large machines, when under tight control, are usually capable of around 0.1 m/s translation rates when working to precisions of 25 μm. A desk top x-y plotter is shown in Fig.9.

Some y-t plotters incorporate bi-directional drive for the t axis (the paper drive) enabling very long lengths of paper to be driven back and forth along the roll in order to produce an x-y form of plot from a y-t format machine.

Graphic Display – Visual Monitors – Many applications require rapid call-up of data that is presented in a way that can be easily read by the operator. It may be quite unimportant to receive it as hardcopy. The cathode ray tube (television) type of display was an obvious choice. Such displays are known as visual display units, VDU for short.

Originally, visual display units were very limited because of the need for a

Fig. 8. Series 500 automatic drafting system by Gerber Scientific.

Fig. 9. The 9862A H-P plotter has a resolution of 0.01% full scale within the 25 x 38 cm area. It will plot vectors at 30 cm per second.

considerable amount of storage with which to generate written and graphical display forms. However solid-state mass data storage is now relatively inexpensive and VDUs in one form or another are now standard peripherals.

The simplest use of VDUs is to display alpha-numeric information — a section of the software programme, a readout of process plant variables, airline arrivals and departures. This is achieved using digital control and data storage to cause the beam of the CRT to deflect, blanking appropriately, to form the appearance of a static written page.

When the operator becomes involved with the data on the screen and is given the ability to manipulate it toward a desired task the terminal is said to be an interactive graphic terminal. An early example of this is given in Fig.10 which depicts a system whereby air traffic controllers are trained using display terminals.

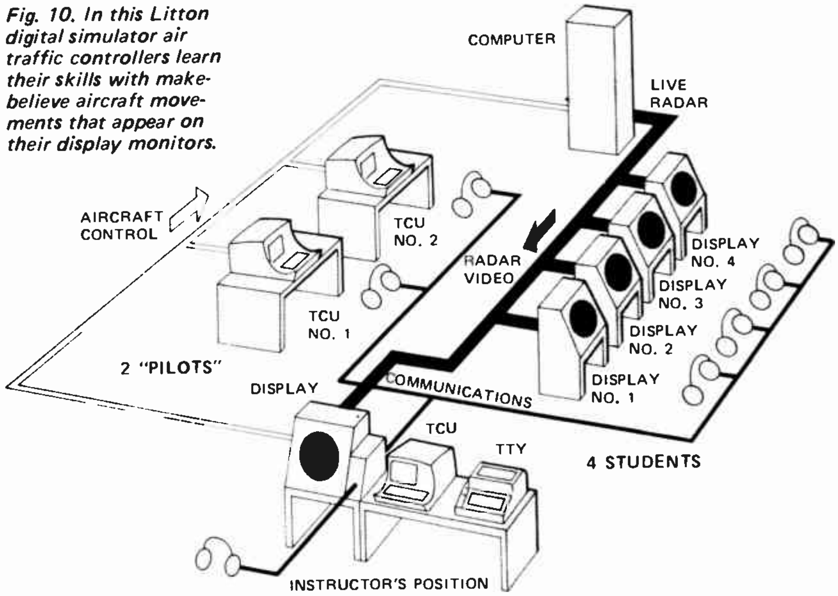


Fig. 10. In this Litton digital simulator air traffic controllers learn their skills with make-believe aircraft movements that appear on their display monitors.

Once it had been realised how the VDU could be used to produce line drawings designers sought ways to 'draw' on the screen. The result was the 'light-pen'. The operator holds a special stylus on the screen of the CRT. Closed loop controls cause the spot to lock onto movements of the stylus. If the trace path is to be retained, the x, y and intensity coordinates values are fed into the digital memory. Once a line is drawn it can be retained and regenerated in this way. Other operations enable the operator to automatically erase sections of line, straighten lines and smooth curves by computer processing. The complete drawing can then be permanently recorded as hard copy on a plotter or as a data set. Interactive methods have saved an enormous amount of time in tasks such as deciding the extremes of a motor-car wheel movement during the many combinations of springing and steering positions within the wheel arch.

Today's graphic terminals are extremely versatile. Completely self-contained units which incorporate a built-in processor are in common use. A recent release is shown in Fig.11.

Improvements in the storage-tubes used to hold the displays of a CRT system have been coupled with the power of modern computing to provide display terminals that have half-tone photographic quality presentation. Figure 12 shows the quality (after our recopying) obtainable. The images shown are entirely reconstructed on the VDU from digital, not analogue data. Colour displays are also coming into use adding yet more dimensions to the interaction available to the operator.

A recent project of the Australian National University gives some idea of the use of the interactive VDU. In the Department of Engineering Physics a team of research workers have developed a colour display terminal that can call-up the data recorded by the ERTS satellite. The computing system has in its memory file copies of the original ERTS data. Using the graphic terminal the operator can select which form of photograph — IR, false colour, etc., to study. He can then rapidly zoom into a particular area using a joystick control expanding the spatial scale as the search becomes concentrated. Other control includes enabling the colours to be digitized into level zones and to be complimented..

Instrumentation Interfaces — When the digital computer has to manipulate measurement and control data from analogue processes, the system must be provided with the appropriate A to D and D to A converters, and the multiplexing arrangement which forms the data logger These interface peripherals were mentioned in the 29th part of this series.

MODEMS and other links — When computer data has to be transmitted over considerable distances it becomes expedient to use telephone lines or

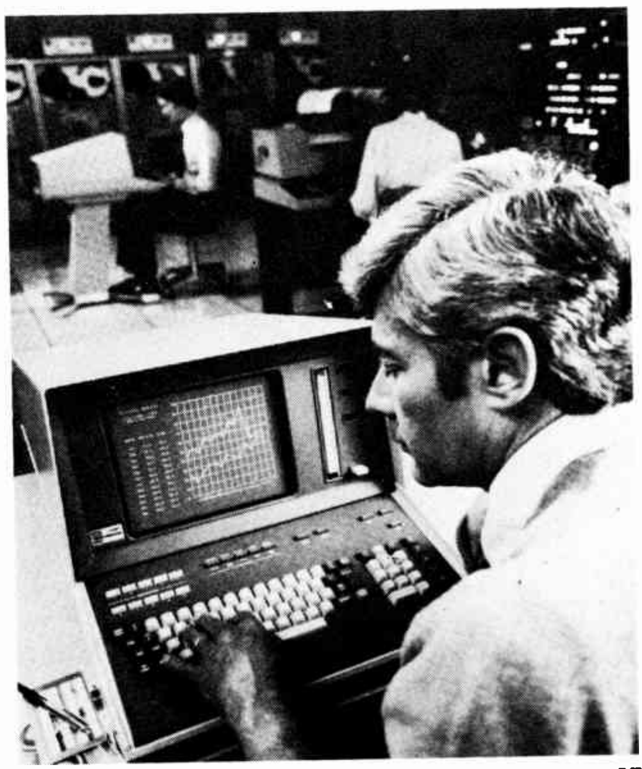


Fig. 11. Interactive graphic units often now incorporate their own processing and memory to form an off-line self-contained unit — 4051 Tektronix BASIC graphic computing system.

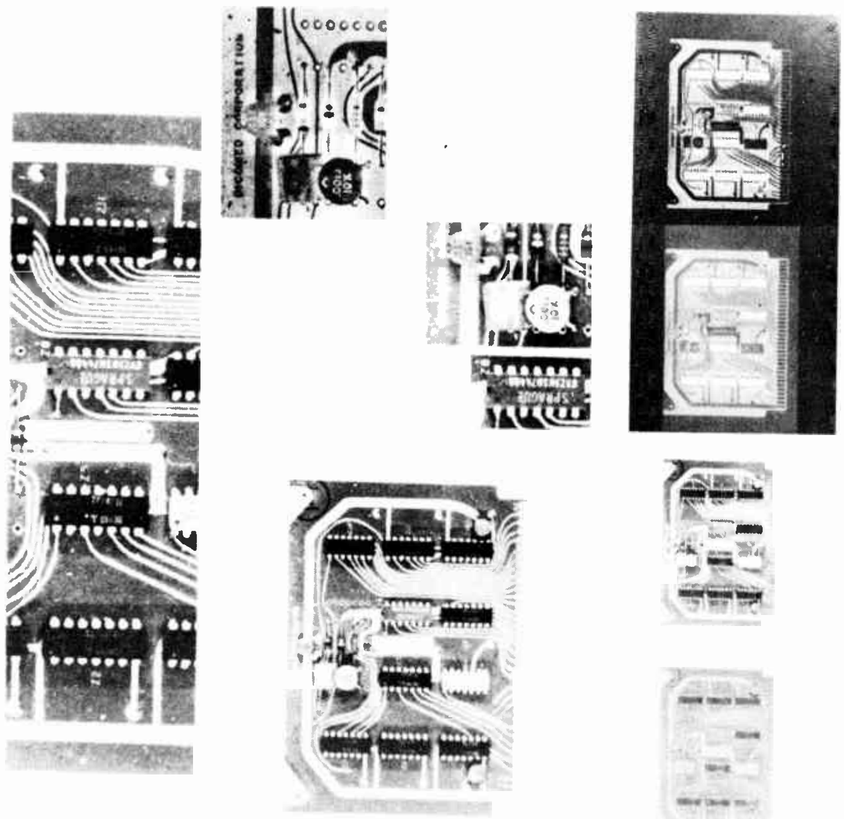
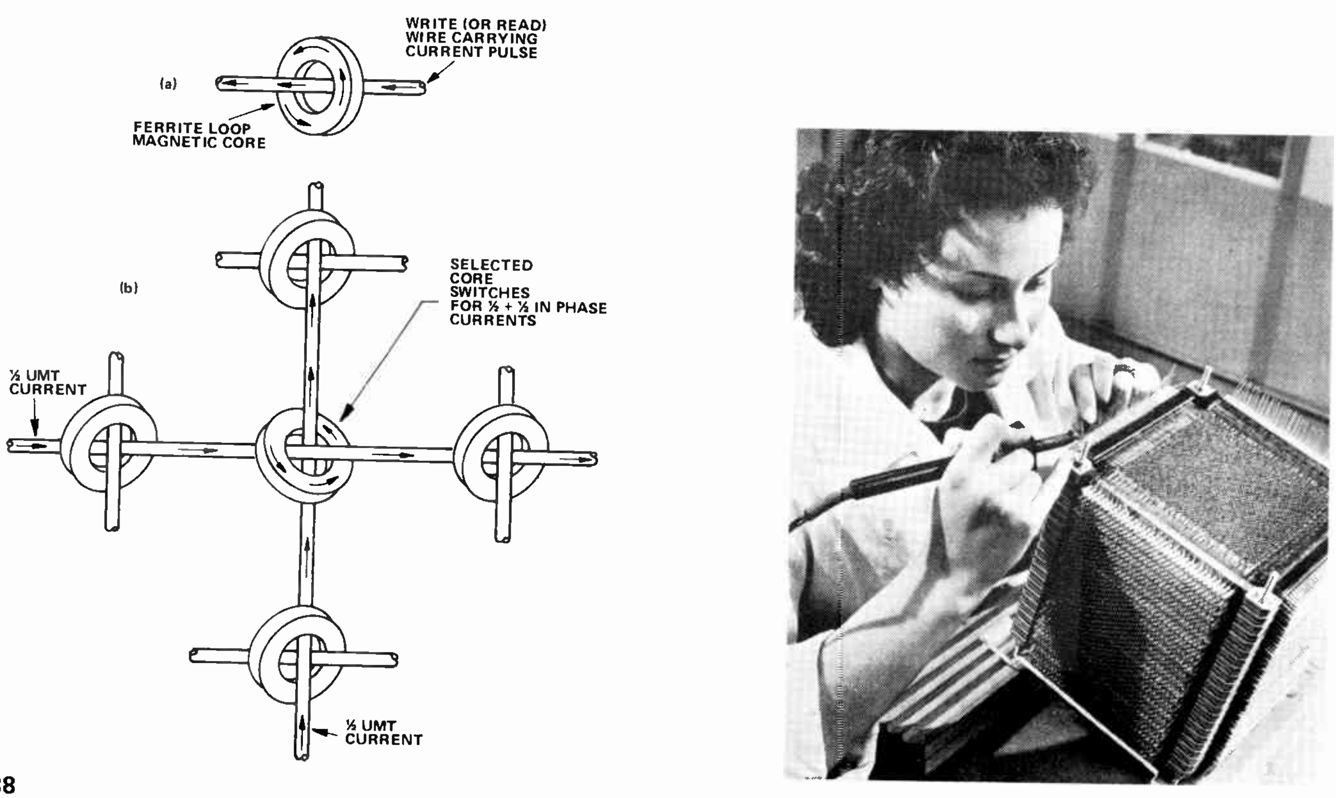


Fig. 12. This multiple image presentation is photographed from the screen of DICOMED digital image display unit.

Fig. 13. (a) When a large enough write current passes in one direction through the ferrite core the core becomes magnetised in one polarity. It thus records a bit.
(b) A second wire is added to act as an inhibitor or enhance line.
(c) Finishing touches being added to a Philips 3-D core store. (20 planes of 64 x 64 cores, one X wire, one Y wire, read and inhibit wires).



microwave links. Units interfacing computers over telephone lines have become known as MODEMS (a word built by combining Modulator and Demodulator). Links are dealt with in Part 32 – in which signals and transmission are covered.

Miscellaneous Peripherals – New methods for communicating with the power of an EDP system continue to be devised in an endeavour to overcome the interface difficulty humans have with electronic machines. We are still a long way from the stage where we need only casually to talk to the machine. Steps are, however, in progress toward this aim with research into spoken word and written word recognition. Neural research into brain waves may one day be coupled with electronic hardware to provide direct thought-links.

Special cases demand special solutions. Work at Warwick University in England has resulted in computer controlled production of braille maps for the blind. Automatic mapping and language translation are other areas where positive progress is being made into very complex human communication processes.

STORAGE

Inside a CPU and external to it will be found a memory of some kind. This is used to store the vast quantities of coded data needed to perform the various tasks.

Memory within the CPU is characterised by the need for high speed access to any data bit needed. The requirement on capacity is less stringent. Memory external to the CPU will, by the necessity of machine organisation, be a little slower to access but it will usually need much greater storage capacity.

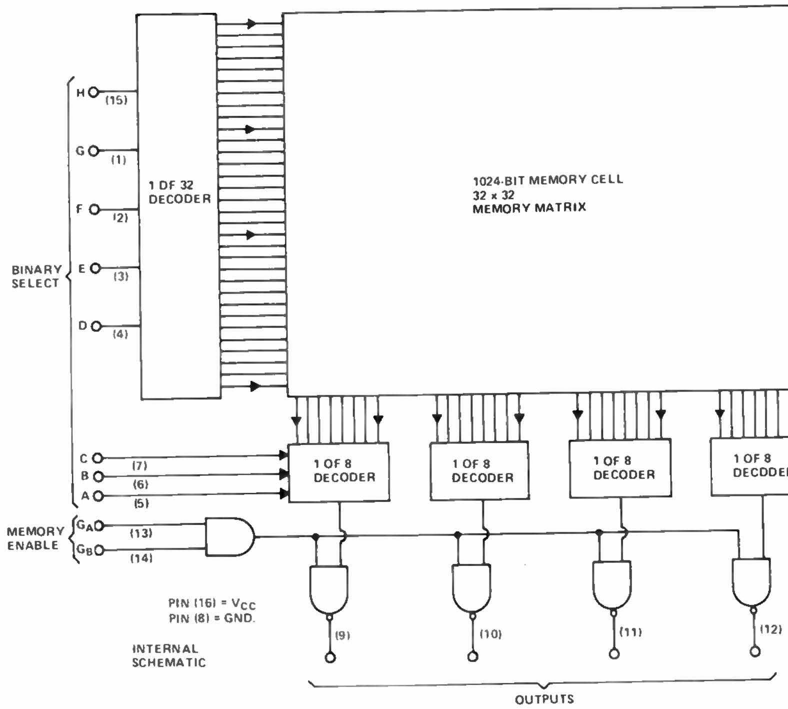
CPU Memory – Core – storage is needed in the CPU to hold important programme instructions and to act as a temporary home for data generated in the course of a manipulation.

There are many options open to the designer but the storage method that has emerged as the optimum for CPU storage is magnetic core storage – known simply as the core store. (This situation will, however, soon change, the preference going to solid-state methods). Magnetic core storage makes use of the fact that magnetically hard materials, such as ferrite, will swing remanent magnetism polarity from one state to the other with the passage of a quite widely toleranced current through a wire passed through the core – see Fig.13a. To make a practical core store it is necessary that any chosen core can be switched on demand. If a second wire is passed through the loop this can be used to prevent or enhance the magnetic switching action by the passage of the current.

A core store comprises a plane of ferrites arranged in a grid as shown in Fig. 13b. Two half-current units appearing in the same direction in a core will switch that core but no other. Thus two lines will select a unique core in the plane as the place to store or readout one bit.

To read out the values it is necessary to interrogate the selected core using input signals in the write wires that will, if switching takes place, induce currents in an additional readout wire. As this process can destroy the data on the core a test means may be provided to rewrite it again ready for reuse. Figure 13c shows a stacked core-plane. Ferrite cores are typically 0.1 mm overall. Planes are either stacked one on the other or mounted flat on a printed circuit board to provide a memory unit. The capacity of core storage varies from thousands to millions of bits. Core-store is more usually quoted in word capacity, words being of 32-60 bit length. The terminology is to refer to capacity as, for example, 32 K of 16 bit words. Note the use of the upper case 'K' – this denotes the multiplier 1024. Don't confuse this with 'k' which denotes 1000. Core storage can be cycled with 100 ns (typically) with some systems taking only 10 ns. The disadvantages of core are relatively high cost resulting from the labour intensive production method and the comparatively large space needed.

Delay Lines – Another reasonably fast storage system makes use of the delay-line concept. It is the property of materials, such as mercury, to pass only waves of acoustic energy at a given rate



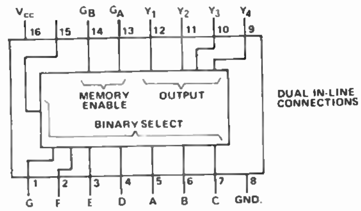


Fig. 14. Internal schematic and dual-in-line connections of National Semiconductor 1024 bit, read only memory. 8 lines are supplied to operate readout from the memory using 8 bit code.

of propagation. Early computers used mercury delay lines in which the acoustic equivalent of a binary word was sent down a tube of mercury to emerge at a later time at the other end. Whilst in transit the word was in storage. The method (if used at all in a computer today) would now be implemented using solid wires or clocked-on registers. It has the severe shortcoming of low storage capacity.

Solid-state – Although core storage still forms part of many computer installations the current trend is clearly toward the use of a solid-state circuitry which stores bits in register style flip-flop systems. Read only memories (ROM), content addressable memories (CAM), random access memories (RAM), and Programmable ROM devices (PROM) are available as IC chips with typical arrays downward from 512 eight bit words – that is 4096 bits on a single IC chip. Figure 14 shows just one of a huge range of alternatives – 1024-bit read-only memory. Memories such as this exhibit a typical delay from address to output of 36 ns. Chips such as these are also available ready mounted as memory cards with as much as 65 536, 16-bit word capacity.

Peripheral Memory – The storage media listed above gives high-speed rapid access but all are expensive. Many others and cheaper forms of storage can be used if short access time requirements are relaxed.

Magnetic Tape – This is basically the same as reel-to-reel domestic tape recording, magnetic tape storage used in computing however records digital rather than analogue data on the magnetic coating of the tape. Reels are generally 10.5 inch in diameter with multiple track use. They are run at much greater speeds than domestic units. They can store around 30 bits per millimetre and maybe run as fast as 25 metres/second. Speeds used are not standardised to any degree. Each track on the tape can only be accessed serially: to obtain a specific data word may involve the whole tape being run through with subsequently long access time. Figure 15 shows a typical reel-to-reel unit.

Magnetic Disks – These are thin disks coated with magnetic recording material. Their advantage is that they can be accessed at any point on the surface by moving the read in read out head to the appropriate part of the disk, as the disk rotates, (at speeds of 3000 r.p.m.). In an alternative procedure the reading is done by a fixed head for each

Fig. 15. The Hewlett-Packard 7970 magnetic tape unit designed as a peripheral to EDP systems.

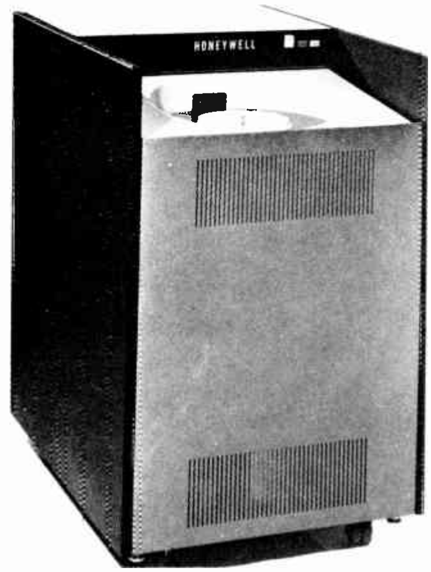


Fig. 16. Honeywell model 4720 moving-head disk storage device.

track. Each track may store 36 000 bits. The moving head disk storage unit shown in Fig.16 can store up to 7.5 million words.

Even greater storage is obtained by permanently stacking as many as 72 disks on top of each other on a common drive spindle. Each surface has its own head giving access to any part of any surface. Such a unit could store 600 million words. Access time is, however, limited by mechanical response times – typically 100-300 ms. Small interchangeable disk stacks are also used. These are known as disk packs. Floppy disks are a variation of the disk memory.

Magnetic Drums – Where better access times than disks are needed, but not at the cost of magnetic core, the magnetic drum may be suitable. A large drum (0.3-0.6 m in diameter) coated with magnetic material rotates continuously at high speed. Reading heads are stacked up the drum. Access time with these is as low as 5 ms. Storage is upward of 2000 million characters.

Other magnetic arrangements include short strips of tape that are individually selected to be drawn through a reading head, and magnetic cards which are held in magazines ready for automatic sorting in a special console. Card systems are not as slow as might be thought – any one of, say, 500 million characters can be accessed in 100 ms by a suitable design arrangement.

MICROPROCESSORS

We saw in the previous part that computers are based upon the availability of a CPU, stores, input/output units and other peripherals. Integrated circuit manufacturing methods become economical only when very large volume sales result and it was to the computing systems market that the IC makers looked around 1970. The main problem, however, was the need to devise a basic general-purpose integrated-circuit that would satisfy a large enough group of users.

At first the trend was to manufacture special-purpose computing systems that were hardwired (connections made permanently) to cause the system to perform a stated computing function – such as a pocket calculator for commercial or scientific computation.

The trend then moved toward another philosophy – the microprocessor. These single card integrated-circuit systems (one is illustrated in Fig.17) possess the ability to be programmed to perform the task needed by the customer. Although the overall system is usually more complex than hardwired specials, the much greater increase in demand has reduced the price to quite unbelievable levels – a few hundred dollars buys a complete basic micro-processor system

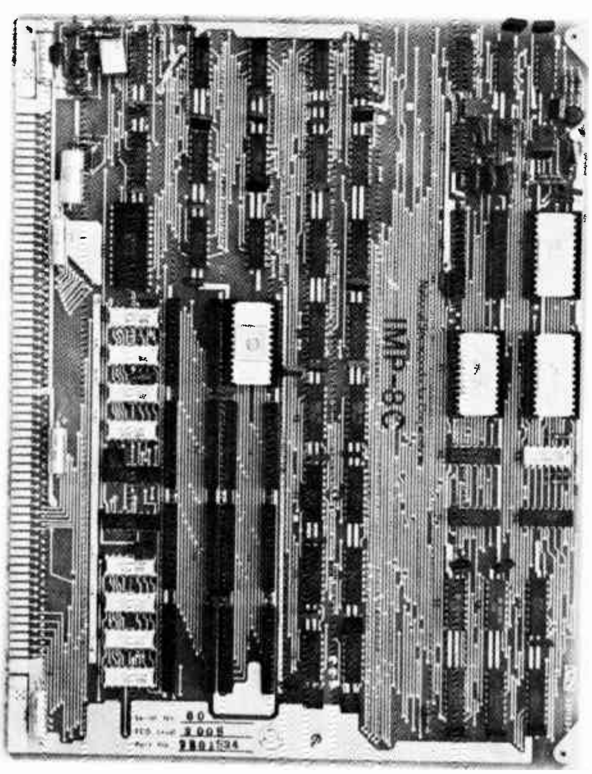


Fig. 17. This National Semiconductor IMP-8C general purpose processor uses MOS/LSI devices.

ABA	Add Accumulators
ADC	Add with Carry
ADD	Add
AND	Logical And
ASL	Arithmetic Shift Left
ASR	Arithmetic Shift Right
BCC	Branch if Carry Clear
BCS	Branch if Carry Set
BEQ	Branch if Equal to Zero
BGE	Branch if Greater or Equal Zero
BGT	Branch if Greater than Zero
BHI	Branch if Higher
BIT	Bit Test
BLE	Branch if Less or Equal
BLS	Branch if Lower or Same
BLT	Branch if Less than Zero
BMI	Branch if Minus
BNE	Branch if Not Equal to Zero
BPL	Branch if Plus
BRA	Branch Always
BSR	Branch to Subroutine
BVC	Branch if Overflow Clear
BVS	Branch if Overflow Set
CBA	Compare Accumulators
CLC	Clear Carry
CLI	Clear Interrupt Mask
CLR	Clear
CLV	Clear Overflow
CMP	Compare Index Register
COM	Complement
CPX	Compare Index Register
DAA	Decimal Adjust
DEC	Decrement
DES	Decrement Stack Pointer
DEX	Decrement Index Register
EOR	Exclusive OR
INC	Increment
INS	Increment Stack Pointer
INX	Increment Index Register
JMP	Jump
JSR	Jump to Subroutine
LDA	Load Accumulator
LDS	Load Stack Pointer
LDX	Load Index Register
LSR	Logical Shift Right
NEG	Negate
NOP	No Operation
ORA	Inclusive OR Accumulator
PSH	Push Data
PUL	Pull Data
ROL	Rotate Left
ROR	Rotate Right
RTI	Return from Interrupt
RTS	Return from Subroutine
SBA	Subtract Accumulators
SBC	Subtract with Carry
SEC	Set Carry
SEI	Set Interrupt Mask
SEV	Set Overflow
STA	Store Accumulator
STS	Store Stack Register
STX	Store Index Register
SUB	Subtract
SWI	Software Interrupt
TAB	Transfer Accumulators
TAP	Transfer Accumulators to Condition Code Reg.
TBA	Transfer Accumulators
TPA	Transfer Condition Code Reg. to Accumulator
TST	Test
TSX	Transfer Stack Pointer to Index Register
TXS	Transfer Index Register to Stack Pointer
WAI	Wait for Interrupt

Fig. 18. Typical microprocessor instruction set.

with as much power as the minis of a decade ago. Predictions, at present, are that they could fall further to a mere $50.

To make a programme microprocessor system, the user has to write a software programme at a basic machine-language level. Each microprocessor has its own instruction set built in – this tells the system what to do with data. It is written in mnemonic code using code letters to denote operations – such a list is given in Fig.18.

The programme, thus written in mnemonic code, is further translated into the circuit binary form (object code) with an assembler. The object form of code is then ready to be fed into the Random Access Memory (RAM) or Read Only Memory (ROM) of the microprocessor system. Figure 19 shows a system design and verification procedure used to produce custom made ROMs. The peripherals are then interfaced to the unit and the complete computing system is ready to go.

The process may sound easy, but as

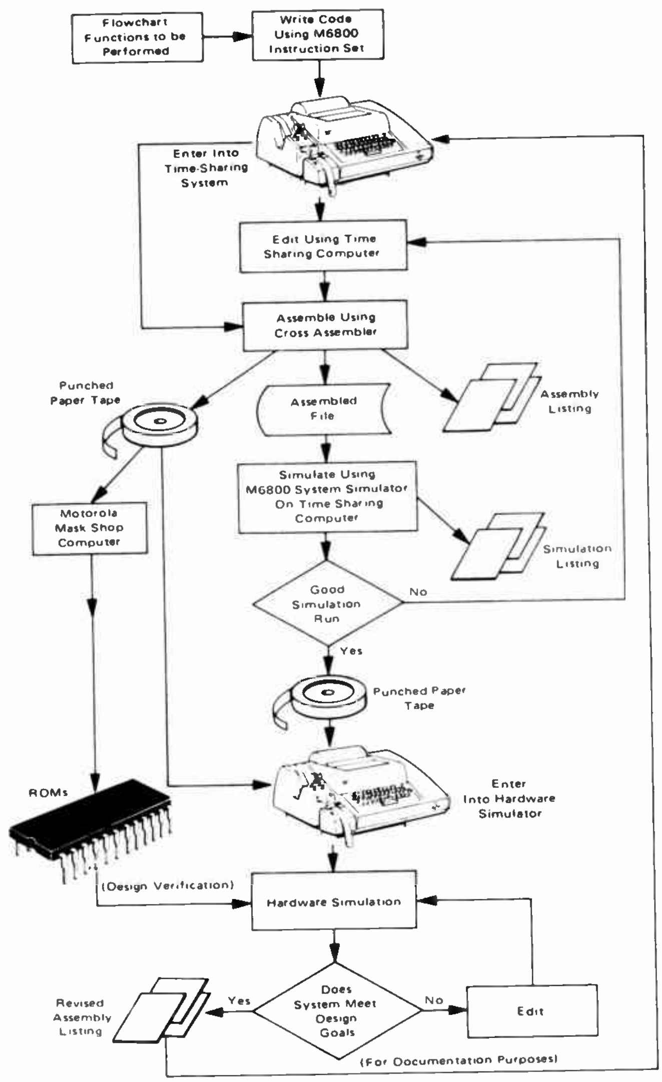


Fig. 19. This flow chart shows how the Motorola Company produces a custom-tailored ROM ready to slip into their M6800 microprocessor system.

can be imagined considerable skill is needed to set up a microprocessor. Such skill demands extensive customer training material – large handbooks, application and hardware – see Fig.20. In fact the stage has been reached where makers will soon have to face a situation where descriptive literature is as expensive or more expensive than the hardware itself. To this end they continually strive to reduce the complexity and to standardise design. We should soon be in the position where microprocessors are sold as standard tools which are each used in exactly the same way to service the enormous number of custom jobs available. Literature may degenerate to a short-form pamphlet. The actual computing part of an Everyman's Complete Microprocessor System is a very small part of the whole. Peripherals and software are now the major cost consideration. It is not feasible to describe microprocessors more fully in this series, however a series of articles devoted specifically to microprocessors was published in ETI from August to October 1976.

REFERENCES – References listed in the previous part provide descriptions and illustrations of computer peripherals and storage methods. "Computers at Work" by J.O.E. Clark, Bantam Book, is a worthwhile discussion on how and where computer interface are used for all manner of needs.

Infotech International of Maidenhead, UK, have recently released 12 000 pages of state-of-the-art reports on computer operation and trends. They are, however, much too expensive for the reader to procure. The sets costs over $2000 on an individual basis!

"Microprocessors – an Introduction" by F. Horne, NS Application Note AN114, 1974, is a basic statement.

"Microprocessors – Why They Evolved and What They Are" by M. Levi, N.S. Imp Brief 1, 1974, is also useful as a starting point.

"New Blocks for the Computer Builder" by D. Aspinall, New Scientist, 18 September, 1975 gives a basic survey including some facts about production. An extensive self-contained introduction is "Introduction to Microprocessors", H. Tireford, Motorola Semiconductor Products, 1975.

Manufacturers of microprocessors will freely supply descriptive data to aid the user of their own style of unit.

The subject of microprocessors has recently been discussed in depth in several electronics publications – Practical Electronics, Wireless World and Electronics Today International (UK edition) have each run introductory series. "Development and Trends of the Microprocessor" by J. Tobias, Control Systems (Sydney), 3, 17-31, 1975 is an extensive study and it includes a summary chart of dozens of systems offered.

Fig. 20. Array of support products for Motorola M6800 microprocessor system.

32 Transmission links and coupling

ELECTRONIC systems consist of basic analogue and digital subsystems interconnected to provide the required overall input-output relationships. It is important for the various subsystems to be interfaced correctly if they are to perform as intended. But with this condition satisfied, one cannot just assume that subsystems merely connect together without need to consider any other parameters in the interconnection process.

In practice the individual circuit assemblies may be geographically apart – such as the remote control of off-shore oil wells by a shore-based computer, the recording of test data from a missile, the control of banking accounts by a central computer centre or the sensors of a refinery which connect to the central control room. Each of these required some form of telemetry system. (Telemetry was introduced in part 5 of this series).

When making connections it is also important, especially when noise sources are present that will interfere with the signal, to ensure that the signal is transferred from stage to stage without significant noise pick-up or signal degradation.

TRANSMISSION LINKS

Several different transmission methods exist in which the signal is confined – open wires, coaxial cables and waveguides, optical fibres etc. Alternatively, information can be transferred via open radiation paths – radio, optical or acoustic links. The required signal bandwidth is one of the primary factors deciding which method is used. In radiation methods it is often necessary to use a carrier frequency higher than the signal bandwidth dictates because low frequency carriers will not radiate as well for the same amount of transmitted power.

Confined Signal Links: The simplest links are formed using an open-wire circuit (supported on insulators) or a multicore cable (such as is used in local telephone distribution).

Although apparently trivial, lines may, in fact, be an important part of the system. They are not as simple as they first appear because they have a frequency response that must be adequate for the signal bandwidth to be transmitted. Open-wire lines would not normally be used beyond 10 MHz. Above that coaxial cables are needed – these are useful to about 5000 MHz.

When currents flow in a conducting line, magnetic and electric fields are set up around the wires. Figure 1 shows these plotted for the various kinds of cable. Open configurations radiate energy, the amount increasing with the frequency of the signal. A line is, in reality, a distributed inductance and capacitance component which also has losses due to the resistance of the wire and the resistance to ground. Figure 2 shows how lines can be considered as a lumped-element equivalent circuit which can be analysed more easily. Depending upon the factors that are negligible for a particular case the equivalent can be reduced to simpler circuits – see Fig. 3. For example, at very low frequencies (less than say 100 kHz) a medium length line may be represented by the series resistance of the cable shunted by the capacitance of the line. Typical cables may have a resistance of around 0.05 ohm per metre and a capacitance of 100 pF per metre. Hence a long length of shielded or open cable could provide a considerable shunting effect that attenuates and phase shifts the signal.

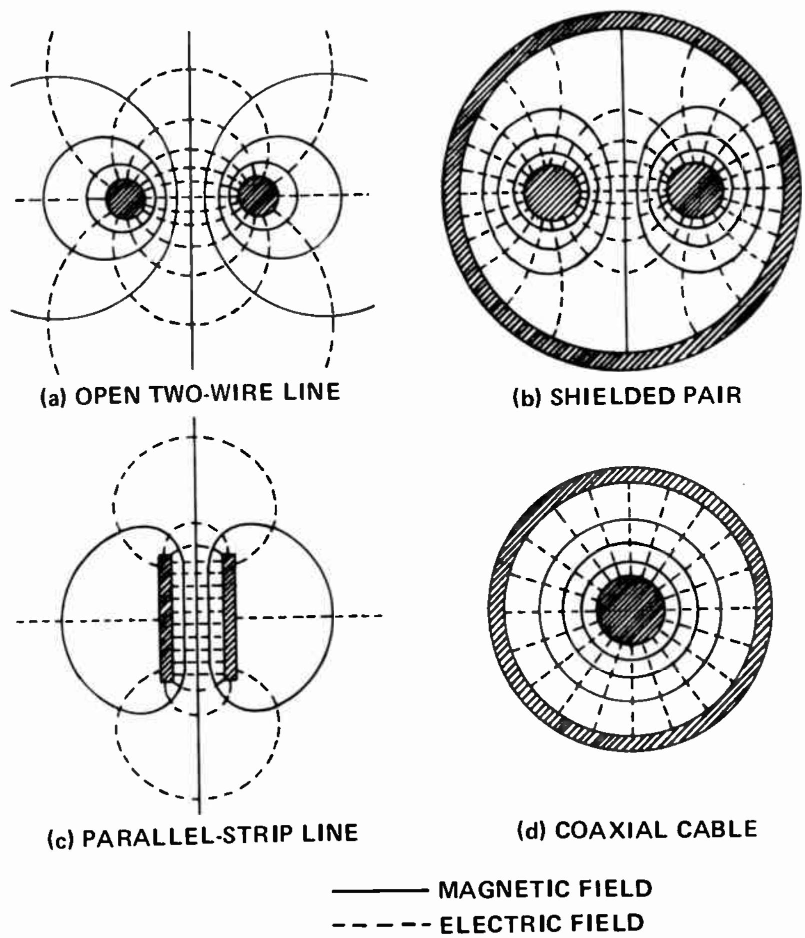


Fig. 1. Currents flowing in signal wires generate electric and magnetic fields. Enclosed configurations can be used at higher frequencies because these fields are contained.

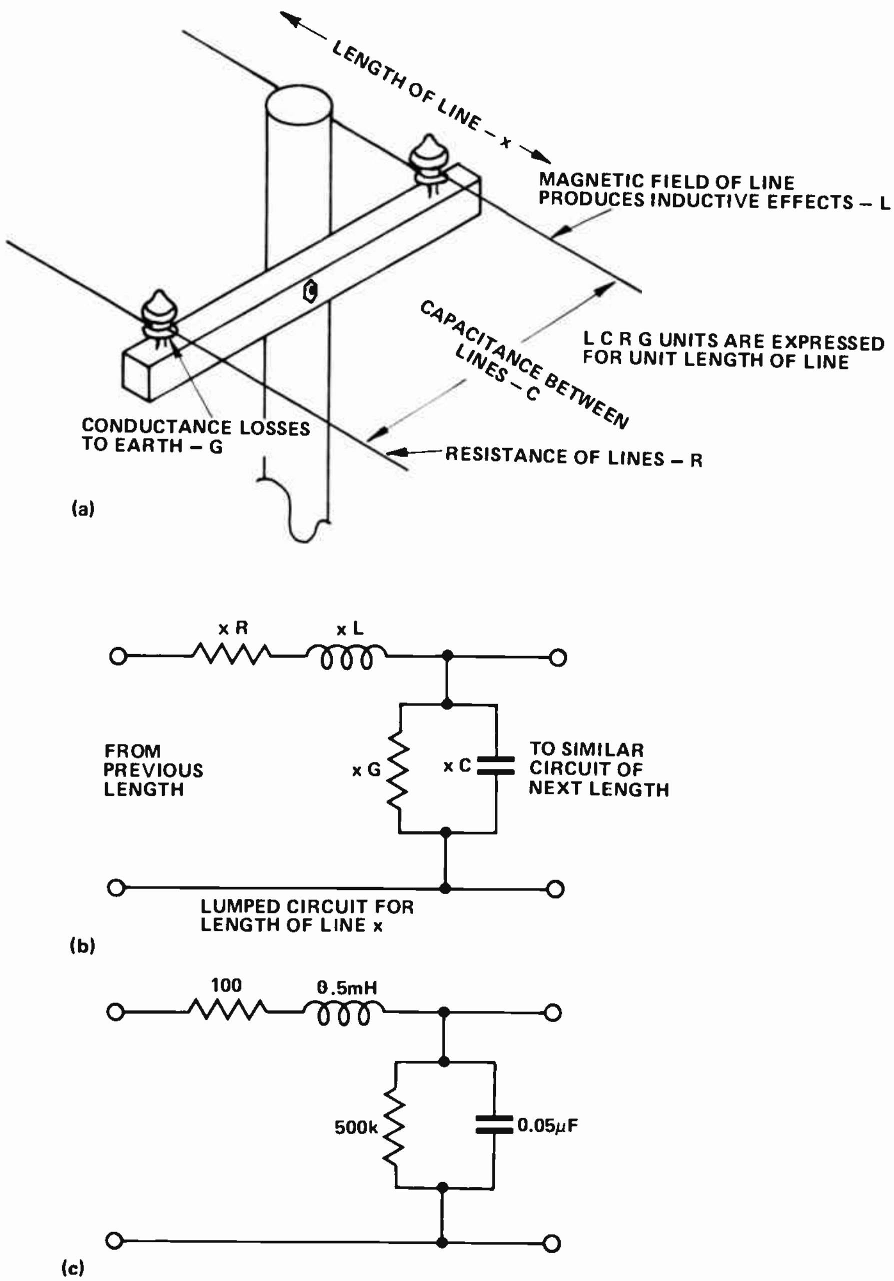


Fig. 2. Transmission links are systems in which R, L and C are distributed uniformly over the length. For convenience we can consider the line as being composed of cascaded lumped-equivalent elements.
(a) A length of low frequency telephone line.
(b) Approximate lumped-element equivalent.
(c) Representation values for 1 km. of medium-size telephone line with earth return. (Actual constants vary widely depending upon design of line).

When connecting high output-impedance sensors to lines, as little as one metre of cable may be sufficient to markedly attenuate the signal. It's a matter of applying Ohms law to the suitable equivalent circuit.

Because of the reactive effects of the cable the higher frequency signals transmitted will be degraded more than the low frequencies — for example, square waves become rounded as well as attenuated. The high-frequency performance of the line may be improved by "loading" it with inductors placed at regular intervals. The inductance value is chosen to tune out the inherent capacitive reactance at the upper frequency where response begins to fall off, a method that extends the bandwidth some way beyond the inherent unloaded upper limit. This is used, for example, to broaden the bandwidth of submarine cables.

The coaxial cable, shown in Fig. 4, by virtue of the surrounding external shield (Fig. 1) acting as the second wire, has no external field and, therefore, does not radiate energy. Because of this a well designed coaxial cable will pass from dc to microwave frequencies — that is, such a cable can have a bandwidth of about 5000 MHz. Coaxial cable is, therefore, potentially able to transfer much more information than open wires. It does however need a common earth connection (asymmetric) and can't be used in a balanced mode (see later). The bandwidth of practical coaxial cables is limited by resistive and dielectric losses. In practice waveguides are generally used at frequencies above 1000 MHz or so.

Waveguides consist of precise pipework — they look as if they had been made by a precision plumber! Waveguides carry travelling electromagnetic waves of very high frequency and behave vaguely in the same way that pipes carry water. They cannot however be used for low frequency transmission.

The cross-sectional area of a waveguide is inversely proportional to the design frequency. As a general rule of thumb guide the upper frequency limit of a waveguide is where the wavelength of the signal becomes one quarter of the guide aperture — millimetre wavelength signals (50 GHz or so) being the practical upper limit.

Beyond this, a still wider bandwidth is obtainable using optical fibre transmission elements which will pass radiation in the visible light region (10^{14} Hz to 10^{15} Hz). At our current state of technology, however, scientists have only been able to detect the frequencies of far infra-red signals (around 10^{11} Hz). We cannot, as yet, monitor individual cycles of light with electronic detectors.

When the losses of the line are insignificant (G=0, R=0, in Fig. 2b) the lumped-equivalent of the transmission lines reduces to L in series and C shunting, as shown in Fig. 3b. The nett result is, rather surprisingly, that the line exhibits only resistance of a fixed value when looking into the ends. This is called the characteristic impedance, Zo, for which Zo = (inductance per unit-length/capacitance per unit length)$^{1/2}$. The line appears to be purely resistive and the Zo value is decided by the design of the line or

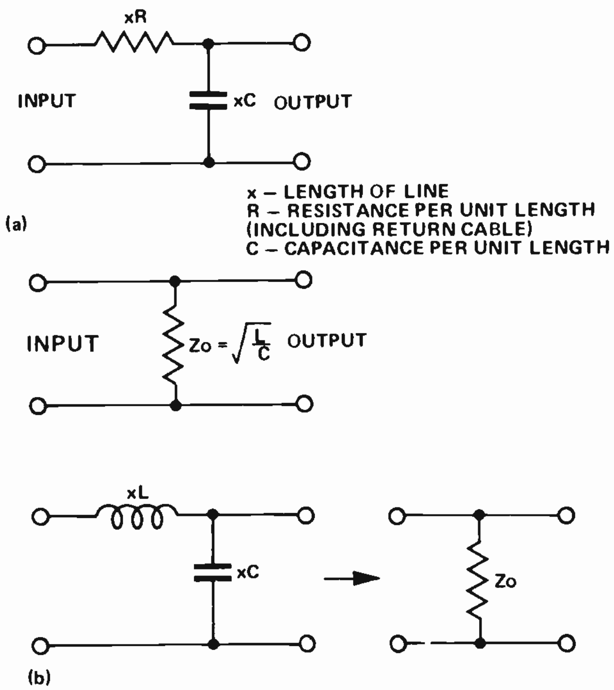


Fig. 3. In certain practical cases the lumped equivalent reduces to simpler situations situations.
(a) Low frequency (negligible L assumption) short line in which only C and R are dominant. The R, C values are found from maker's data.
(b) High frequency lossless line (negligible R and G assumptions). The input and output impedances of the line are equal and constant regardless of length.

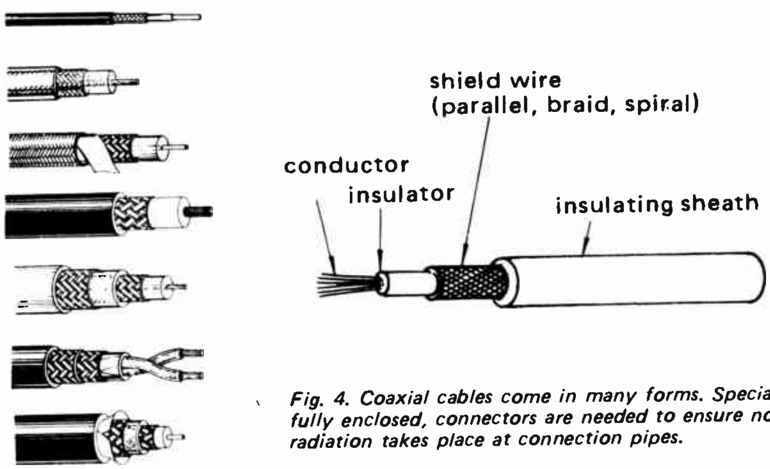


Fig. 4. Coaxial cables come in many forms. Special, fully enclosed, connectors are needed to ensure no radiation takes place at connection pipes.

cable, not by its length! Examples are 600 ohm telephone lines, 75 ohm colour TV coaxial feeder cable.This means, in practice, that we can interconnect units on the basis of matching all connections to the Zo of the cable without having to worry about the cable length. If this rule is observed, no high-frequency energy will be reflected at the termination to change the information being transmitted. (The need for correct matching was also mentioned in the previous discussion about filters). However, if the line is very long matching must still be applied to obtain maximum transfer, but account must now be taken of losses. For example a typical 75 ohm coaxial cable will have losses of the order of 2 to 5 dB per one hundred metres.

Radiation Links: Electrical signals fed into open wires radiate energy out into the surrounding medium. As well as this radiated energy there also exists a "near field" that remains established, storing energy. This is the field we associate with, say, an electromagnet. As the frequency rises, the ratio of radiated energy to stored energy increases. For this reason we are able to build efficient radio systems provided the frequency is kept above 100 kHz or so. Lower frequencies can be used as transmission systems but the power input needs rise enormously for the same distance radiated in free space. (The Omega navigation system uses extremely powerful VLF signals because of their ability to penetrate deep into the waters of the ocean) Beyond the gigahertz frequency region, circuitry becomes impracticable with current technology.

Even though the radiated energy must be at a very high frequency to operate efficiently we may not necessarily need to use the bandwidth available on the carrier. The modulation techniques we met briefly in Part 5 are used to super-impose a relatively narrow bandwidth signal onto the carrier. It might be thought that optical and infra-red links use extremely high carrier frequencies (330 000 GHz for red light) but in these applications the carrier is not modulated on an individual cycle basis but rather as variation of a continuous dc link. Figure 6 shows what might be the first electro-optical link – its bandwidth would have been barely 150 Hz. In contrast, Fig.7 is a modern link designed to transmit television plus speech commands – a bandwidth of 7.5 MHz. Acoustic links using soundwave propagation operate with frequencies as low as 10 Hz to well above the 10 MHz region. These can be modulated on the individual cycle basis.

Skin Effect: The alternating magnetic field produced around a wire has the effect of causing the current flowing in the wire to flow at a greater density in the outer region of the wire. The higher the frequency the more pronounced this so-called skin-effect. At the very high frequencies so little current flows in the centre of the cable that the centre is often omitted completely, thus a tube is used as a conductor. For example, at 1 MHZ the majority of the current flows in a copper cable to a depth of only 60μm whereas at 60 Hz the distance would be 8.6 mm depth. This also means that the effective resistance of a wire rises significantly with frequency – by factors of 100.

Process Industry Telemetry Links: Process plants such as oil refineries, paper mills, brick kilns, power stations and aluminium refining plants are monitored by using hundreds of sensors connected to the control-room area via instrumentation links. These are invariably wired using shielded wire or coaxial cable. Because of the extreme electrical noise level of such plants and low output signal level of the sensors these links could pick up significant noise thus degrading the

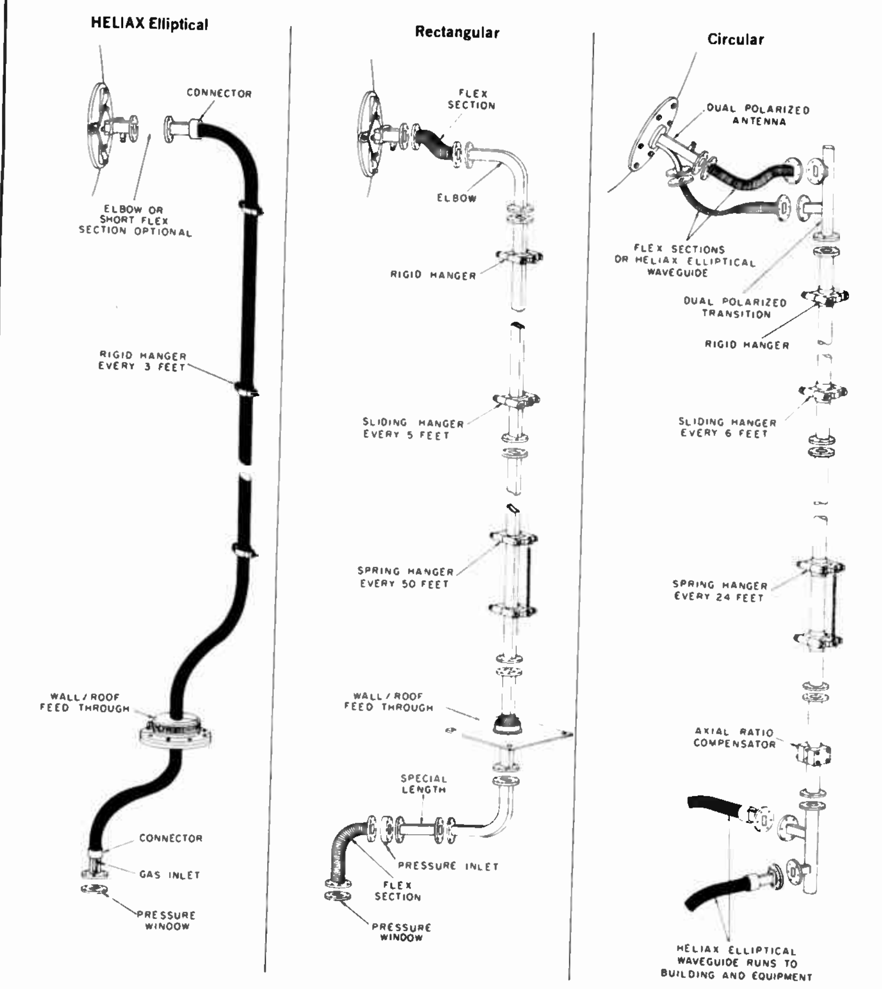


Fig. 5. Waveguides transmit electro-magnetic energy via travelling fields that are conveyed through the pipework. These are three forms of antenna feeder offered by Andrew Antennas.

sensor information. Over the years process instrument suppliers have standardised the design of the control systems, and their installation and noise pick-up by the cable has been avoided by several methods.

The first strategy is to superimpose the information signal onto a standing current or voltage thus raising the wanted signal level above expected noise levels. The two systems commonly used transmit the signal range of the data through 4-20 mA dc or 10-50 mV dc systems. An 0-20 mA system is also common. Current transmission has the advantage that the circuit is of low impedance — a few ohms — which reduces the level of induced noise power. Figure 8 is an example of these practices — Honeywell's arrangements used to test the temperature and pressure of natural gas wells in the Leman Field of the North Sea.

Safety Precautions: Often the sensor has to be placed at a location where an explosion could result from a spark or excessive overheating of a malfunctioning sensor circuit. The most obvious way of overcoming this is to place the whole unit in an explosion-proof enclosure. This, however, has disadvantages: the cost is high, and testing and maintenance difficult due to the need to shut off the power when the enclosure is opened.

The alternative, more modern, method is known as intrinsic safety. As inflammables require a specific level of energy to ignite them, explosion can be prevented by ensuring that the sensor stage cannot, under any conditions, provide enough ignition energy. No enclosures are

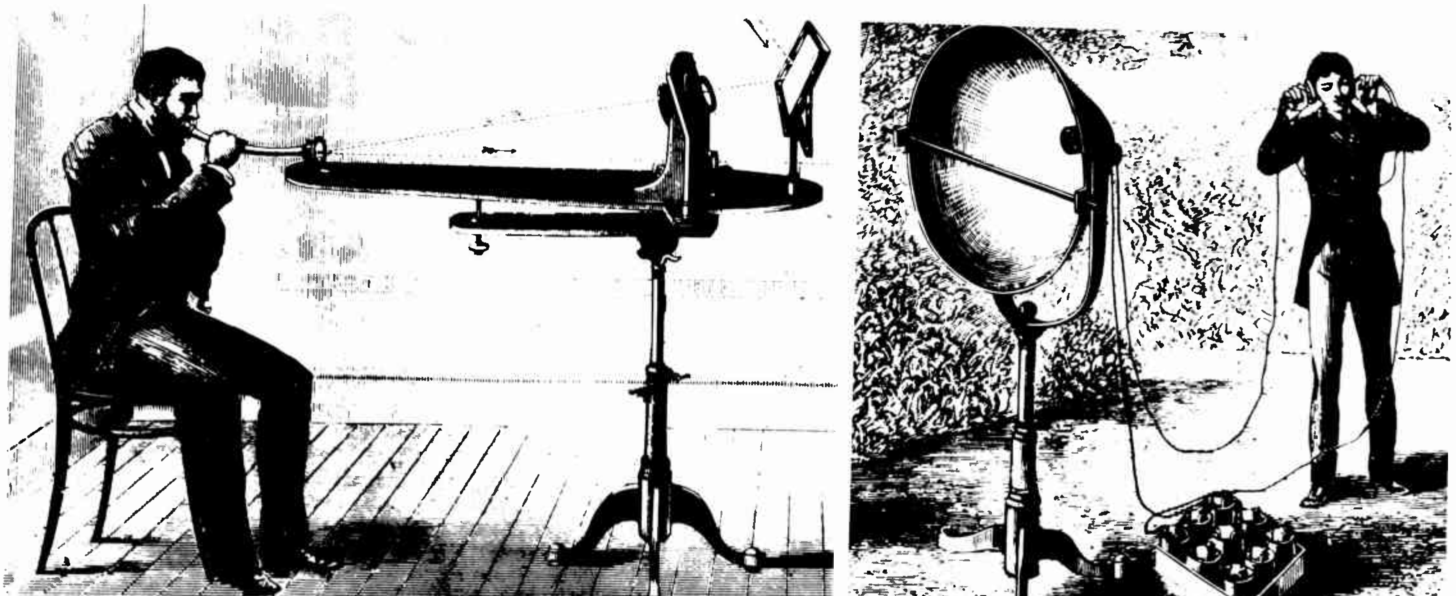


Fig. 6. This Photophone of Bell and Tainter was designed in 1881. Sunlight reflected to the receiver was modulated by acoustic waves vibrating the speaking tube mirror. Detection was with a selenium photo-electric cell driving earphones.

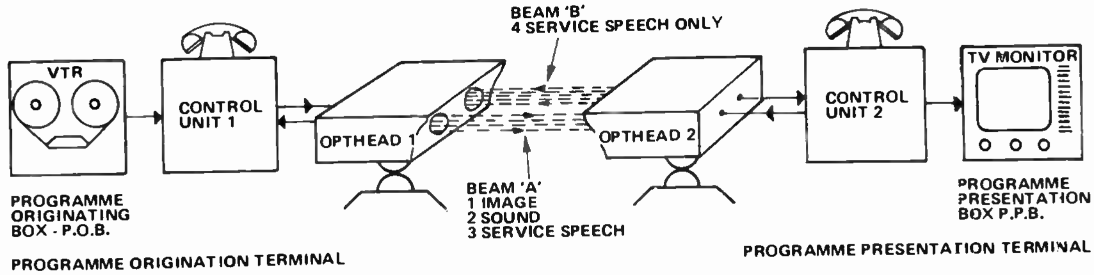


Fig. 7. Schematic with photograph of the optohead of a Leevers-Rich opto-electronic communication link. The output from a light emitting diode is modulated by the incoming signal. The receiver detects the modulation with a solid-state photo-detector.

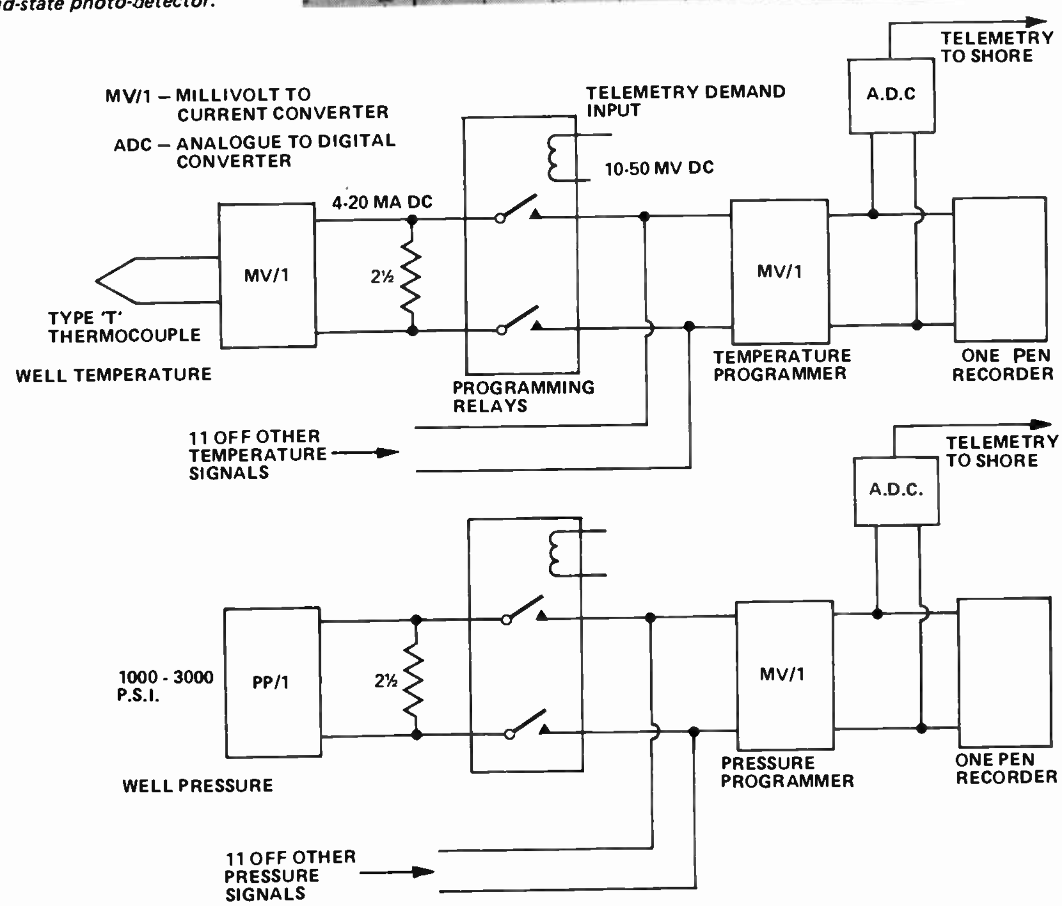


Fig. 8. Process measurement link arrangements used between the oil-well and the off-shore platform. The A/D units send the data to a shore-based computer by digital telemetry links.

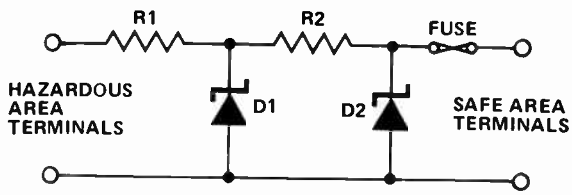
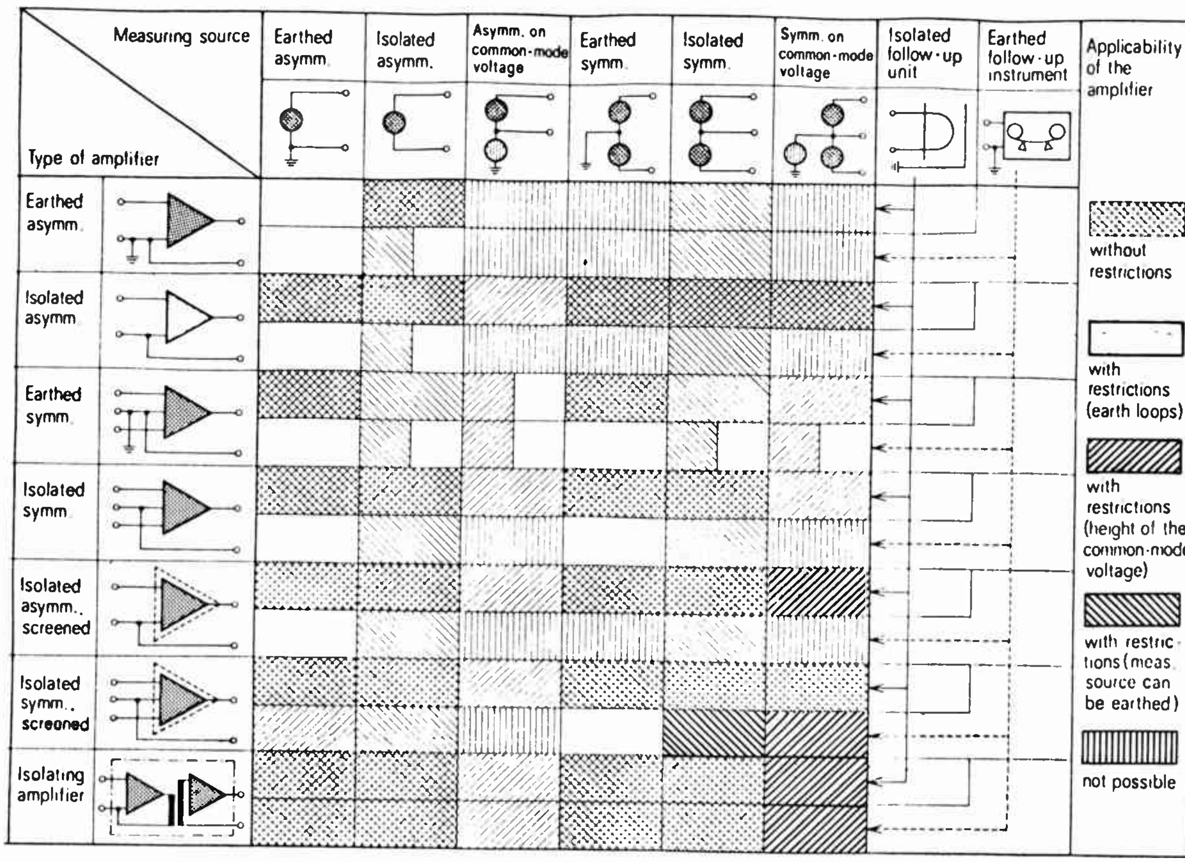


Fig. 9. Circuit used in a Zener barrier.

Fig. 10(a). Chart showing common combination possibilities of various output to input cascaded schemes incorporating amplifier stages of various kinds between the first stage and the two commonly used output recording/monitoring connections. (Courtesy Siemens Industries).



needed and the circuit can be maintained whilst it is operating. Originally the concept was implemented by ensuring the sensor circuitry could not draw, or produce via storage, more than a specified power level. This level was found by experiment in a test rig set up for the situation involved.

The more recent idea is to use "safety barriers". At the exit from the declared hazardous area, the cables terminate into a zener-diode and attenuator arrangement which ensures that the current and voltage entering the area are limited to safe values. Figure 9 shows the circuit of a zener barrier. Another safety device uses a solid-state closely-coupled electro-optic link which provides dc electrical isolation between its input and output, the information being transferred from a light-emitting-diode mounted next to a silicon photo-diode detector. These ensure that overvoltage or induced earth-loop currents cannot enter the isolated hazardous area.

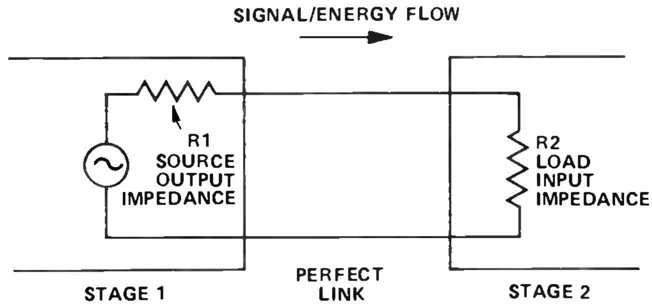
In electro-medical instrumentation, safety precautions of another kind are vital to ensure the sensor does not act as a pathway for a dangerous level of electric current into the patient. At 240 Vac the human body's resistance, hand to hand is around 2000 ohms — 100 mA will flow. If totally connected (as by a conducting fluid) the resistance reduces to 200 ohm — 1 A will flow. About 75 mA through the body will produce heart fibrillation; only 150μA, through the heart itself, is needed to produce this effect. A person can usually hold (with the fingers) and release as much as a 10 mA, 240 Vac current — beyond that the muscles become paralysed. Skin moisture largely decides the hand to hand resistance. When dry it will be (at 240 V) 2500 ohms and moist, 1000 ohms. *Thus a hand-to-hand 240 V encounter will provide a shock at least double the fibrillation level!*

The instrumentation must, where the metal parts are earthed, be wired with the active, neutral and earthing wires connected correctly. Double-insulated systems avoid this problem. Earth-leakage balanced — core breakers are worth using. These detect minute difference currents in the active and neutral, tripping a breaker if they rise above milliamperes.

The sensor attached to the patient must not be capable of providing a lethal level of energy by means of feedback from the instrumentation. The reading list provides more detail of safety factors in electro-medical instrumentation.

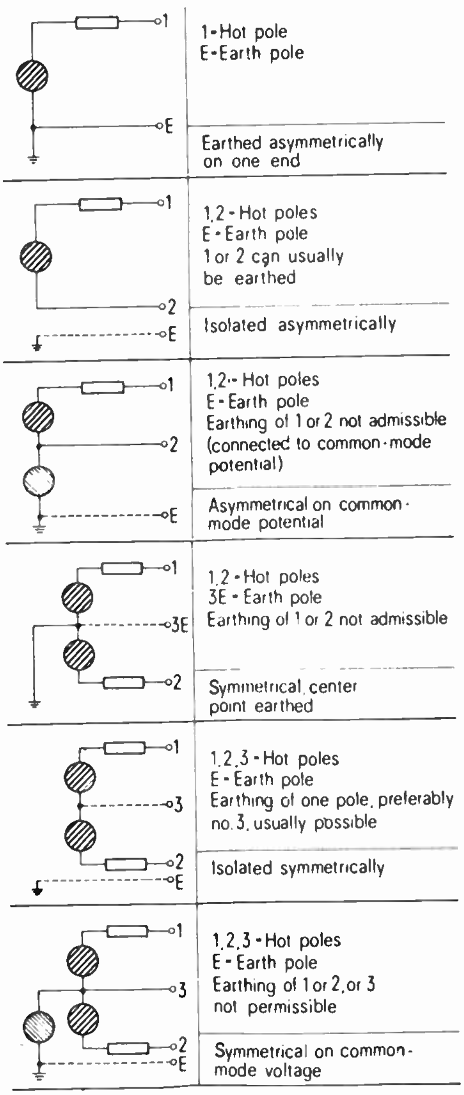
COUPLING STAGES

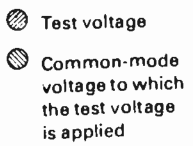
Connection arrangements: As was pointed out in the discussion of meters, in Part 3, electronic sub-systems must be cascaded



Matching requirement	relative R1, R2 values
Maximum energy transfer	R1 = R2
Maximum voltage signal transfer (least voltage loss across R1)	R2 >> R1
Maximum current signal transfer (least current change due to R2)	R1 >> R2

Fig. 11. Summary of impedance values for various matching requirements.





(b) Extra detail of source arrangements.

intelligently or loading of the output of a stage by the input impedance of that following may degrade the signal. Output configuration of the various stages involved in instrumentation can take many forms depending on how the earth is connected and if the signal is symmetrical or assymmetrically connected. The six commonly encountered source output schemes are shown at the top of Figure 10. On the left-hand side are seven common kinds of amplifier connection (any other form of black box could be regarded similarly). On the right-hand side are leader lines that show a link between the output of the chosen amplifier and one of the two most commonly used instrument connections — fully isolated circuit with case only grounded, or one pole grounded to earth. Using the legend, the chart shows the applicability of connections between chosen combinations of source arrangement, amplifier and output device. Not-possible situations usually arise because the earth connection shorts out one of the source arms.

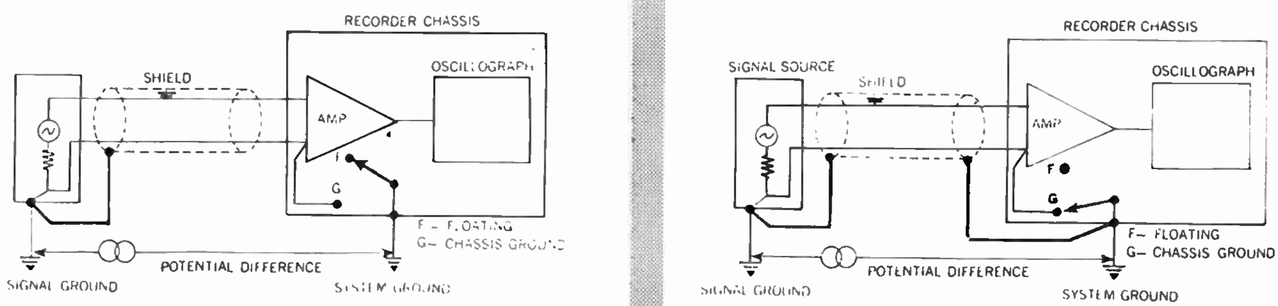
Matching: Three basic matching criteria exist when connecting two stages together. Figure 11 summarizes these.

If the need is for maximum power transfer, as when driving a loud-speaker from an output stage of an amplifier, the output impedance (usually thought of as an average value of resistance) of the driving stage must equal the input of the stage being driven. When maximum voltage transfer is required, as occurs when a pick-up cartridge or other voltage generating transducer is used or when measuring a voltage in a circuit, the rule is to ensure the connecting stage has a much higher input resistance than the output resistance of the stage producing the voltage signal. A factor of ten to one hundred times is usually sufficient.

The opposite situation, that is, loading a high output impedance stage with a low input impedance, arises when the maximum current transfer is required.

In many cases the appropriate buffer amplifier is required to provide the desired matching condition. In certain ac coupled systems — those which do not require a dc path between stages — a transformer can provide an adequate impedance match in an economic way. Transformers, however, have limited frequency response and must be chosen carefully to suit the signal requirements.

Eliminating noise: In the ideal situation any circuit added after another should add no more noise energy to the signal than is fed to it. We specify the ratio of the two as the signal/noise or S/N ratio. In practice



RIGHT

WRONG

Fig. 12. Correct and incorrect methods of joining a sensor to a recorder. Most output instruments offer the user the choice of leaving the instrument floating above ground or grounding it.

all circuits, including connections, will add a finite amount of noise – degrading the S/N ratio. An amplifier or other cascaded stage should ideally increase or modify the signal amplitude without reducing the S/N ratio (input noise will be amplified equally with the signal).

A common unit used to describe the degree of degradation is the noise-figure NF which is calculated as NF = 10 log (Signal in/Noise in, divided by Signal out/Noise out); the ratio being expressed in decibels (dB). The perfect additional stage has NF = 0 dB, So-called low-noise amplifier stages will have noise figures better than 3 dB (S/N ratio reduced to half). NF is a function of device characteristics, frequency of operation, source resistance and temperature – the correct choice of components to yield a low NF is a skilled task.

Connections between stages are most common source of noise addition. Observing several basic rules will usually greatly reduce the noise pick-up in wiring between and within stages.

Grounding and Shielding: When wiring circuits and inter-connections the circuit diagram shows a signal ground. (Terms ground and earth are used somewhat synonymously). This line is assumed to be at exactly the same potential at all points where a ground symbol is indicated. From the electricity supply authority's viewpoint any good low resistance connection to mother earth is a good ground or earth point. But this is not so for instrument stages operating at millivolt and microvolt signal levels. Signals as large as volts can be induced, or dropped, between two points of a metal chassis! The rule for avoiding this ground loop problem is to attach all circuit points required to be grounded to a substantial size copper bus bar – the circuit ground – that is grounded to earth at one place. Better still, use a single common connection point.

Shields of cables are too often assumed to have the same potential at each end, both ends being presumably at ground potential. This is often incorrect for the shield becomes an earth-loop having a finite resistance when both ends are grounded. Only one end, the input end, should be earthed and the shield should be insulated against earth at all other points. Figure 12 shows the right and wrong ways to connect two stages together with a shielded two-core lead. Special quality low-level signal cables are available. These incorporate an inner twisted-pair that is wrapped inside a multi-layer metal foil along with a bare copper drain wire, the whole being well insulated.

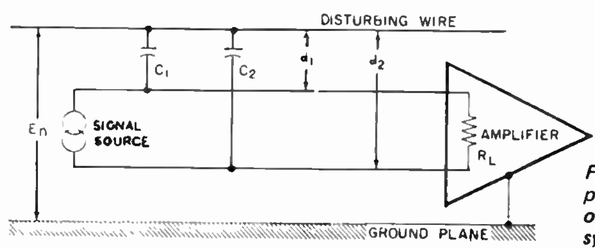


Fig. 13. Common-mode pick-up of noise is balanced out by the differential system.

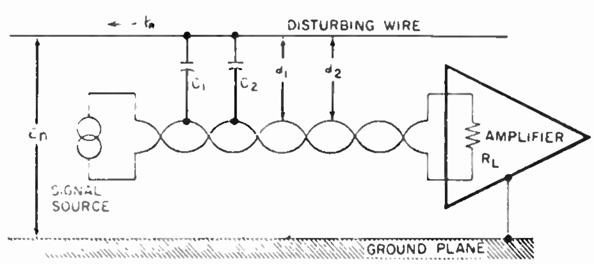


Fig. 14. Twisted wires provide best common-mode rejection when making connections between stages in which a symmetrical signal source is joined to a differential input amplifier.

Common-mode rejection: Before other aspects of connections with cables can be appreciated we need to study the principle of common-mode signal rejection.

We begin by looking at the noise pickup from supply mains radiation by two open wires used to complete a link, as shown in Fig.13. If both wires are at the same potential above earth, that is, neither is earthed, the noise pickup in each wire will be closely similar. One wire, however, passes signal currents in the opposite direction to the other so noise induced in each wire will add to the signal in one wire and subtract in the other — the result is that the noise just about balances out. This is known as a common-mode rejection arrangement.

It is a balanced system as far as unwanted signals are concerned because of the use of a differential arrangement.

The same concept is used in low-noise, high gain, dc amplifiers — see part 11 — to eliminate transistor defects. A slight disadvantage of differential configurations is that many testing instruments operate with one grounded input. Connecting an oscilloscope to probe a differential-mode circuit may short out a line to ground in certain connections. For such work a differential input amplifier is essential in the oscilloscope.

Once the signal level has been amplified well above the ambient noise levels the symmetrical dual output can be converted to a single pole with earth output, using a suitably connected operational amplifier.

For the best low-level signal transfer, wiring between stages should observe the common-mode principle, the aim being to make each wire of the pair appear as identical as is possible to the interfering noise sources present. Figure 14 demonstrates why the twisted pair is better than two separate lines to connect a symmetrically-connected source to a following differential input stage. The distributed capacitances of the two wires are different (with resultant different pickup noise) in the open-wire case than they are in the twisted line.

Shielded two-core cables used with a symmetrical outputs source should have the shield grounded at the source, not at the following stage. The latter option degrades the common-mode rejection capability.

Common-mode principles must be carried through completely in exacting low-level signal applications, even to providing identical terminating conditions at the wire ends — similar length open wire ends, similar, dissimilar-metal, conditions at terminal posts with identical temperature for each to ensure identical thermo-electric currents are generated in each lead.

Active devices, such as amplifiers, have a limit to the common-mode

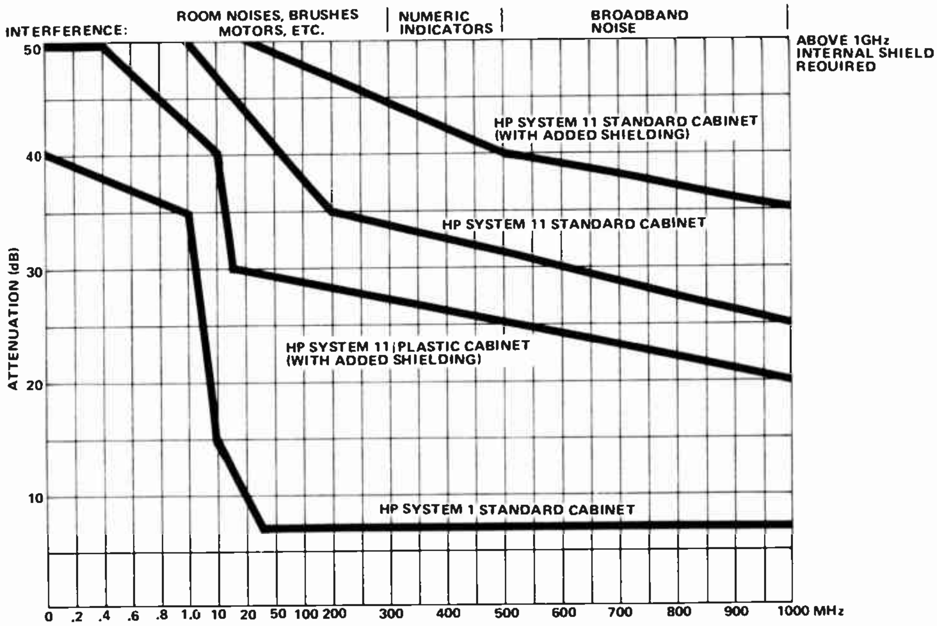


Fig. 15. Shielding of RF energy by various designs of enclosure used for H.P. instruments. The actual value of a particular unit depends upon the need for holes and shafts through the panels.

signal levels that they can handle. If the induced signals are too great in amplitude they may saturate the amplifier, removing its ability to operate correctly. It is, therefore, always best to reduce interference at source rather than attempt to eliminate it by common-mode rejection alone.

RF Shielding: Mains frequency interference (50 Hz) is comparatively easy to eliminate from or retain within equipment by using low conductivity enclosures. RF interference, however, tends to penetrate the best designed enclosure — remember waveguides transmit RF — through apertures of size similar to wavelength. Cracks, where covers join, may act as waveguides for UHF signals. As modern circuits operate with transition times of nanoseconds they too generate considerable quantities of RF energy. By way of example of what can be achieved by careful mechanical construction Figure 15 compares different instrument enclosure designs of a manufacturer. Slots introduced into frame elements form wave-traps (as opposed to wave guides) when the metal covers are bolted in. Modern instrument enclosure design is as much a case of containing RF radiation *inside* the unit as it is to prevent it entering.

REFERENCES

In depth discussions of the theory and practice of transmission methods are to be found in works on telecommunications. "Transmission lines and Networks", W.C. Johnson, 1950 is a theoretical treatment of the use of lines.

Modulation methods are discussed in most general text books on electronics. Safety in medical electronics practice is the subject of "Hazards of Electrical Apparatus", J.M.R. Bruner, Anaesthesiology, 28, 396, 1967, and "Electrical Hazards Associated with Cardiac Pacemaking", R.E. Whalen and others, Ann. N.Y. Acad. Sci., 111, 922, 1964.

A useful short guide to the practice of interconnecting system units is to be found in "Elimination of Noise in Low-Level Circuits", Clevite Corporation (Brush Instruments Division). A more extensive treatment is "Grounding and shielding techniques in instrumentation", R. Morrison, Wiley, 1967. The topic is also covered in a treatment on low-level techniques by PAR — see ETI, February 1972 for "Signal to Noise Ratio — its Optimization in Precision Measurement Systems" by T. Coor, A similar article is "Modern Signal Processing Technique for Optimal Signal to Noise Ratios", R.D. Moore and O.C. Chaykowsky, Princeton Allied Research Corporation, Technical Bulletin, No.109, 1963. "Taking noise out of Weak Signals", R. Brower, Electronics, 41, 80-90, 1968 is also relevant.

33 Oscilloscopes

OF THE MANY INSTRUMENTS required to service, test and maintain electronic systems, the cathode-ray oscilloscope must be the most versatile and useful. Other names are derivatives from the full name – the C.R.O., CRO (pronounced crow), oscilloscope and scope. Early works also refer to it as an oscillograph.

The basic workings of the oscilloscope were introduced in Part 4 where it was explained how they are used to view signal-level variations as time-versus-amplitude graphs drawn on the screen by electrons. This part provides a deeper coverage of the capabilities both simple and advanced units. We do not delve in the circuitry; although electronic system builders will always need to use an oscilloscope they will rarely need to build one.

THE CATHODE RAY TUBE

The first cathode-ray tubes were experimental, designed to investigate the nature of beams of particles produced in thermionic-diode arrangements operating at extremely-high voltages.

Figure 1 shows the three stages in developing the basic cathode ray tube. Fig. 1(a) is a thermionic diode – a valve diode. The cathode, heated by the current passing through it, emits electrons into the space around it. These, being negatively charged, are attracted to the positive anode.

The greater the voltage between the cathode and anode the greater the velocity of the electrons. If a hole is made in the anode, as in Fig. 1b, many of the electrons will pass through, forming a diverging beam on the other side of the hole. When a phosphor powder is placed on the inside of the tube the electrons reaching it cause it to glow as they give up their kinetic energy. The powder re-emits this energy as photons of visible light. Early researchers' tubes did little more than this. The nature of cathode rays was studied in the early 1900s by such famous names as Goldstein, Braun, Crookes, J. J. Thompson, Rontgen, Coolidge and Dumont. Experiments showed that the beam could be deflected by a permanent magnet and by electro-magnetic and electrostatic fields. Prior to 1897 interest had been in physical-science investigation – not in the measurement of electronic signals. Then in 1897 K.F. Braun produced the first basic measuring device from the CR tube.

FURTHER DEVELOPMENTS

However the CRO to become a useful, practical instrument more development was needed. From Fig. 1b it can be seen that the beam of an elementary device is badly defined and floods over the entire area of the phosphor. A tube or grid arrangement placed between the cathode and anode causes the beam to pass through the anode more cleanly, because of the negative repulsive effect of this tube assembly. The whole assembly – cathode, anode, grids and tube – is called the electron gun. Its full design is quite complex: Other elements are used to make electron-lenses (akin to optical lenses and light) to focus control and an intensity control, the former adjusts the spot shape and size on the screen, the latter the current flowing in the electron beam.

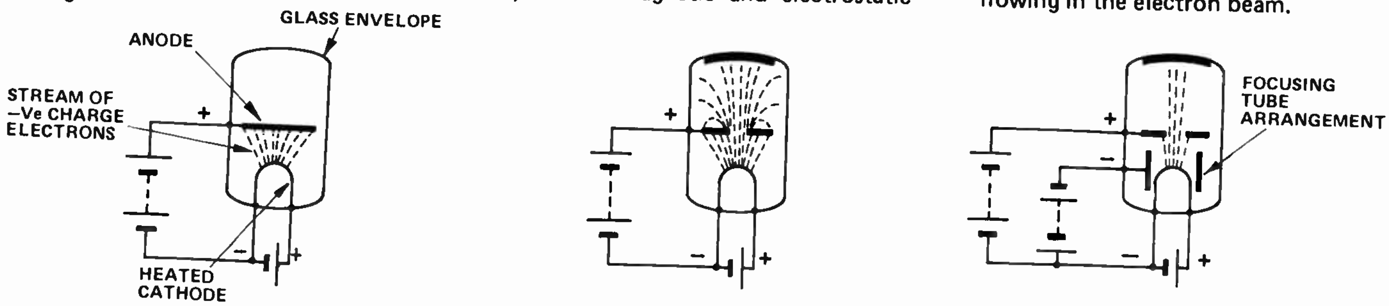


Fig. 1. The basic cathode ray tube developed through stages to provide a gun aimed at a screen.

(a) Thermionic diode in which the cathode is self-heated to cause liberation of electrons which move to the positive anode.

(b) A hole in the anode allows some electrons to pass through to the phosphorescent screen.

(c) A tube arrangement or grid is added to form a more concentrated and smaller size beam.

Phosphor European/ U.S. code	Fluorescence	Phosphorescence	Persistence	Burn resistance	Relative luminance	Comments
GP/P2	Bluish-green	Green	10 μs–1 ms	Medium high	55%	Medium speed oscillography
GM/P7	Purplish-blue	Yellowish-green	100 ms–1 s	Medium	35%	Low speed oscillography.
BE/P11	Blue	Blue	10 μs–1 ms	Medium	15%	Best photographic writing speed.
GH/P31	Green	Green	10 μs–1 ms	High	100%	General purpose oscillography. Brightest available phosphor.
GR/P39	Green	Green	5–100 ms	High	50%	Sampling oscillography.

Fig. 2. Chart showing characteristics of oscilloscope screen phosphors.

The choice of phosphor on the screen determines the persistence (the length of time the spot glows after removal of the beam) of the display. The storage effect of various phosphors enables CROs to be made so that beam energy can be dispersed as light over time durations varying from microseconds to milliseconds. Fig. 2 is a guide to the selection of a phosphor. Manufacturers often offer a choice of screen persistence values to suit various applications. Fast moving spots, where the spot is likely to reappear on the same point in a short time, require short persistence. Long-persistence screens are suitable for slowly changing signals. (See the discussion of storage methods in the next part.)

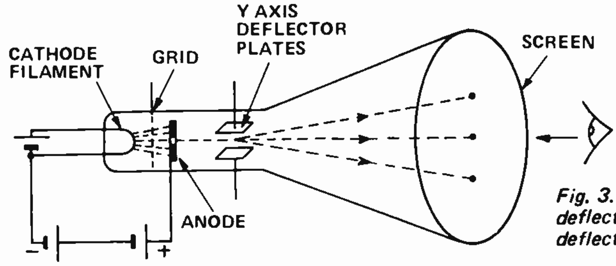


Fig. 3. The electron beam can be deflected by voltages applied to deflection plates.

ELECTROSTATIC DEFLECTION

The next refinement provides a method by which the beam can be made to deflect under the control of electrical signals. Fig. 3 shows how this is done for one axis, using the electrostatic method. A voltage difference of zero between the deflection plates allows the beam to pass along the tube axis undeflected. Any voltage differential will cause the beam to be deflected towards the more positive plate. Thus we have a way to cause the beam to move in the vertical direction (called Y-axis or Y plates). A further two plates set at right angles to these (the X plates) will cause the beam to deflect in the horizontal plane when a voltage is similarly applied to them. Beam-intensity control by electrical means is defined as the Z control.

Electrostatic deflection is the easiest to deploy for voltage measurements because deflection is proportional to applied voltage. Small cathode ray tubes usually use electrostatic deflection. Large tubes, such as those used in television systems or large-screen teaching oscilloscopes, usually use magnetic deflection because electrostatic deflection would require very high deflection voltages. These do not have deflection plates set inside the tube, but make use of magnetic fields created by electromagnet coils placed around the neck of the tube. The deflection in this case is approximately proportional to the current in the coils.

Cathode ray tube design (for CROs and TV) has remained reasonably static since the late 1930s, the only obvious differences being in the linearity of beam sweeps and the shortness of tubes for a given screen size in television applications. Figure 4 is a modern oscilloscope with the cover removed to show the tube. From the instrument viewpoint the differences have been improvements in frequency response, spot control, linearity of sweep and a wider choice of phosphors. In addition the development of tubes with more than one gun and deflection system (some dual-beam oscilloscopes, but not all, use separate beams for each channel) and storage tubes which enable the effective persistence to be varied at will have greatly improved the versatility of today's instruments.

TURNING THE TUBE INTO A MEASURING INSTRUMENT

In the majority of cases the CR tube is used to produce a graphical display with the amplitude of a signal being expressed in the vertical (Y) direction and its variation with time being along the horizontal (X) direction. The following description is to be read in conjunction with the schematic of Fig. 5 and the panel layout of a portable unit shown in Fig. 6.

Time-base: If the X plates are driven by a signal voltage that increases proportionally with time the electron beam will be deflected across the tube at a steady speed. When the signal returns to its original value the spot returns to begin the next sweep. The waveform required to produce such linear deflections is a sawtooth. (During return the beam is normally blanked out.) This provides a sweep function. The period of the sawtooth determines the time taken to cross the screen; this is expressed in the units of time per division (screens are divided into a grid of centimetre squares by means of plastic graticule or by engraving the inner face of the tube). A selector switch in the time-base section of the panel (an example is

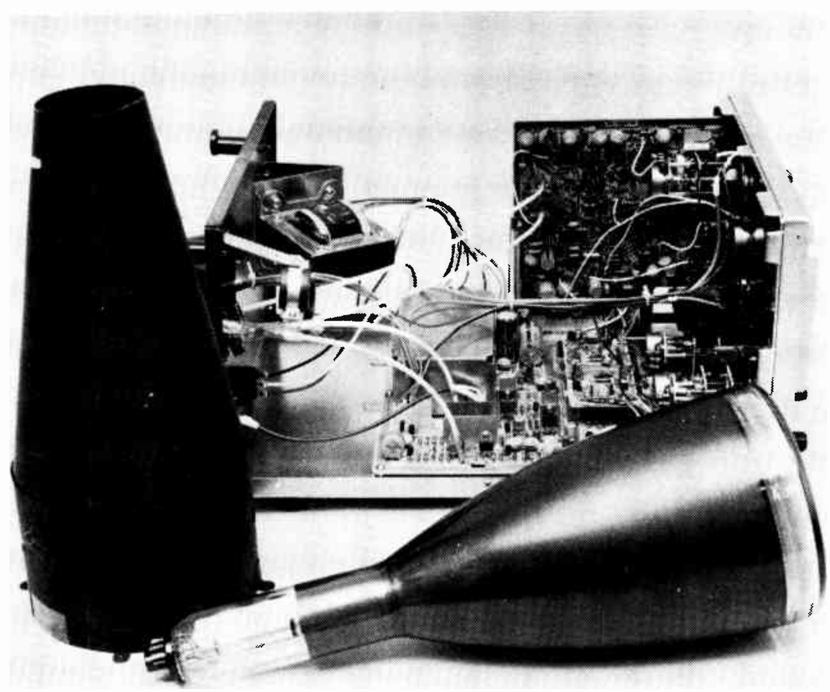


Figure 4. The insides of a modern oscilloscope – the Trio CS-1562. The cathode ray tube and its cover have been removed. Note the tube is much longer than the tubes used in TV sets (when you consider the small screen area). The cover screens the electron beam from the influences of stray magnetic and electric fields.

ELECTRONICS–it's easy!

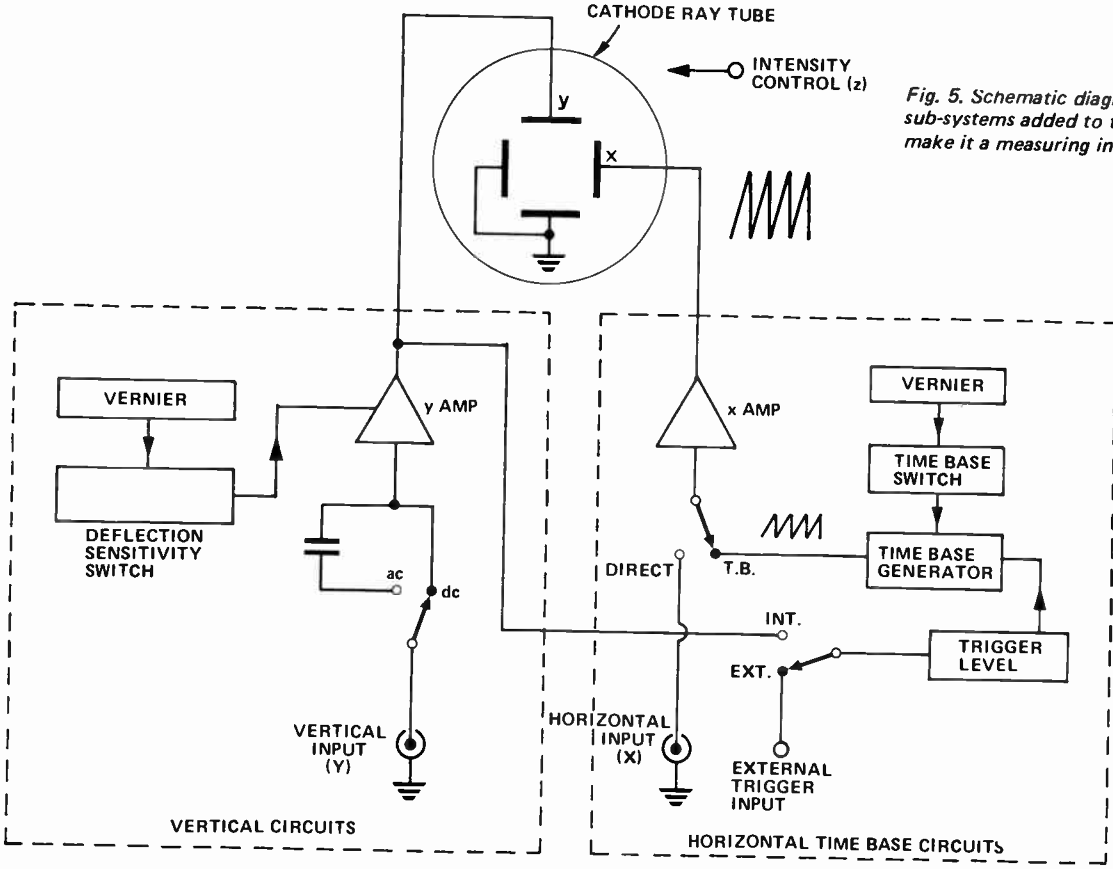
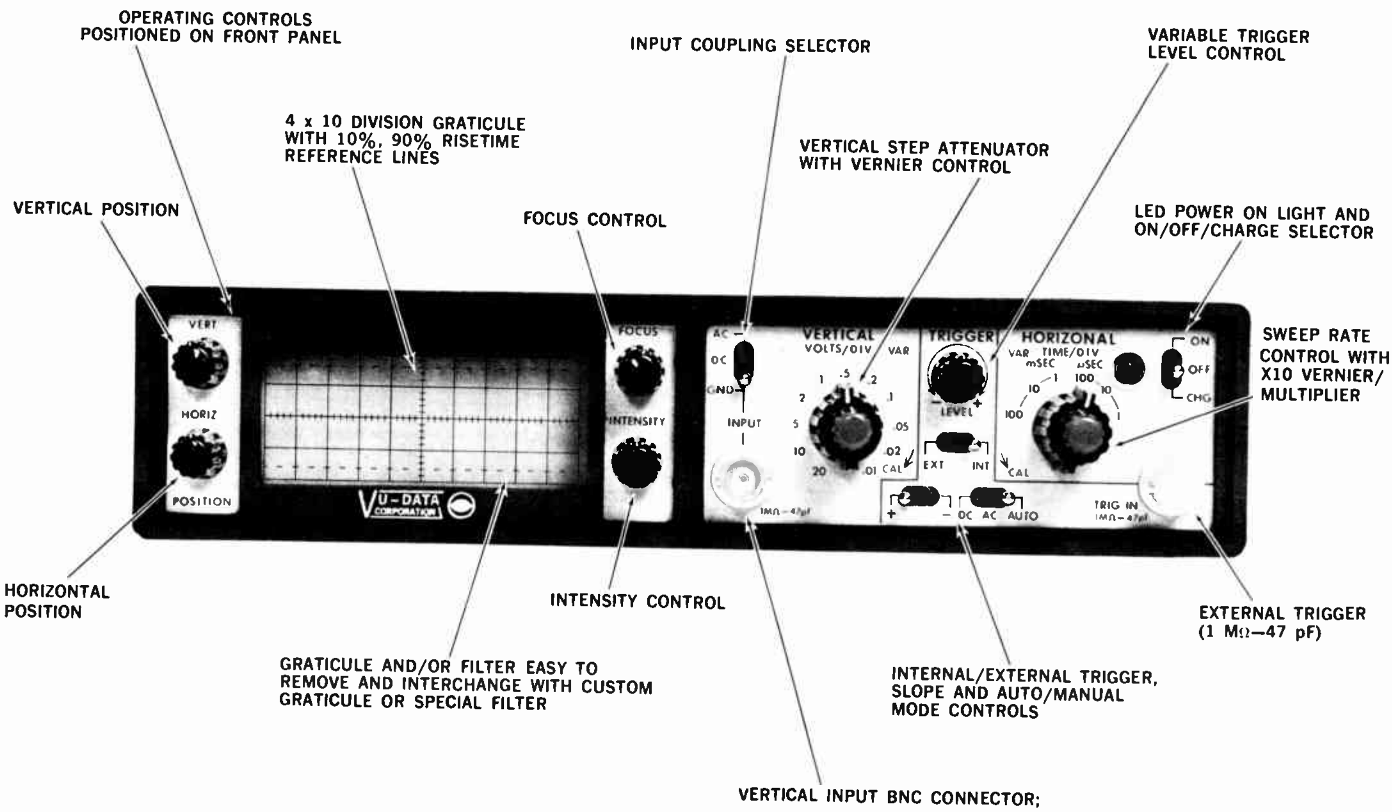


Fig. 5. Schematic diagram of essential sub-systems added to the CR tube to make it a measuring instrument.

Fig. 6. The controls of this mini-portable oscilloscope are typical of basic units. The position of controls around the screen will vary from make to make.



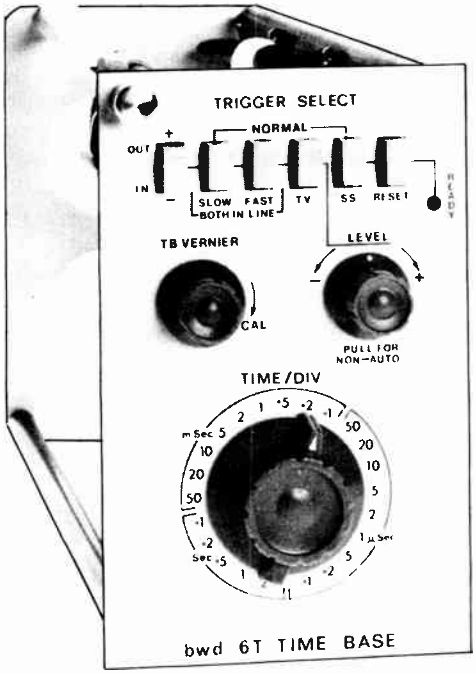


Fig. 7. Controls of a basic time-base unit include those shown on this plug-in. Terminology is generally the same for all makers but layout and controls will vary.

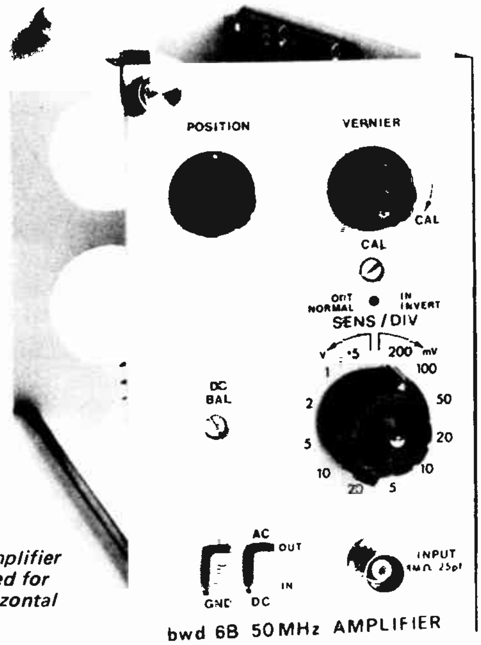


Fig. 8. Single-trace amplifier unit. These can be used for both vertical and horizontal amplification.

given in Fig. 6) enables the sweep rate to be chosen to suit the period of the signal being examined. Basic units will have time bases which range from 0.5 μs to 0.1 seconds per centimetre; sophisticated units can go as slow as 10 seconds per division to as fast as 1ns per division. (Special "sampling" plug-ins can provide 10 ps/division.)

The time-base sawtooth generator is an integral part of all CRO measuring instruments. The accuracy of the rates are determined by circuit components – more expensive units can provide more-accurate information. A further control in the time-base section (See Fig. 7) allows the switch-selected sweep rate to be varied continuously. This is usually referred to as a vernier control. When making time measurements, such as waveform period, it is important to set the vernier control to the calibrated position.

To obtain a static display (where each cycle of a periodic signal overlays the previous one) the time-base must be synchronized with some point of the input signal. That is, the time-base is caused to begin its sweep across at the same point on the waveform being viewed. The circuit which does this is called the triggering circuit. Triggering can be taken from either an internal or an external source. When switched to 'internal' it is possible to vary the voltage level of the signal operating the trigger. Thus the sweep may be adjusted to commence at a chosen point on the waveshape. An 'auto' control position provides automatic selection of the voltage level for most reliable triggering.

Time Base Amplifier: The voltage required to deflect the beam over the full X (or Y) traverse is of the order of hundreds of volts. The time-base generator therefore requires an 'X' voltage amplifier between it and the plates.

In certain applications the X plates are used with signals in the same way as Y plates – that is without a time-base signal. In such cases considerable amplification may be needed. More versatile CROs offer plug-in facilities for the X input to give the user a wide choice of functions from the one unit. Simple units however, have the 'X' amplifier wired in permanently.

Vertical Inputs: At the same time as the time-base circuits sweep the line across the screen the 'Y'-plates are driven with a voltage proportional to the amplitude of the signal of interest. This causes the beam to be deflected in the vertical direction whilst it is swept across the screen. The result is the graphical display of signal amplitude versus time.

Again an amplifier is needed to increase the signal level so that a useful vertical deflection results. Such an amplifier must be able to amplify the incoming signal without distortion to provide vertical sensitivities up to 10 mV/centimetre (typically the most sensitive range of educational units), or maybe as high as 10 μV per division (in sensitive oscilloscopes). The insensitive end of the range will usually be around 50 V/cm but special units (for electrical supply authority use) provide for much higher levels. (Attenuator probes also enable high voltage signals to be investigated.)

The application needed from Y-amplifiers can, therefore, rise to 100 000 on the most sensitive range. In addition it is important that the gain be constant over the bandwidth of the signals being monitored.

Basic units provide amplifier response flat from dc to a megahertz or more. (Bandwidths are defined between points 3 dB down from maximum.) Magnetic-deflection display monitors will only reach 20 kHz whereas sophisticated high-quality instruments have bandwidths rising to 350 MHz. Sampling plug-ins provide bandwidths equivalent to dc to 1 GHz.

Vertical amplifier controls are usually grouped together on the front panel, as are time-base controls. Figure 8 shows the panel layout of a 50 MHz bandwidth amplifier. From this it may be seen that the vertical sensitivity is selected by a switch and that the y amplifier has a 'vernier' sensitivity control which must be at the 'calibrate' position when measurements of signal amplitude are being made.

The position of the trace on the screen depends upon the standing voltage applied to the plates. On both Y and X axes extra controls, (as shown in Fig. 6) enable vertical and horizontal shift of the trace position by adjustment of the bias applied. When using the CRO to probe circuits involving ac signals combined with standing dc levels – as is the case in ac amplifiers for example – the dc level on the Y signal causes the trace of the ac signal to be displaced vertically and, perhaps, to go right off the viewing area. This difficulty is overcome if you couple the circuit signal to the Y-amplifier via a capacitor. The ac signal then centres itself on the screen at the position chosen by setting the vertical shift control. This method is acceptable provided frequencies below the cut-off of the RC filter produced are not wanted. Measurement of very-low frequency to dc signals must be dc coupled on the ac/dc selector switch provided. A further switch position enables the input to the plates to be brought to its dc zero position. This helps the operator to establish where this level is on the screen. The switch for this function is in the middle of the unit shown in Fig. 6.

Signal Input Connections: Oscilloscopes for use with frequencies below about 1 MHz can make use of separate plug-in/screw-down banana-plug terminals. More usually, however, the input to the Y amplifier, and perhaps to the external trigger, will use standard BNC connections. These are designed for use with coaxial cable and coax should be used for all except the shortest end connections to the circuit. The input impedance characteristics are usually quoted – 1 megohm with 20-100 pF shunting capacitance being typical values. In some applications the CRO must be matched to reduce reflections – in such cases the input might be 50 Ω or 600 Ω. For fast rise-time studies it is necessary to ensure that the capacitive value presented does not reduce the overall bandwidth by shunting the device to which the CRO is connected. In exacting cases, needing high input impedance and small capacitance, special probes are used. These are described later.

Calibration of the Time Base and Y-Amplifier: The value of electronic components may drift with time, altering the sweep rate and vertical amplifier values from those indicated by the selector switch. To enable the operator to check these, more advanced oscilloscopes incorporate a special circuit that provides a fixed-frequency, fixed-amplitude square wave signal for calibration purposes. A typical signal would be 1 volt peak-to-peak. As it is derived from the mains frequency (50 Hz or 20 ms period) its time duration is also quite accurate.

MULTIPLE TRACE OSCILLOSCOPES

Measurement situations involving oscilloscopes more often than not require display of comparative information between two points in a system – the relative input and output signals in an amplifier response test, or the phase shift between two signals across a filter stage. Single-beam oscilloscopes are very limited because they cannot provide as much information to the user as a unit that can compare the waveforms at two points simultaneously. Three distinct alternatives are available to provide dual beam operation:

Separate gun: These use two, physically-separate, electron beams and deflection systems that are mounted inside the tube envelope. The beams may be generated by splitting the beam from a single gun. These are generally referred to as dual-beam units (dual-trace is a term reserved for the next method described).

Each beam has its own Y-input panel with a complete set of controls as described earlier. Dual-beam units drive both X-scans with a common set set of deflection plates (as in Fig. 9) but some (rather rare) oscilloscopes enable each time-base to scan at a different rate.

In general, dual-beam units are less common because of the higher expense compared with the next method.

Electronic switching – chopped mode: The deflection response of an electron beam is rapid enough to allow it to be directed from one position to another at a speed exceeding the scan rates used with the signal being viewed. Fast electronic switches are used to switch the common single beam between two (or three or four) Y-inputs. Figure 10 illustrates this. Appropriate blanking (that is reduced Z intensity) is applied when needed, when the beam is chopping from one trace to the other. If the chopping rate is chosen to be at least 100 times faster than the highest frequency to be viewed the two traces appear as separate traces. Hence the name "dual-trace" for this method. In reality the traces are not continuous but are made up of dash-spaces. A hundred dashes across a screen produces a virtually continuous trace to the eye. The limit of usefulness is reached when the inbuilt chopping rate comes close to the upper frequency being viewed thus producing a dashed-line trace in which the dashes are of length equal to wanted signal features. At this point information is lost.

As far as the user is concerned there are still two groups of Y controls – just the same as for a dual-beam arrangement. The difference arises as

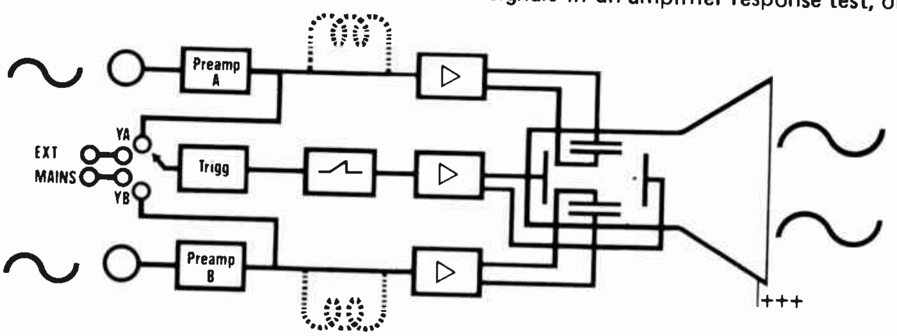


Fig. 9. Schematic of Philips 3232 dual beam oscilloscope. Common x plates provide scan for both beams, separate y plates deflect the two distinctly separate electron beams that are derived from a common gun.

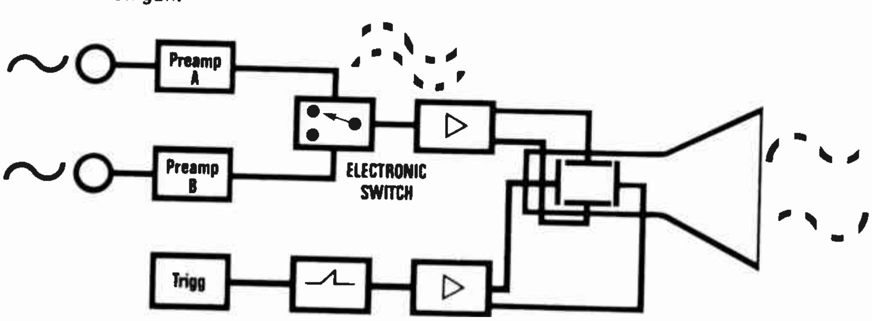


Fig. 10. Electronic-switching enables a single-beam and deflection system to provide dual-trace operation.

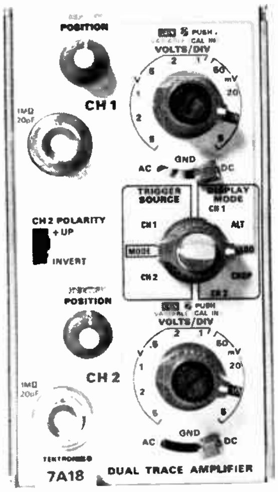


Fig. 11. Example of dual-trace plug-in showing the basic controls.

the position chosen on the selector switch where a 'chop' mode must be selected — see the panel of a dual-trace plug-in shown in Fig. 11.

Chopped operation ensures that the time relationship between the two signals is faithfully presented: phase measurements are also accurate (that is, providing the input amplifiers to each are identical).

Chopped operation will also display two simultaneous, non-recurrent signals, such as transients induced at various points when a complex resonant system is excited by an impulse. It is quite suitable for low-frequency signals but less convenient as the frequency rises.

ELECTRONIC SWITCHING-ALTERNATE MODE

Switching can also be employed on a full alternate trace-by-trace basis. The first trace is a scan of channel 1, the next of channel 2 and so on. This does not suffer from the dotted defect with high-frequency viewing but it suffers from another deficiency in that the phase relationship between the two signals may not necessarily be as indicated on the screen.

The method is unusable for observation of "once-only" dual events because the second transient signal may have gone to zero by the end of the trace of the first simultaneous transient signal. The panel shown in Fig. 11 is typical of dual trace units. The selector switch enables choice of alternate, chop, channel 1, channel 2, and channel 1 plus channel 2 modes.

With two channel operation it is necessary to decide which input will synchronize the time-base scan. A switch provides the choice of appropriate internal triggering. Although only channel 2, for example, may be being viewed there are

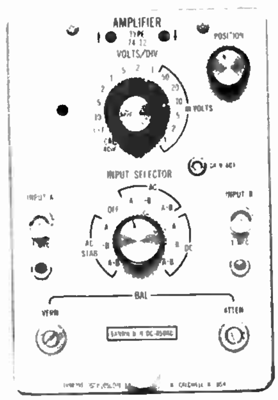


Fig. 12. Differential amplifier plug-in panel.

circumstances where it is desirable to trigger from the channel 1 signal.

The electronic-switching method enables more than two traces to be displayed— three and four-trace units are available.

DIFFERENTIAL AMPLIFIERS

Generally the dual-trace oscilloscope is recognised by two sets of input terminals. There is, however, another two-input unit that is for single trace operation. This is the differential input amplifier unit; it is normally provided as an optional plug-in.

Two inputs are amplified by the high-gain differential arrangement of a dc amplifier — refer to Part II for the basic concept. These are used when common-mode noise rejection is needed and when the difference between two fully floating inputs must be studied. Figure 12 is the panel of a high-gain differential amplifier. Position and gain controls are used as for an asymmetric input Y-amplifier. The input selector provides choice of Channel A — channel B, A-B; each of these being ac or dc connected. The balance controls enable the two amplifiers to be trimmed to remove offset and gain differences.

FINDING THE TRACE

Even experts can experience temporary difficulty when confronted with an unfamiliar oscilloscope — especially when it is complicated. Naturally it takes training to get the best from a unit. A basic difficulty is often finding the trace! These steps provide an efficient procedure that

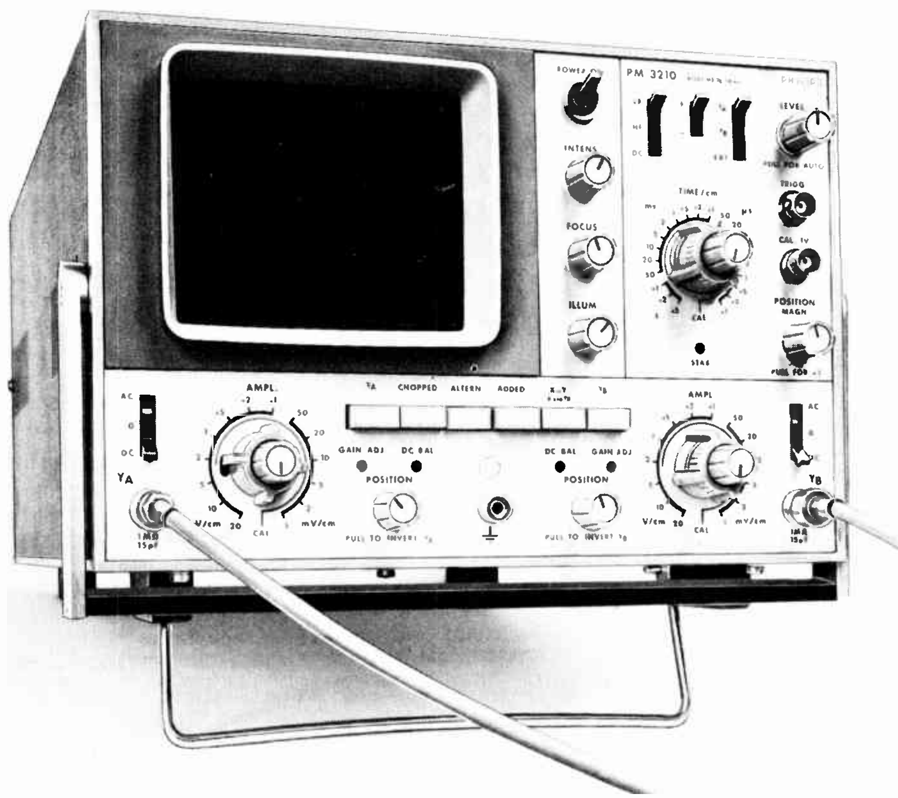


Fig. 13. Front panel of an oscilloscope that represents medium sophistication. The text contains a self test about its controls.

should be learned. Begin with the input to the Y-amplifier unconnected.

1. Ensure that the power is on. The on-off switch control is usually built in with the intensity knob, but not always.

2. Turn the intensity to 75 per cent clockwise.

3. Switch the time-base (horizontal) to a medium speed – say 1ms/cm. This ensures that the trace displays a full line across the screen rather than a point which occurs when the scan speed is on the slow ranges.

4. Switch to auto triggering. This may be a marked position of the trigger control or a separate switch. This ensures that the trace is being triggered.

5. Switch to internal triggering. This is necessary for (4) above. Relying on an external signal to trigger the scan is unreliable – it may not be of adequate magnitude.

6. With this done slowly vary the vertical position control about its mid range point widening out to get the trace on screen.

7. The above may still not produce the trace. If not put the vertical position in its middle point and the gain at an insensitive value and begin a scan of the x-position control. This should be somewhere mid range. Too much x-shift can cause the trace to slide off screen.

Complicated oscilloscopes will invariably incorporate a variety of controls that may also need adjustment to find the trace. Space prevents a full guide to spot finding. Fortunately the more expensive units often provide a spot-finder button. Press it and the spot appears on screen enabling the controls to be adjusted accordingly to bring it back from the direction it flies to when the button is released.

When the trace is located in mid screen the intensity and focus are then adjusted by switching the scan to the slowest rate to produce a spot. These should be adjusted to produce a small round spot without halo: stationary spots on screens should be avoided as this shortens the life of the phosphor at that point.

BASIC MEASUREMENTS USING AN OSCILLOSCOPE

Voltage amplitudes of a waveform: The amplitude vernier control is set to the largest position. The range selector is set to obtain the largest practical vertical size of the waveform. Set the waveform with the lower value on a convenient graticule line using the vertical position control. Read off the number of graticule divisions overlapped by the waveshape, to the nearest estimated tenth. Multiply the number by the range setting value to get the peak to peak voltage. Example . . . 7.8 divisions for range setting of 20 mV/cm = 156 mV pp amplitude.

Voltage level of a dc signal: Connect the two probe leads together. Set the horizontal scanned line, created with a suitable time base setting, to a lower graticule line using the vertical position control. Set to calibrate position, as above. (Advanced units have a control that shorts the probes internally). Attach the probes to the circuit. The dc level is given by the number of graticule divisions above or below the line used as the zero defined by the first step.

To read the dc level of an ac signal superimposed on a dc signal, carry out a similar process but use the centre of the ac waveform as the vertical reference point.

Currents: As they stand, few CRO units can measure currents. They are built to measure voltages so some conversion method is needed. Two methods are:–

(a) Addition of a small value series resistor. The circuit is broken and a resistor added in series. The voltage across this gives the current using Ohms Law. It is, however, not always possible to use this method because an oscilloscope is usually grounded at one terminal meaning that one end of the resistor must also be at earth potential to preserve proper operation. If that is not possible then a differential input CRO is needed.

An alternative, but not very accurate method is to probe the voltage at each end of the resistor taking the difference of the two voltages obtained.

(b) Current probe. Special probes are available that clip over the circuit line without the need to break the line. The current passing through induces a voltage that feeds the CRO. For 'ac only' measurements a transformer device suffices but for dc measurements it is necessary to use something more sophisticated, such as a Hall effect probe. All probes have the conversion constant clearly marked on them.

Waveshape: Controls are adjusted to obtain the largest possible waveshape on both time and amplitude scales. It is sometimes easier, however, to use a slower time base value to obtain more stable triggering. The shape is then transferred to a hard copy form by use of a camera, by longhand, copying using the graticule lines as indicators, or by tracing the waveshape directly on the screen.

Frequency: Again the waveshape is set to have the largest size with stable triggering. The horizontal distance from peak to adjacent peak is a measure of period. Frequency is calculated from this by dividing the period into unity.

Measurements made with the CRO are only of low order. The linearity of scanning and the estimation of traces, with the finite width trace, limit reading accuracy to around 5-10%. The CRO is not the tool to use for accurate measurements.

SELF TEST

If the above facilities are understood it should be a straightforward matter to master any oscilloscope with the degree of sophistication that has been introduced so far. Figure 13 is the front view of a modern unit. It is given as an exercise.

1. Locate the controls that are associated with the quality of the dot produced by the beam.

2. Does it provide two trace display, and if so by what method?

3. Which is the time-base control area?

4. Where is a Y-amplifier vernier control to be found?

5. What are the input terminating conditions?

34 Oscilloscopes – the refinements

MANY MEASUREMENTS IN electronics can be handled by the relatively unsophisticated oscilloscopes described in the last part of this series. More capability can be provided at greater cost and this can be valuable if the user understands how to make the most of it. This part describes refinements that will be encountered in more advanced oscilloscopes.

IMAGE STORAGE

Screen persistence: Repetitive signals, such as a sinewave signal, can be made to repeat on the screen overlapping the previous trace produced. If the time-base frequency is sufficiently high – from thirty or forty hertz upward – the screen provides an apparently stationary signal of constant and adequate intensity. This is primarily because the eye cannot detect individual scans (as in motion pictures and television) and secondly because the phosphor, at frequencies above a few hundred hertz, re-energizes before its light emission due to the previous scan, has decayed away.

Phosphors with large time-constants are available (such as P2, which takes one second to reduce to 10% of original brightness and P7 which takes three seconds) and oscilloscopes have been manufactured which use these to enable signals of less than one hertz to be studied. This feature, however, largely restricts the use of the instrument to low frequency work because medium and high-frequency signals that are not well synchronised will produce separate traces which remain and add up with time to produce an unclear picture. This method of studying slow-transient phenomena has not been developed to any great degree because of this and other factors (such as poor resistance to burning). In addition the retained-image times are still inadequate for many applications.

CAMERAS

Storage requirements fall into two classes – those where the transient is unique and therefore needs to be recorded only long enough to allow the trace to be studied and those where a permanent record is needed.

The oscilloscope fulfils both these needs. Until the advent of the Polaroid-Land process this involved a time-consuming development process before the operator was certain of having even recorded the trace. Most oscilloscope makers now offer specially built trace-recording cameras that fasten onto the large bezel surrounding the screen.

Such cameras use a Polaroid-Land film pack of some kind and often incorporate a 35 mm roll film facility also. A Dumont unit is shown in Fig.1. The user sets the CRO controls until satisfied that the trace will be as needed. This is done using the viewing aperture which reflects the screen image to the observer via a mirror. It is essential that the camera has the correct focal distance set for the CRO concerned, so in general cameras relate to specific units. Some models incorporate adjustable object-image ratios; a few are fixed ratio. With experience it is even possible to capture multiple trace events (by multiple exposure) for comparison purposes.

A considerable amount of film and patience can be consumed trying to record once-only events. Cameras can be quite expensive – several hundred dollars – but they do provide a permanent record for reports which no other storage system can provide, and the price of a camera is not as great as the extra cost of the variable – persistence storage units to be discussed later.

STORAGE OSCILLOSCOPE

Most of the objections to the above storage methods, with the exception of permanent photographic reproduction, are overcome by using an advanced form of the basic CRO tube. It is called a variable persistence storage tube and is a development of early 1950's more basic storage tubes in which the waveform could only be held at a constant intensity (without the feature of gradual fade out). In fact variable persistence is a feature of tube operating circuitry not the tube itself.

The construction of a typical storage tube is given in Fig.2. The phosphor viewing screen (having 0.1 s persistence time from P31 material) and the writing electron gun shown are similar to those used in the simple cathode ray tube. Additional components are the flooding electron gun system, a storage mesh which is coated

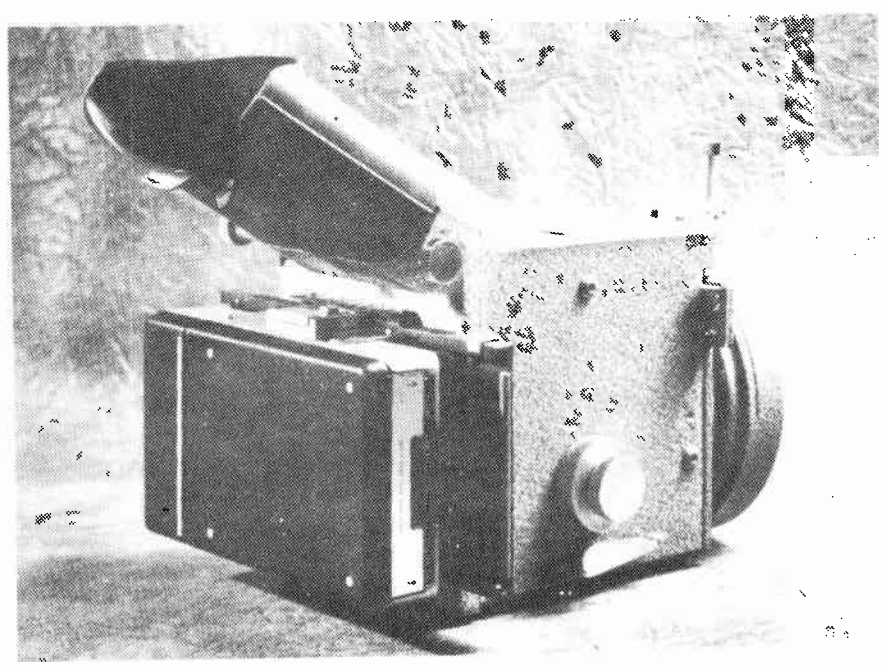


Fig. 1. Recording camera using Polaroid film pack.

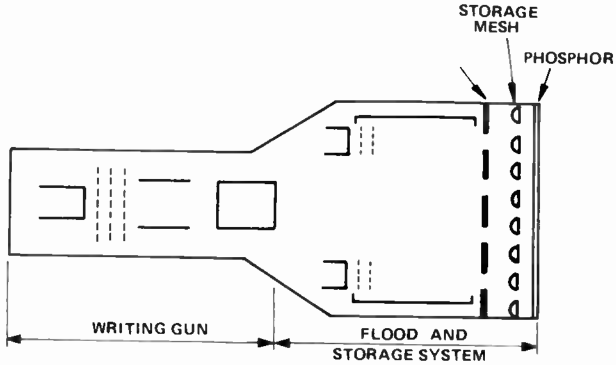


Fig. 2. Schematic construction of storage cathode ray tube.

with a non-conducting, highly-resistive material such as magnesium fluoride, and a collector mesh which is held at a positive potential.

To store a trace the writing gun is scanned over the storage surface. Where the beam strikes the storage mesh electrons are knocked loose leaving a positive-charge pattern. The high-resistivity of the surface prevents the charges moving toward a neutral state: the scan is thus stored – and can be held for at least an hour (one maker offers four hours) in a reduced intensity mode.

To make the trace visible, low velocity electrons are sprayed by the flood guns onto the entire mesh surface. These electrons are allowed to pass through to the phosphor in proportion to the amount of positive charge at each aperture of the storage mesh. The positive field pulls many electrons through causing them to pass on to hit the phosphor.

The collector mesh is provided to help accelerate the flood electrons; to repel the positive ions generated by the flood guns; (which would otherwise write the whole screen bright) and to absorb the emitted secondary-emission electrons produced whilst writing is in operation. It is not possible to store the trace in the view mode for as long as in the store mode: one to ten minutes of viewing time is typical for various makers' designs.

Erasure is done by applying a large positive voltage to the storage mesh which charges capacitively to the same value. The mesh voltage is then brought back to a small positive value whereupon the flood guns reduce the voltage to zero. A small sudden negative excursion is finally applied to the mesh making it ready to write. (This procedure is automatically initiated at the single action of a switch.)

Variable persistence is incorporated by changing the time taken to erase the picture. In the Hewlett-Packard unit, shown in Fig.3, this is achieved by using a variable-width pulse generator that applies erase voltage pulses to the storage mesh. The positive-ions created by the flood-guns limit this mode to a maximum of 10 minutes persistence.

Storage oscilloscopes can be used as conventional units by applying about 30 volts to the storage and collector meshes. Long persistence has many virtues – it enables successive traces resulting from adjustments to a system response to be overlaid together for comparison purposes. It also allows us to see very low-frequency scans, and to plot scans of spectrum analysers. Long persistence also finds use in time-domain reflectometry where the time between send and receive pulse needs measuring.

By stacking sweeps on top of each other a long persistence time can be used to integrate or average a set of traces. Variable-persistence storage oscilloscopes are extremely versatile but the high price restricts their use to large laboratory groups.

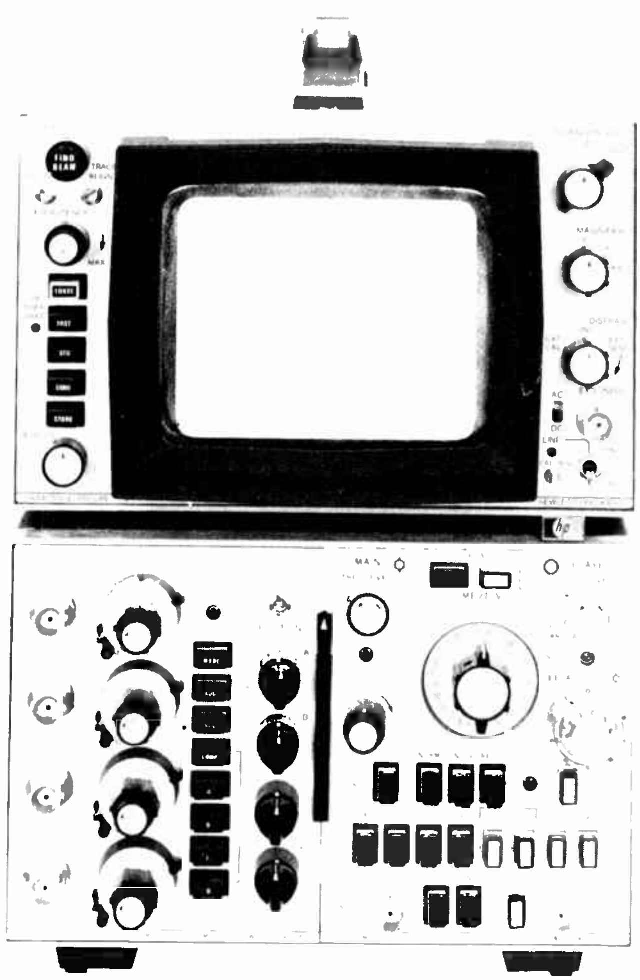


Fig. 3. Hewlett Packard Model 184A variable persistence and storage oscilloscope. Controls for storage are on the left-hand side of the screen.

STORAGE USING DIGITAL MEMORY

Figure 4 shows a unit marketed around 1972. The transient recorder unit accepts the analogue signal, converts it to a digital equivalent with respect to time and stores the values in digital registers. Readout can be obtained by using digital-to-analogue conversion of the stored increments which are scanned sequentially, the resultant analogue voltage being fed to an oscilloscope or chart recorder. Digital print-out is taken direct from the scanned store locations.

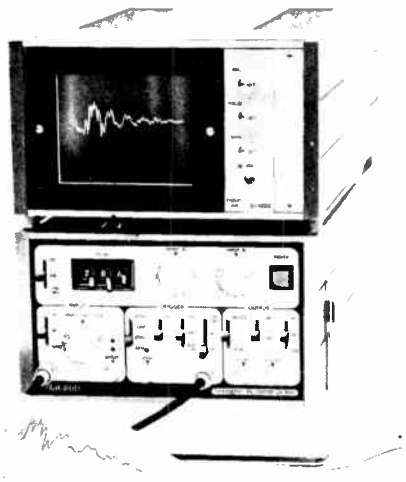


Fig. 4. Storage can also be obtained using a digital memory to capture the event which can be displayed at will on an oscilloscope or recorder (Datalab DL 905).

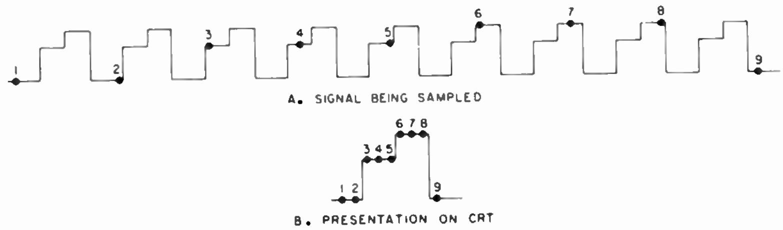


Fig. 5. The sampling oscilloscope builds up the waveform on the screen from sampled values taken from the original.

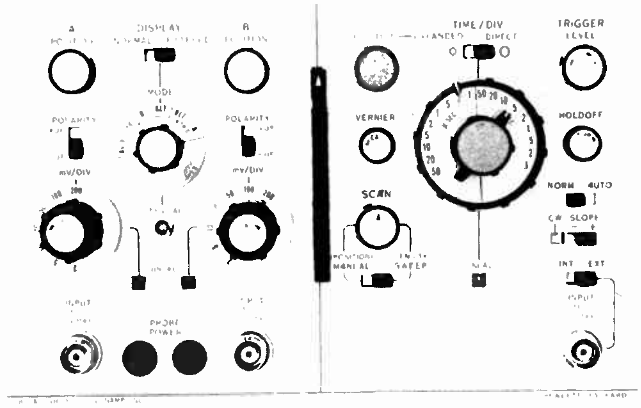


Fig. 6. 1 GHz dual-channel sampling plug-in. (Hewlett Packard Model 1810A). The controls are designed to provide operation as for those on normal real-time oscilloscopes. With such a unit it is possible to view nanosecond rise time signals of repetitive nature.

This method is less common than the storage oscilloscope alternative but the ever-reducing cost of digital methods may put this technique into a competitive price region.

Another method of capturing difficult to see, once-only transient signals, and very slowly changing waveforms is to record the level of the signal, increment by increment, as the signal occurs, using a digital memory. The concept is simple and the method offers certain advantages. These include ability to speed up or slow down the timescale of the original event, ease of providing a permanent numberical printout and the facility to process the signal before display.

SAMPLING OSCILLOSCOPES

How to capture a very fast repetitive event, say near to the GHz region where scan times of 0.1 ns/division are needed, is a problem because the electron beam cannot transfer enough energy into the phosphor to obtain a useable trace brilliance. Further it becomes increasingly difficult to deflect the beam at such speeds. The sampling oscilloscope offers a solution to these problems.

The sampling oscilloscope makes use of the stroboscope concept to look at a waveform, which must therefore be repetitive (as shown in Fig.5). The beam is set to illuminate the screen at point 1 in the diagram, waiting there until the next cycle where it moves to point 2 – and so on. The trace therefore gradually works its way through the complete cyclic waveform and because the scan speed is slower than with a conventional sweep system the cathode-ray tube system can operate with a lower bandwidth than the signal. The waveform produced is an average of many so the display is not only sharper but more uniform. (This may be a disadvantage in some applications for the sampling unit is effectively smoothing the unknown true original signal). Sample and hold methods were discussed in the previous part discussing D-A and A-D conversion.

In practice a sampling oscilloscope is a normal high quality scope which can accept a sampling plug-in. Figure 6 is the panel of a dual sampling unit.

DELAY FACILITIES

Often one needs to study a certain part of a repetitive waveform – the very beginning, for instance. An example is the ringing of a non-ideal square wave shown in Fig.7a. The trace is triggered, to begin the sweep, by a fast-going edge. Due to circuit response-times, the trace does not

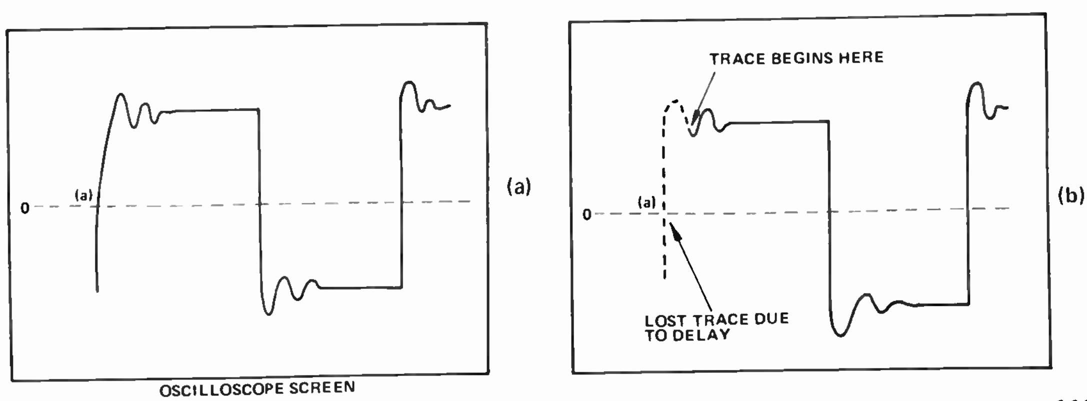


Fig. 7. Inherent trigger delay, if not compensated for, will lose the leading edge of a waveform.

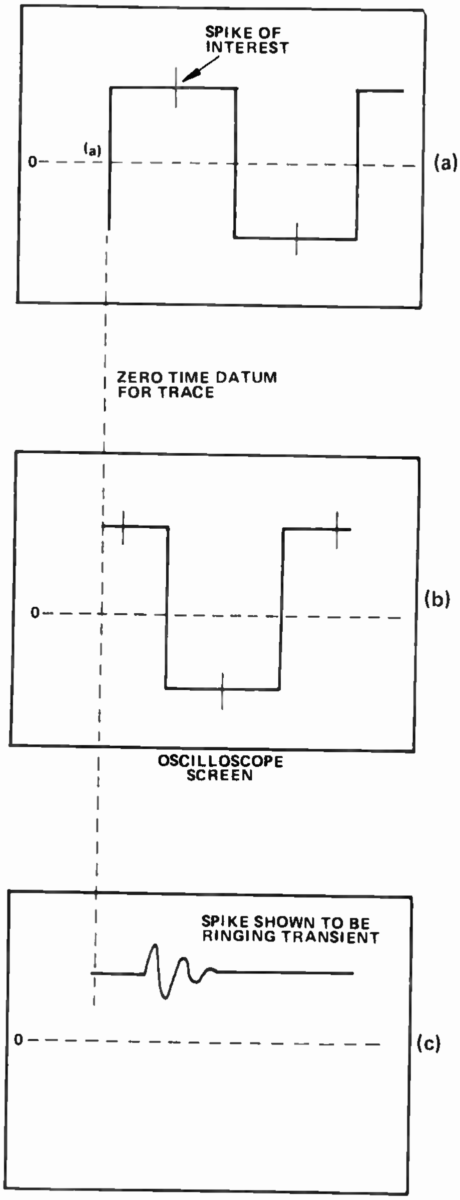


Fig. 8. Use of introduced delay in triggering to enable an event away from trigger transient to be investigated.
(a) Original spike on pedestal of square wave.
(b) Delay introduced to bring spike back to time origin.
(c) Scale expanded to reveal true nature of spike.

begin to sweep at exactly that time but begins a little later. The result is loss of the leading edge region of the wave as shown in Fig.7b. The following waveform may provide the information sought but attempts to widen the waveform in the horizontal direction lead to the second front disappearing. The simplest solution to this problem is to incorporate an appropriate fixed delay into the triggering circuits and this is often provided within the circuits. A slightly better method is to provide an adjustable delay control on the trigger panel.

A more difficult problem is capturing a point on the signal train that is remote from the triggering transient. Consider the signal shown in Fig.8(a), where the problem is to investigate the spike transient on the pedestals of the square wave. Triggering is best achieved by using the edge (a). But this means that scale expansion puts the spike off scale when the horizontal expansion scale is great enough to provide information about the spike structure.

Variable delayed sweep is the answer. The trigger circuit is set by the (a) edge but trace scan does not begin until after a period, as in 8(b). Thus the trace captures the spike at the left-hand side of the screen and scale expansion will now be possible as in 8(c).

To make this workable in practice the operator must know just where triggering occurs for there may be several somewhat similar events along the trace. It is vital to know which one is being viewed. A refinement provided in variable delay circuits is to brighten the original display from the point where triggering will begin. Taking the idea one step further leads to a second delay that effectively decides where the trace stops Figure 9 shows the waveform brightened to show the portion that will be expanded and the second trace of the dual-beam unit is used to show the expanded part. Another useful feature

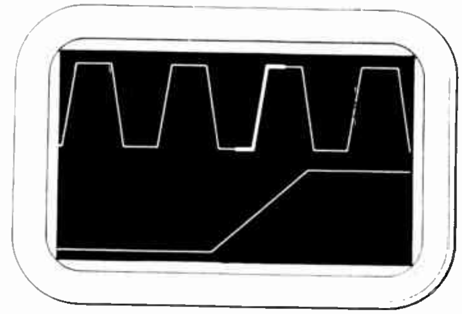
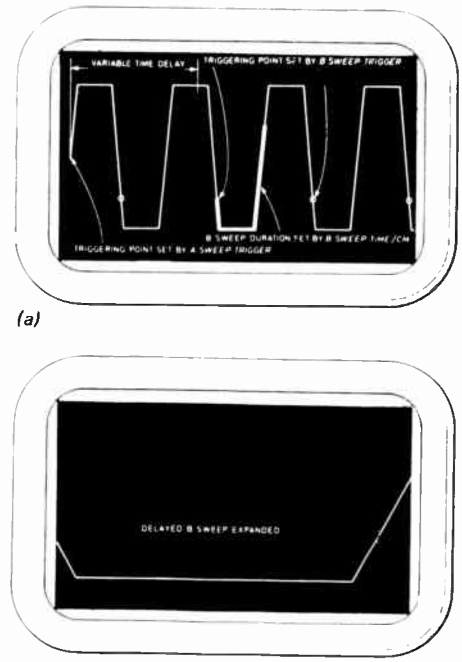


Fig. 9. Trace brightening is used to show which part of the waveform is to be expanded. In this display the expanded portion is also displayed on the second trace of the CRO.



(a)

(b)

Fig. 10. Use of dual delayed triggering point. (a) original (b) expanded.

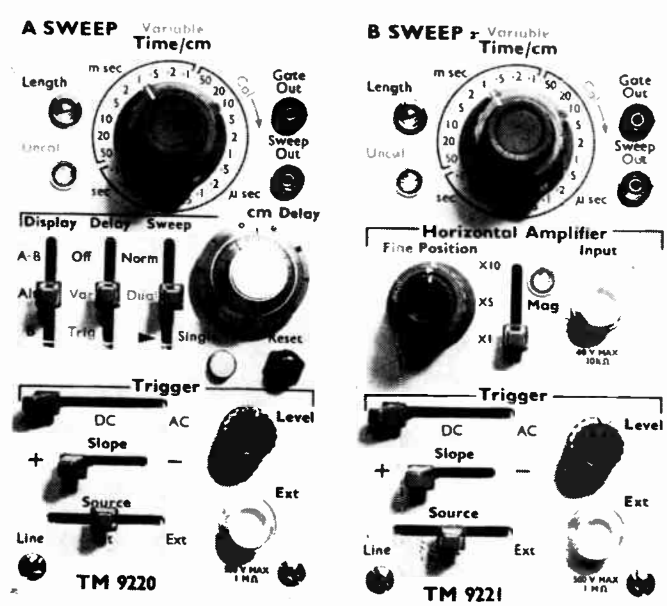


Fig. 11. Marconi TM 9220, TM 9221 sweep units.

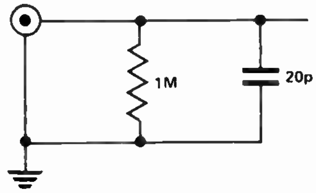


Fig. 12. Most oscilloscopes have this input equivalent circuit. Although the values seem insignificant, at high frequencies they become dominant requiring the use of special probes.

is to be able to use a trigger point not on the origin of the first trace set up — as in Fig.10. Here a marker dot is provided to help the operator. Figure 11 shows the panels of Marconi plug-ins which provide these and other variable delay features.

PROBES

Passive probes for voltage measurement: In part 32 the importance of providing the right matching conditions between two electronic systems was stressed. This is also important when connecting an oscilloscope to a circuit, for each output and input has certain resistive and reactive conditions which must be properly combined to get realistic signal transfer.

The oscilloscope can be represented as an ideal termination shunted by a large R and an adequately small C value — or at least they appear this way at first sight. Figure 12 is the most common approximate equivalent circuit. (Others used include 50 ohms with negligible reactance in certain applications). Referring to the chart in Fig. 13, it can be seen that with 20 pF at 10 mHz the circuit being measured must have an equivalent output resistance of no more than 8 ohms!

For high frequencies, those above 100 kHz say, we therefore need a better connection method. To further compound the problem the oscilloscope input leads can easily increase the equivalent C value to 100 pF — leads for 1:1 connection must therefore be carefully designed to ensure known loading conditions which can be allowed for in signal measurement corrections. It is very bad practice to use any piece of coaxial cable and connector for frequencies beyond 100 kHz.

The first improvement is to use a probe which has 10:1 attenuation built in, for these are designed to have a lower effective cable capacitance — see Fig.14(a). Still better is a special correction arrangement that balances the shunt against series capacitance to provide a wider bandwidth — see Fig.14(b). By the use of inductive tuning a further improvement in bandwidth can be obtained — Fig.14(c).

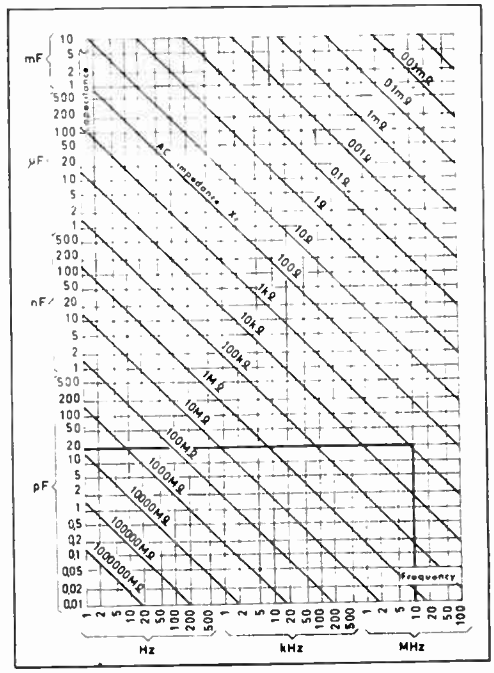


Fig. 13. Chart for obtaining reactance of capacitors at various frequencies of operation.

Probes with division ratio of 100:1 also are manufactured — these can provide equivalent termination conditions of 5K/0.7 pF, 10M/1.8 pF, 1 M/1 pF. The reason for different pair combinations arises from the need to alter the trade-offs between rise time and signal loss in high-frequency and very fast transient measurements.

There is no easy answer to the question of which attenuator probe to use. These guides are the start. For

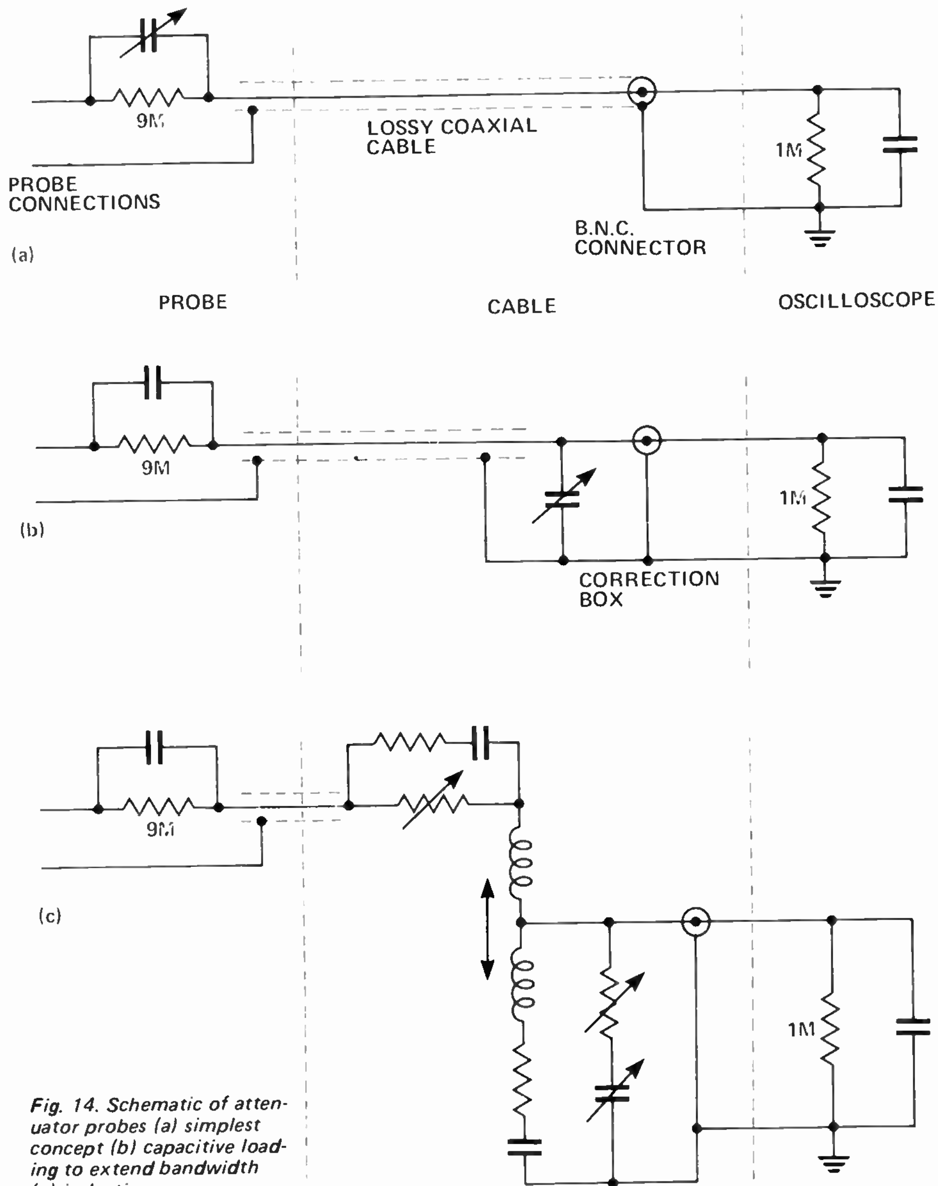


Fig. 14. Schematic of attenuator probes (a) simplest concept (b) capacitive loading to extend bandwidth (c) inductive.

amplitude measurements select a minimum-impedance source point to measure from. The best probe to use here is one with the highest impedance at the frequency of interest. Capacitance is less important here than resistance for it alters edge shapes, not amplitude.

For fast risetime measurements again select a low impedance source point and use a probe with lowest effective capacitance – signal attenuation is less important than transient edge shape changes.

ACTIVE PROBES FOR VOLTAGE MEASUREMENT

The above probes make use of passive matching arrangements. But for the extremes of frequency and/or risetime measurements the values of components required in passive probes become impractical. However active amplifiers interposed between the circuit and the oscilloscope can be used to improve performance by increasing input resistance and lowering capacitance (short loads). FET probes are marketed to meet this. Figure 15 is a Phillips PM 9354 FET probe which can be used for dc to 1 GHz measurements. These need an additional power supply to operate, and problems of dc drift (0.5mV/^{o}C) and added amplifier noise (60 μV – 1.5 mV) may be disadvantages in certain applications. As well as being the choice for very high-frequency work, FET probes also find useful application in low-frequency situations where measurements across high source impedances are needed.

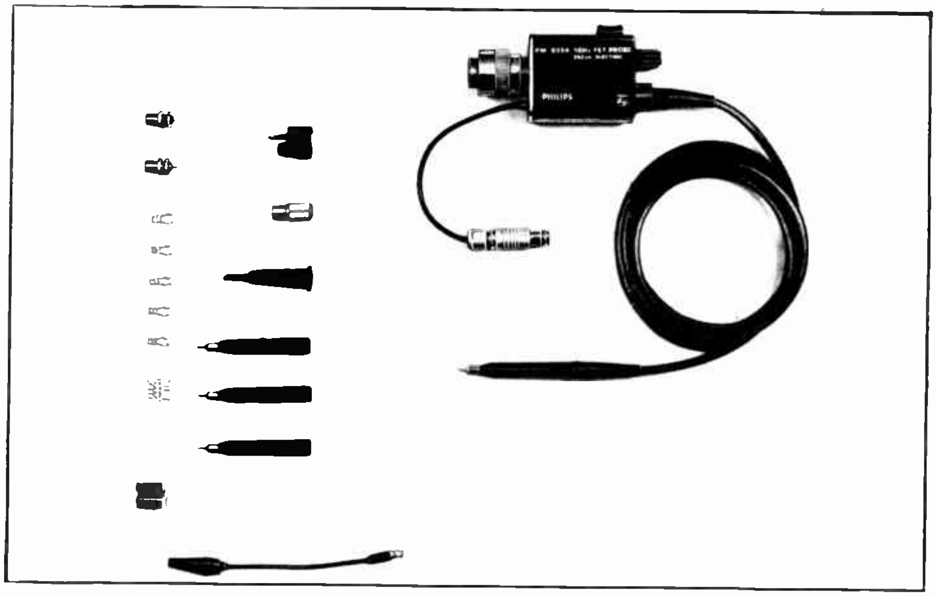


Fig. 15. This FET probe uses active coupling to overcome oscilloscope connection problems.

OTHER PROBES

Voltage measurements are by far the most frequent measurements made but in some instances it may not be possible to determine voltages, and current measurement is used instead. An example is the current flowing in a direct-coupled Darlington pair configuration where no significant resistance exists over which a voltage can be developed. DC current probes (see Fig.16) clip over the wire in question coupling the dc magnetic field created by the current flowing in the wire into a Hall effect transducer which generates a voltage equivalent to the current flowing. These will also measure ac currents. The maker specifies the conversion constant – typically 1 mV/mA. AC only, current probes are also made using a current-transformer principle.

Probes for use in digital circuits are also available. These may incorporate a logic gate that combines the outputs from up to 6 circuit points as shown in Fig.17. Power for the gate is obtained from the circuit under test.

Fig. 16. D.C. current probe.

SPECIAL PLUG-INS

The oscilloscope, due to its extensive flexibility, can form a major part of many test systems, thereby reducing the overall price of advanced measurement systems where a suitable CRO already is available. Special plug-ins are offered (to suit certain mainframes) that will convert an oscilloscope into a spectrum analyser or into a semiconductor characteristic-curve tracer. Another plug-in is offered that converts the CRO into a four-trace unit.

A basic need in manual measurement is the provision of output form that best suits the operator. In many tasks a visual output in the form of a picture or graph is better than having

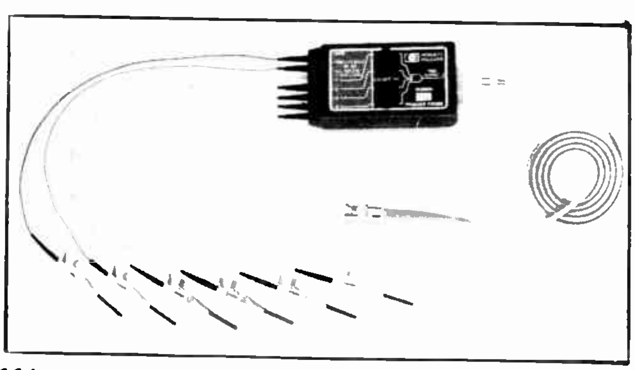


Fig. 17. Digital circuit logic probe.

Fig. 18. Display monitors can now provide an exhaustive arrangement of visual output forms. The next step is to provide this capability for routine use in oscilloscopes by making plug-ins available to go into suitable main frames.

to view many traces of a time sequence taken over the whole system. Already we have logic analysers which display space-plane information on the CRO screen, multi-meter CRO units that write digital values on the screen and units that provide axes information on screen graphs. With the reducing cost of advanced processing it will not be long before the microprocessor and memory (already in use in very sophisticated units) are introduced into quite moderately priced oscilloscopes for converting the information taken from the circuit into better forms of display. Display monitors are already available with many display forms — see Fig.18. The next stage must be the marrying of the basic CRO unit to such capability via a wider range of sophisticated plug-ins. The colour oscilloscope will also soon be with us extending the information rate at which the operator can be informed about a system via a CRO.

The only weak link in present systems (as far as robustness, life and cost is concerned) is the CRT itself for it is just about the last remnant of thermionic device technology remaining in general use. This too will soon be replaced by a solid-state equivalent. Perhaps this will take the form of a matrix of three-colour, LEDs in a flat display — making maximum use of the low-cost production advantages of LSI techniques.

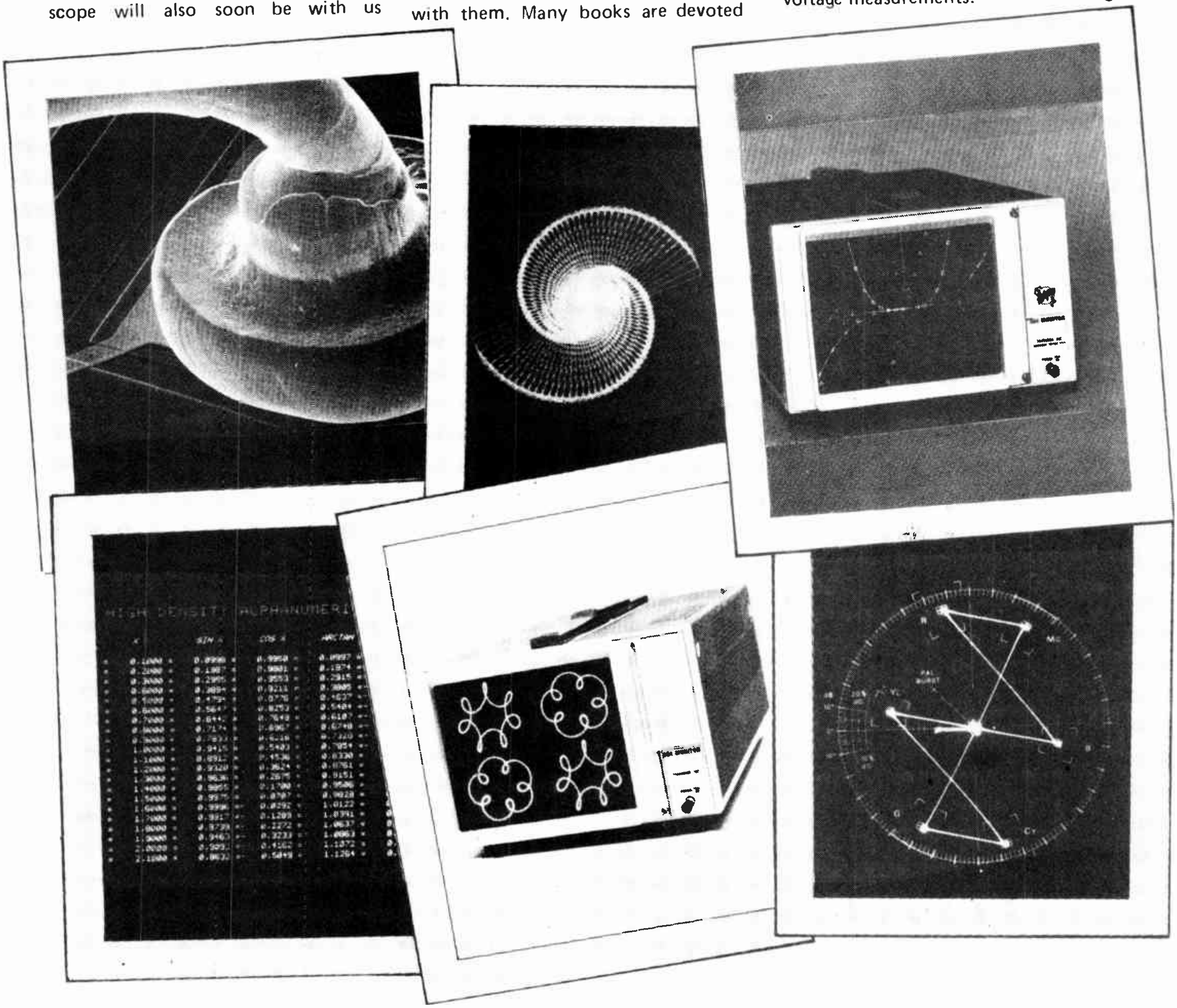
REFERENCES

Due to the versatility of the oscilloscope most books on electronic instrumentation include basic descriptions of how oscilloscopes work and how to perform basic measurements with them. Many books are devoted entirely to the oscilloscope.

General considerations are discussed in "Test and measuring instruments — 1974 Catalogue", (Philips). Tektronix, Hewlett-Packard, Dumont and Marconi outlets also provide basic articles on the selection and use of oscilloscopes.

The principles of storage tubes are explained in the Philips catalogue and in "Variable persistence increases oscilloscope's versatility", R.H. Kolar, Electronics, November 29, 1965, 66-70.

The subject of correct probe choice is quite extensive. The Philips Oscilloscope series referred to above has two articles by PFW Zwart on probes (Nos. 3 and 4 of series). V. Bunze in "Matching oscilloscope and probe for better measurements" Electronics, March 1, 1973, p 88-93 discussed voltage measurements. ●



35 Chart recorders

IN GENERAL, chart recorders are designed to accept electrical voltage signals as these constitute the majority of signals produced by sensing equipment. Occasionally the chart recorder is more appropriately connected to a mechanical output without electrical signals being involved: in some circumstances there is no need for electrical circuitry.

Chart recorders are, therefore, electronic system units which accept a voltage signal converting it to an equivalent graphical representation on paper. The recorder can be put to use in any application where an electrical signal is produced. Examples are measurement of fluctuations of the power mains voltage, records of body currents in medical diagnosis and changes in temperature in a process plant. The earliest chart recorder was probably Lord Kelvin's 19th century paper-tape siphon-recorder used to record electric telegraph signals. Because of the large and varied demand for chart recorders, manufacturers have developed numerous alternatives. Figure 1 shows a number of recorders installed to monitor an oil rig.

In fundamental terms chart-recorders are electro-mechanical converters — electrical signals are changed into equivalent mechanical ones which are used to make a permanent record on a paper-chart. For this reason there are two aspects to a chart recorder — its mechanical design and its electrical design. For convenience we look at each more or less separately but in designing and operating the recorder the two are so closely related that the response depends on adjustment of both disciplines of thought.

Fig. 1. Chart recorders are used in many varied applications. The panels of this control room contain a number that are used by the operators to see how the process is behaving.

Chart Recorder Formats:— Chart recorders are designed to display a signal in a graphical form that is convenient to the user. There are two basic types: those which record one or more variables with respect to time (commonly called x-t recorders) and those which plot one variable against the other (x-y recorders).

Strip-chart:— In these recorders a continuous roll of suitably scaled paper is motor driven at constant speed past the marking head. The paper drive is usually driven by a synchronous or stepping motor as this ensures accurate paper-speed. Where mains supply is not available dc governed-motors and clockwork alternatives can be used. Chart speed changes are commonly obtained by altering gear ratios. Figure 2 shows the construction of a typical panel mounted strip-chart x-t recorder. The module shown withdrawn from the housing is the paper drive unit, the housing contains the electronic amplifier driving the pen which contacts the top of the paper when the drive unit is plugged in.

Strip chart recorders designed for bench top use are also common — Fig. 3. Some strip chart recorders take up the used paper by rolling it or by folding it in a concertina. The latter, known as z-fold, is very convenient when the need to refer to the record arises. Chart speeds vary widely — from metres per second in fast-writing recorders used to capture kilohertz bandwidth transients, down to millimetres per hour for industrial process and slow-scientific phenomenon recording. It is not usual, however, to find a range as wide as this in the one unit.

Process industry strip-chart recorders generally run at one speed only; units for scientific use usually have switched speed capability. The choice is decided by matching the resolution required with the amount of paper consumed.

Paper sheet:— The flat-bed style lends itself to x-y operations where the

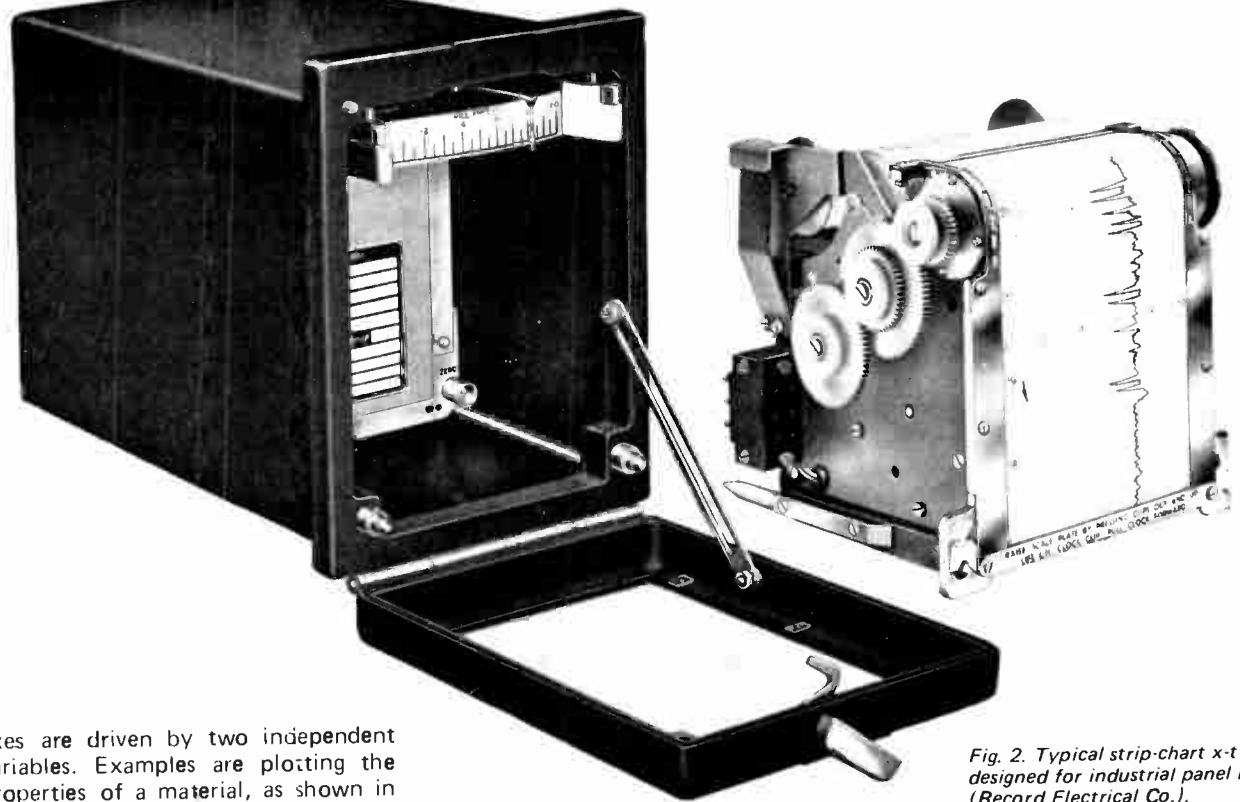


Fig. 2. Typical strip-chart x-t recorder designed for industrial panel mounting. (Record Electrical Co.).

axes are driven by two independent variables. Examples are plotting the properties of a material, as shown in Fig. 4, and charting antenna field strength versus position. In this style the recording paper is a single sheet which is attached to the platen. The pen moves both in the x and y directions. The paper may be held by clips or by electrostatic attraction. If the x axis input (horizontal) is fed with voltage that rises linearly with time (a ramp function) the x axis will move across the chart with time

Fig. 3. Flat-bed strip-chart recorder (Philips).

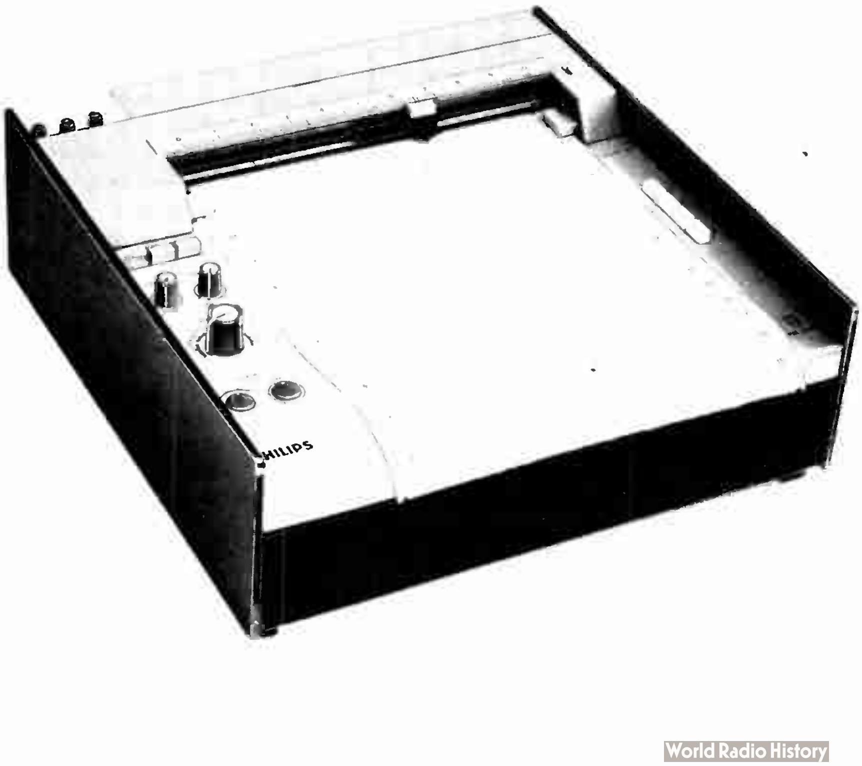
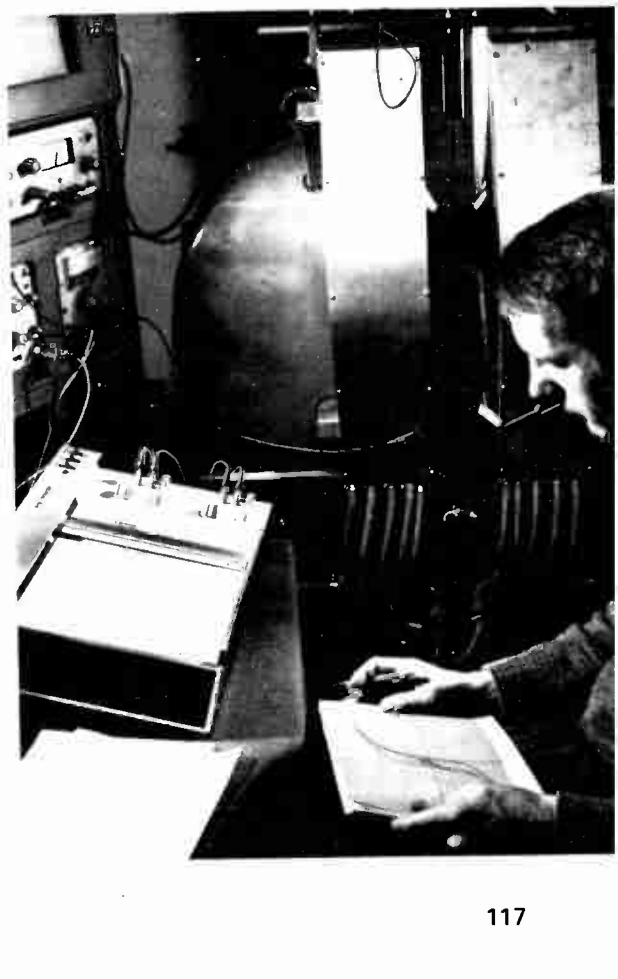


Fig. 4. Plotting an hysteresis curve for material under test in the large magnet shown at the rear.



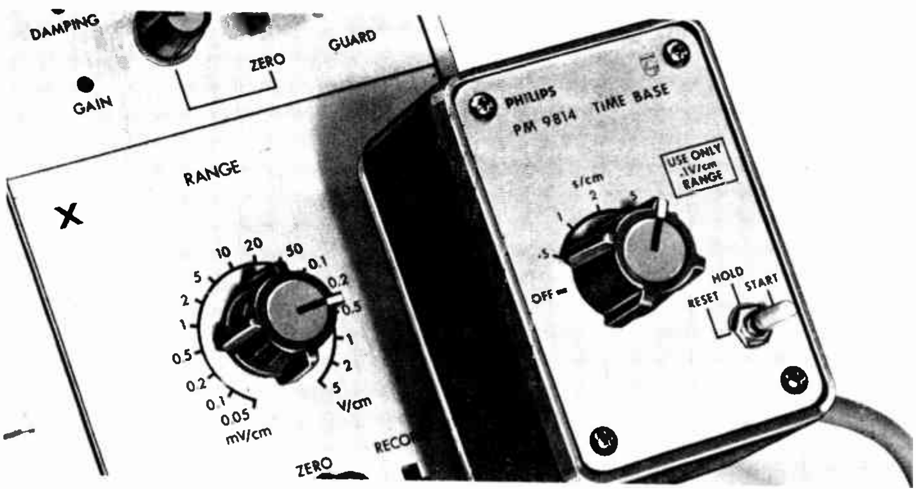


Fig. 5. Plug-in used to convert x-y flat bed recorder to x-t mode of operation.

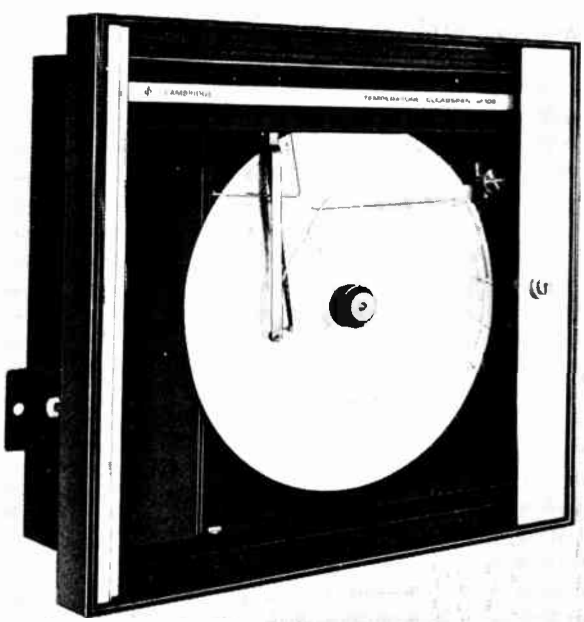


Fig. 6. Circular chart-recorder.

making the unit an x-t format recorder. Plug-ins generating appropriate ramps are often provided as an accessory – one is illustrated in Fig. 5.

Circular:– Where the Geometry of the measurement task is circular, such as recording out-of-roundness of a ground shaft, or where the measure has a cyclic time function, such as daily temperature changes, a circular form of chart is easier to use. The chart rotates under the marking device at a rotational velocity locked to the geometrical position or the appropriate sub-unit of time – hours, days, weeks and months. An example of a circular-chart recorder is given in Fig. 6.

The size of chart papers varies greatly from recorder to recorder. Strip charts are used from 50 mm width to around 800 mm with lengths as much as 150 m. The duration of the maximum record that can be taken on a roll is decided by the chart length and the chart speed. Flat bed units begin in paper size at about 200 by 300 mm ranging to huge computer-controlled automatic-draughting units with beds as much as 6 m x 4 m. Circular charts rarely exceed 300 mm diameter.

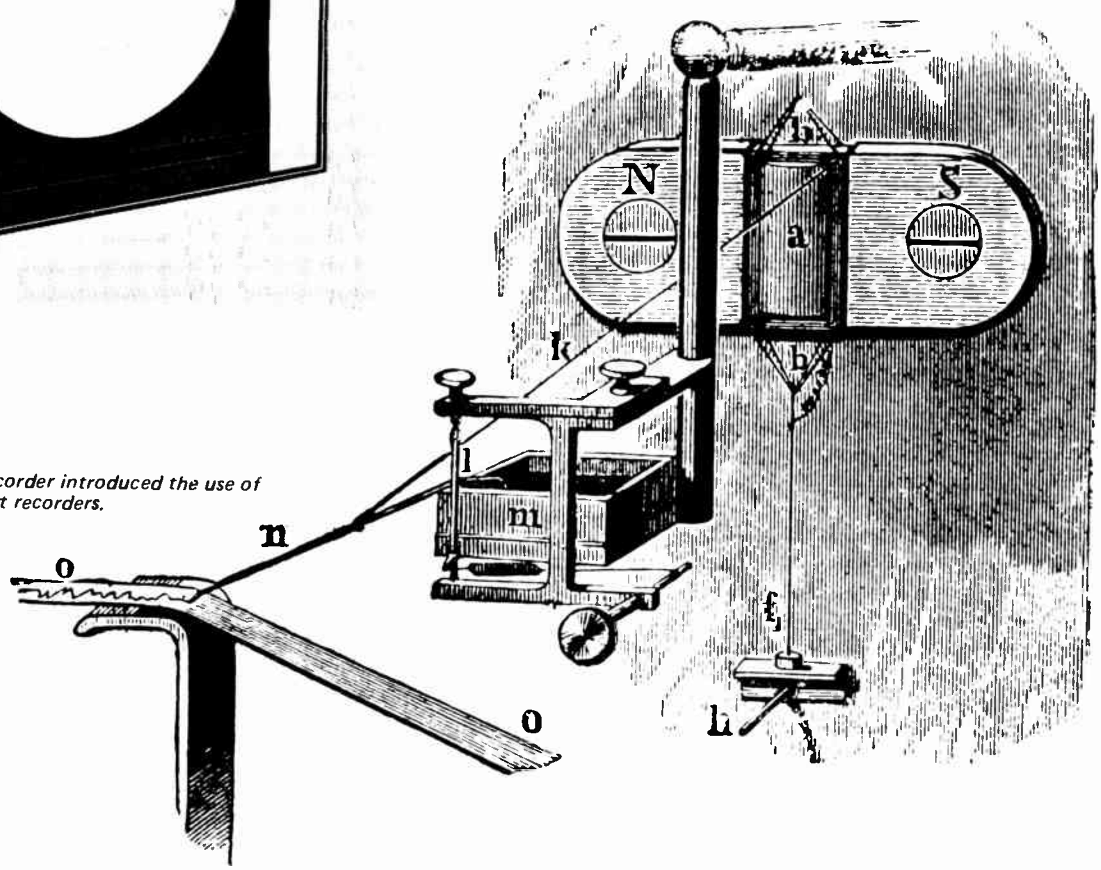
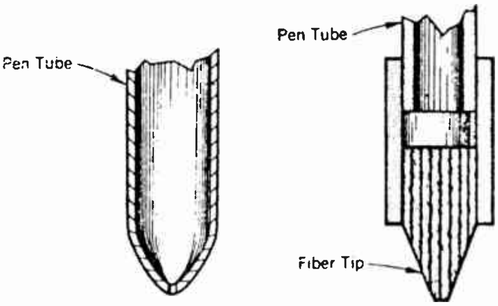


Fig. 7. Kelvin's 1873 siphon recorder introduced the use of continuous ink marking in chart recorders.



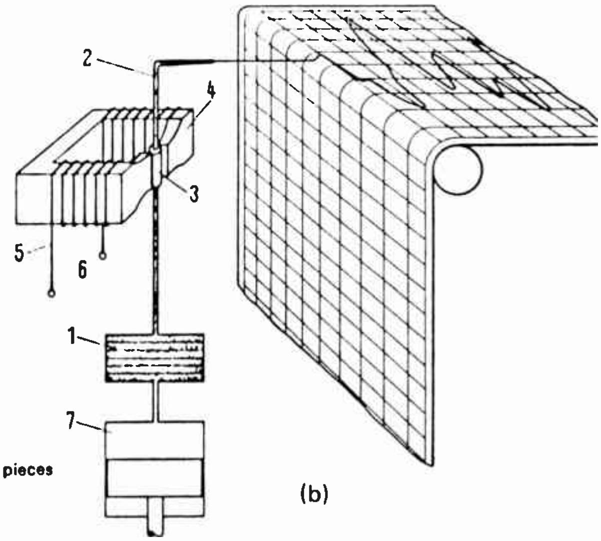


Fig. 8 (a). Capillary action ink pens.
(b) Schematic of pressurized ink recorder (Siemens).

Supply of chart papers can be difficult at times because stockists find difficulty in holding large stocks of the numerous options available. It is wise for the operator to hold a generous supply in hand at all times.

When reading values from paper charts care must be exercised in ensuring that inaccuracies caused by paper size changes, paper wander across its platen and marking mechanism offsets are allowed for. Good quality charts are a necessity with high-quality measurements.

PAPER MARKING TECHNIQUES

In these units an electronic amplifier coupled to a mechanical drive moves a mechanical point across the chart. It is then necessary to mark the paper in order to show where the point has travelled. Five commonly used techniques will be encountered.

Ink pen– Samuel Morse's telegraph recorder shown in Part 5 (Fig. 1) used a pencil to mark the paper strip. A limitation is that the lead wears away making a feed mechanism necessary. Ink can flow from a reservoir continuously: Kelvin introduced the siphon system in 1873 – see Fig. 7. This system is used extensively today in one form or other. Ink feed rate is a factor of the bore of the pen, paper absorbency and ink viscosity. Figure 8a shows pen details.

A second ink feed method uses a combination of gravity feed and capillary action through small bores. These are the ballpoint and fibre-tip pens. A third ink method pressurizes the ink, recording being performed by a very fine ink jet. This method is suitable for fast writing speeds (as high as 60 metres per second compared with around 1 m per second for unpressurized ink feeds). There is no

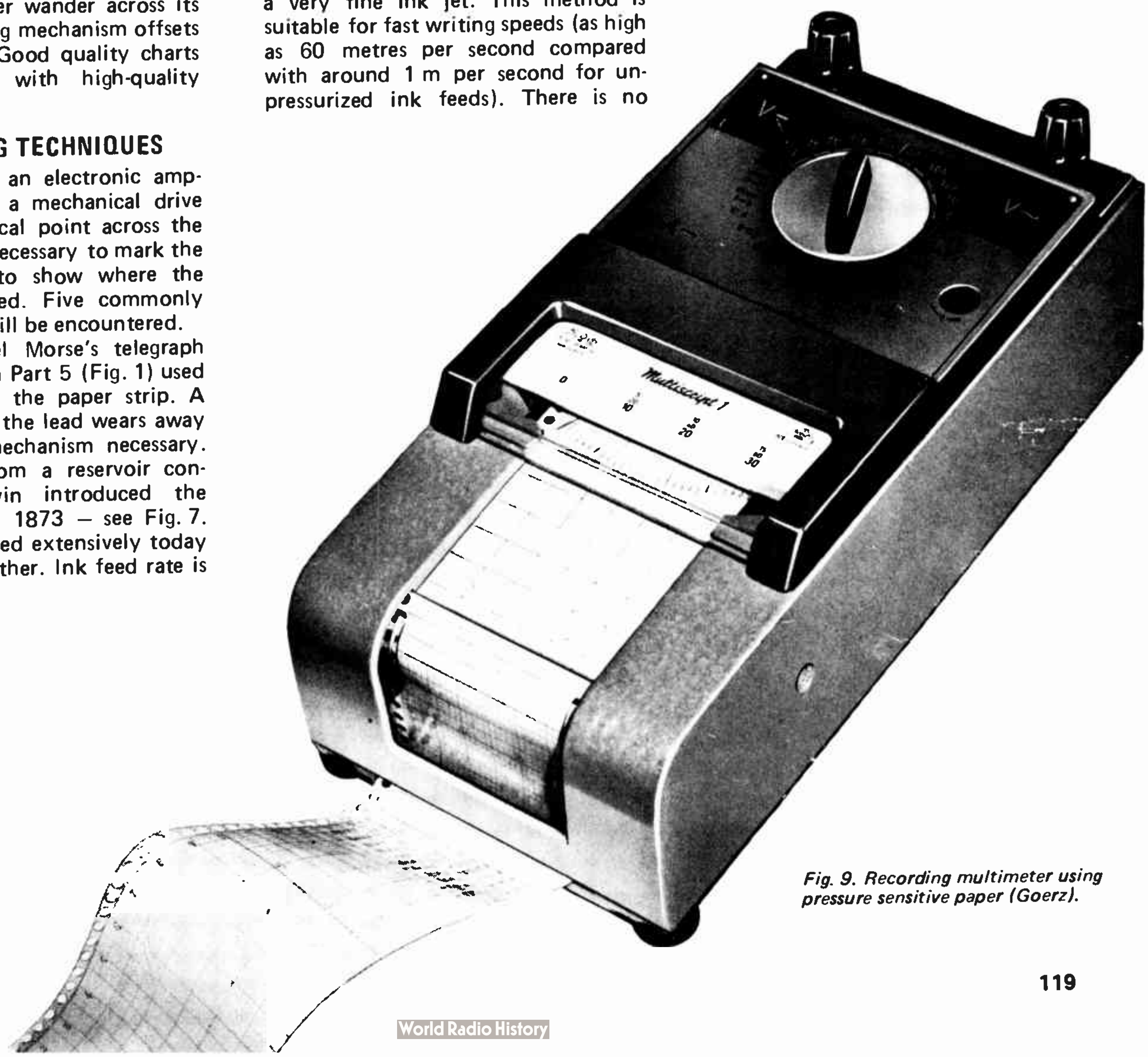


Fig. 9. Recording multimeter using pressure sensitive paper (Goerz).

Fig. 10. UV recorders provide traces by exposure of photographic paper. Further exposure is needed to bring the latent image into view.

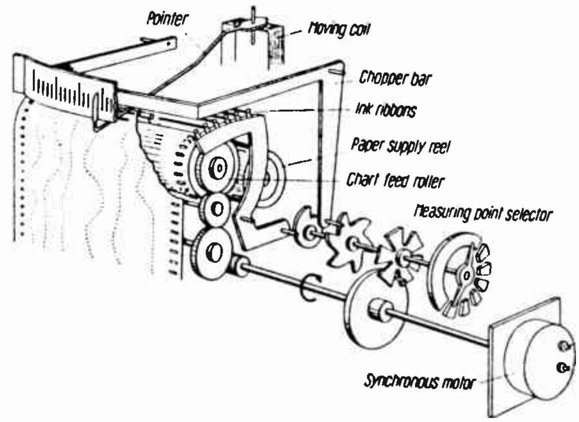


Fig. 11. Dotting recorders offer the advantages in slow-speed applications of being suitable for multi-channel multiplexing.

mechanical contact with the paper in pressurized systems, the fast writing rate arising because of the very small size of nozzle built into the deflecting system. Figure 8b shows the schematic of such a recorder. The pressure is automatically adjusted to suit the chart speed set.

The correct choice of ink and paper for the speed of operation is essential. Water-based inks are to be avoided as the record can be destroyed by accident. Fast drying inks are needed or else the trace may be rolled-up before the ink is dry. In short, although the alternatives to ink offer certain advantages we are still forced to use ink as the best all-round choice in many applications.

Pressure sensitive papers-- Black paper treated with tiny wax beads appears white until the beads are flattened to form a transparent cover window thereby exposing the black. Pressure sensitive papers are marked by the action of a gentle pressure exerted by the stylus. The relatively high contact-force needed restricts these to slow response application. Pressure-sensitive papers are more usually used with marking mechanisms that are periodically pressed against the paper to form a dot. Figure 9 shows a recording multimeter which uses this latter method of marking. Another limitation is that the record can be marked during handling.

Electro-sensitive papers: Some recorders use paper which is marked when an electric current is passed through it. The earliest was carbon impregnated; dielectric breakdown producing the mark by applying a high voltage between the stylus and the platen.

Another method electroplates onto the surface of paper made conductive by saturation with salts. It requires wet paper use but will operate with lower voltage levels than the above carbon paper method.

Zinc oxide reduced to free zinc is the process used in another kind of recording system. Metallized papers in which the metal film is fused to its paper backing are another. Yet another is based on providing a change in the paper surface which takes up toner (similar to the Xerox process) – it is fine for very fast systems but not those that occur slowly.

Heat senstitive papers: Yet another method of making the record is to use a heated stylus melting a wax-like coating on black paper. These papers can be manufactured with greater resistance to marking (during handling) than the pressure sensitive papers. Stylus temperature can also be varied with ease to suit the writing speed concerned.

Photographic paper: The earliest photographic systems used negative film. Such systems are still in use today but the majority of the highest speed recorders (30 kHz is possible) use ultraviolet light to expose specially treated paper. Exposure produces a latent (invisible) image which needs further exposure to form the visible image. This is shown in Fig. 10: the fluorescent lamp intensifies the traces.

Continuous versus dotting mechanisms: Fast writing speeds require continuous marking and for these the writing mechanism functions continuously. For very slow speed needs, as are found in process plant monitoring an alternative, in which a dot is produced on the paper at regular periods, has certain advantages. Figure 11 shows one form of mechanical arrangement. A separate motor, or pick-off from the chart drive causes a point to periodically press on the paper, marking it by the appropriate method used. By incorporating a geneva mechanism (one that rotates a shaft in steps) the input signal can be switched sequentially over a number of different signal channels (six and twelve are usual). Also synchronised to the channel changing action is an inking system that steps from colour to colour to provide a different coloured dot for each channel. Inking may be as shown (different ribbons) or may be provided as individual pads each soaked with ink. A multipoint dotting head wipes through this ink. One maker uses a multicolour single ribbon, akin to a typewriter ribbon.

Multi-channel operation is also provided in some continuous trace recorders. This is almost always achieved by incorporating separate recording heads for each signal. Figure 12 is a four pen recorder of the type in which the pens do not overlap: each trace is contained within a quarter of the full chart width. Multi-trace recorders in which each trace has the full paper width capability are also available. Mechanical drives have the disadvantage in that the traces must be slightly out of phase so that the pens can pass one another without fouling. Optical recorders do not suffer from this drawback.

RECORDING MOVEMENTS

We now look at the methods used to transduce the electrical input signal into an equivalent mechanical movement.

Moving coil mechanisms: Basically these use modified moving coil and pointer. The end of the pointer carries an ink pen or acts as a marking point when forced onto the chart paper in dotting styles (see fig. 11). Simple systems trace an arc across the chart giving a non-linear record. (curved markings on the paper overcome this but complicate the platen design). This can be linearized to provide better accuracy by various means such as that shown in Fig. 13.

Optical recorders also use a moving coil unit on which a mirror is mounted to reflect a high intensity focussed beam across the paper. These units have their origin in practical oscillographs designed by Duddell (to Blondel's ideas) at the turn of the century. The choice of galvanometer unit largely decides the frequency response. Today they are supplied as robust plug-in units like that shown in Fig. 14. The application, in many units, decides which galvanometer is used and the optimum terminating resistance value in order to know the deflection and sensitivity for a given frequency of signal. (Refer to reading list for guides). These recorders offer the ability to modulate the trace intensity producing 2-D half-tone chart records.

Potentiometric recorders: Around 1898 Professor Callendar devised his recording resistance pyrometer (Fig. 15) and in doing so provided instrumentation with the potentiometric or self-balancing recorder. This method makes use of a closed-loop system that causes the pointer to follow input signals. Referring to Fig. 16 the recorder has a drive motor mechanism which translates the pointer in one direction or the other depending upon the polarity of the signal driving the motor. Attached to the shaft driving the pen is a rotary resistance balancing potentiometer, as shown in Fig. 16a. Schematically this can be shown as a linear equivalent (the more recent design style used) as shown in Fig. 16(b). The potentiometer wiper moves across in unison with the pen and generates a changing

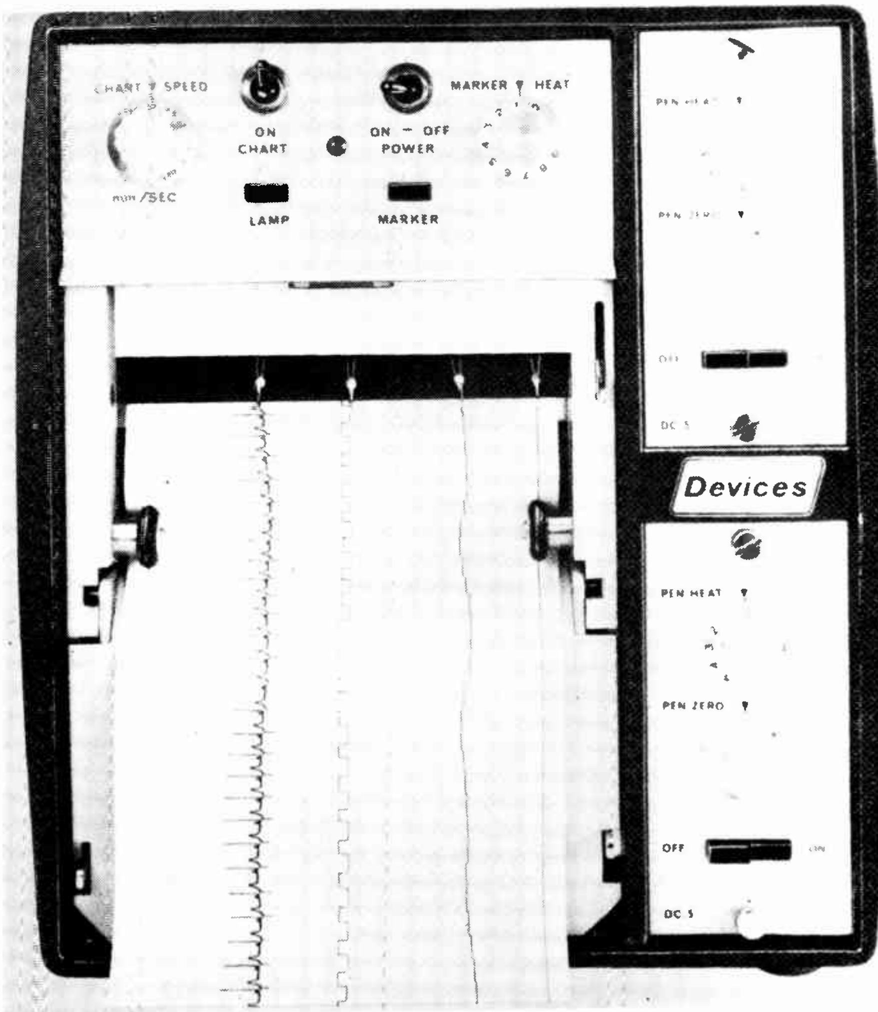


Fig. 12. High-speed four pen recorder. In this style the pens do not cross over each other limiting the trace width to a portion of the paper width.

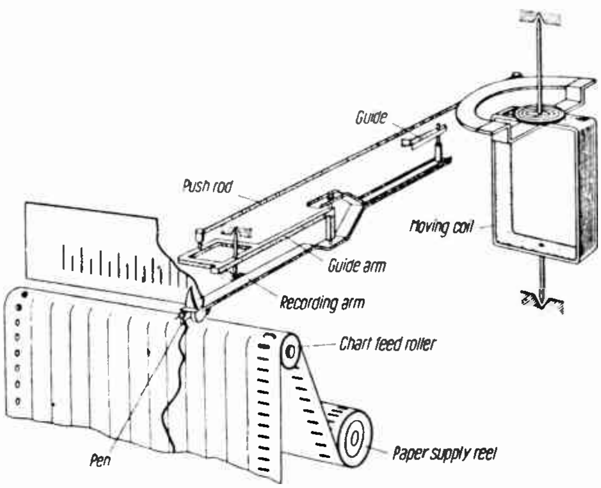


Fig. 13. Special linkages are used to linearize the non-uniform movement produced by a moving coil pen drive.

Fig. 14. Galvanometer unit for UV recorder (Hathaway Instruments).

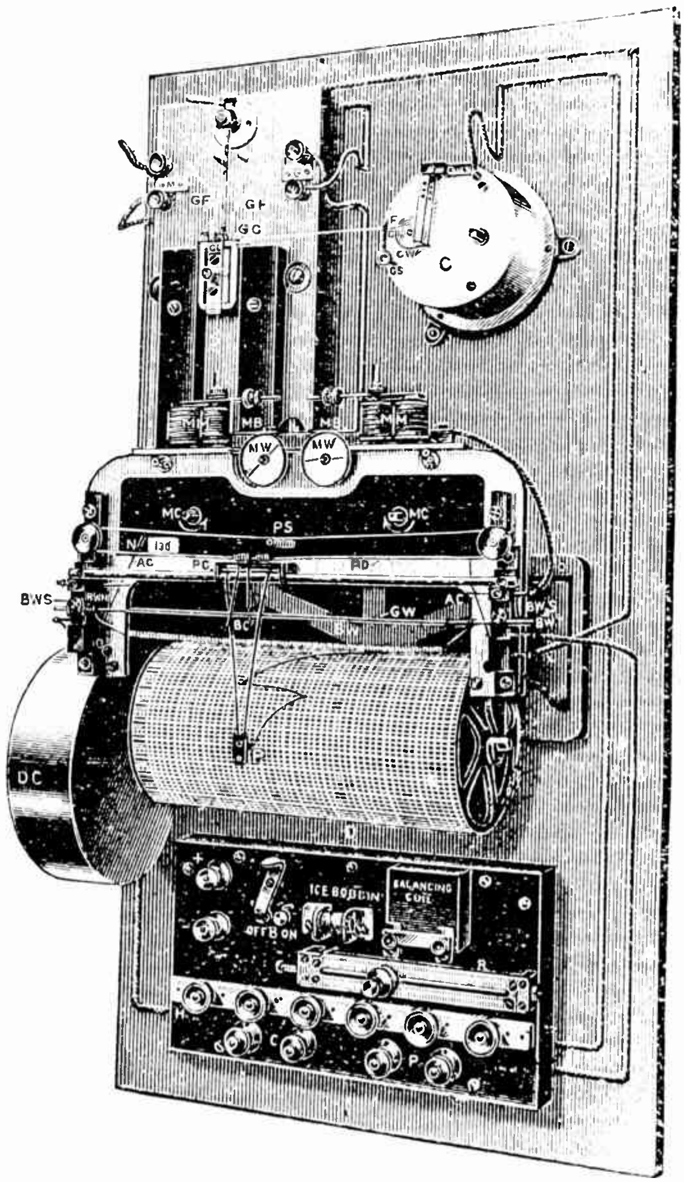


Fig. 15. Callendar's original potentiometric recorder was devised around 1898 to record furnace temperatures.

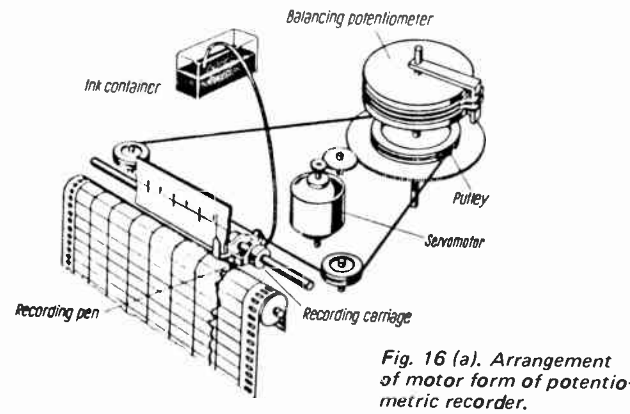
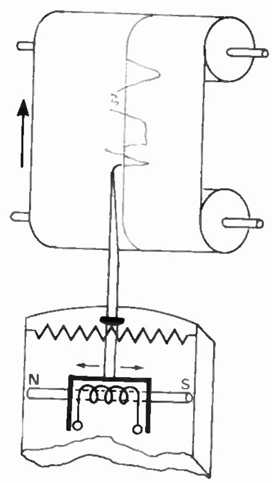
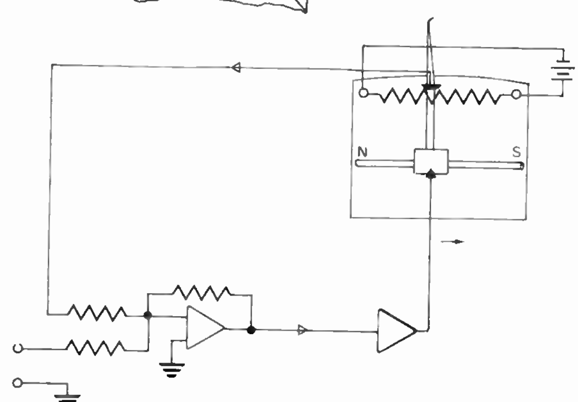


Fig. 16 (a). Arrangement of motor form of potentiometric recorder.



(b) Schematic using linear motor.



(c) Circuit schematic of potentiometric method (simplified.

value signal. The potentiometric system circuit layout is represented in Fig. 16(c). A reference voltage is supplied across the potentiometer. Voltage from the wiper is compared with the input signal voltage to be recorded. If a difference exists this constitutes an error which causes the drive motor to move accordingly to correct the error. The input signal and reference signals are suitably attenuated to provide the sensitivity needed at full-scale deflection.

The advantages of recorders such as those described above are that the mechanism plots a linear scale, and there is considerable power available to move the pen against frictional forces. The system, being potentiometric, draws little current once the unit has achieved balance and, as considerable drive power is available under closed-loop control, the pen response can be made tighter than for the open-loop pointer-type moving coil units. Sensitivity is decided more by the amplifier gain than mechanical constants. The majority of flat-bed recorders use this principle: at full trace movement their writing speeds can reach several metres per second. The method also overcomes the restriction on traverse length suffered by rotationally driven recorder mechanisms. Although a simple dc servo control is shown, potentiometric recorders, especially those built before around 1970 more usually used ac control systems.

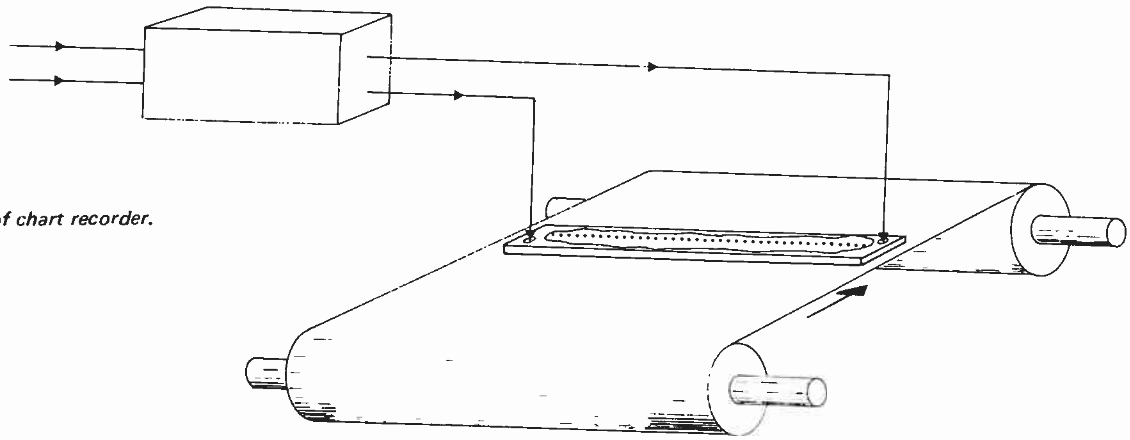


Fig. 17. Future style of chart recorder.

CRT – Fibre Optic Recorders: A recent design concept couples a CRT linear sweep trace to photosensitive paper via an optical fibre connection. This provides the highest response of all chart recorders so far available – dc to 1 MHz.

DYNAMIC RESPONSE

A point commonly overlooked is that chart recorders have a certain dynamic response and are effectively low-pass filters of the input signal. The response of a recorder to a sine signal, that is, the recorded trace, will look like the original but will lack adequate amplitude if the pen cannot follow fast enough. When quoting response rates it is therefore necessary to state amplitude as well as frequency. For example, moving-coil recorders with short pen arms – as in Fig. 12 – have a typical response that is flat from dc to 100 Hz at 10 mm peak-to-peak deflection for a sinewave. If the frequency is increased the recorder will still operate but the amplitude of a sinewave record falls off. Plots of complex waveforms may be severely distorted for the fundamental may be recorded at full amplitude with harmonics attenuated progressively. A square-wave input may be recorded as a near sine-wave if the response is inadequate. It is better to use a smaller signal amplitude in such cases.

Simple moving-coil chopper-type recorders will roll off from as low as 1 Hz. Ink jet units extend to 800 Hz: beyond that optical recorders are needed providing up to 1 MHz in the CRT design. Frequencies above this must be viewed by oscilloscopes using cameras to record the image.

Faithful response is also a function of amplifier characteristics. With the exception of simple moving-coil recorders most units have built-in amplification because the majority of signals to be recorded, have insufficient power to provide an adequate response. Recorder sensitivities may be fixed in manufacture, as in process industry dotting recorders, or have adjustable ranges. The manufacturers of recorders usually provide the amplifiers as part of the recorder, the purchaser only has to make the selection.

Event-marking recorders: In many recording applications the variable remains constant for more of the time than it varies. An example might be recording rainfall in dry areas. If the record must provide fine time-resolution the chart must run fast which means using immense lengths of paper for little data recorded. An approach, slowly finding acceptance, is to use a time/date printer which prints a value each time an increment of event occurs. Each increment print-out causes the chart to advance a unit. The result is a record chart completely filled with non-zero data. It is harder to interpret but much more efficient for spasmodic data situations. At present, however, this form of equipment is hard to procure commercially.

THE FUTURE

The design of recorders is decided largely by cost, reliability, sensitivity, and packaging to suit the application. Response and accuracy cost money. The weakest points of inexpensive recorders seem to be the reliability of the marking arrangement, and poor response. Optical recorders eliminate marking problems but still (with the exception of the CRT types) require fine electro-mechanical mechanisms to deflect the trace. We can confidently expect to see solid-state "deflection" systems marketed in the near future which are based on semiconductor technology. Units, like that depicted in Fig. 17, will use a linear array of LEDs to expose a spot on photographic paper in the appropriate place. This method would eliminate mechanical manufacturing problems, have excellent response characteristics and be readily multiplexed to provide multi-channel traces. Using LSI manufacturing methods, the cost of the array head and analog-to-digital converter would be minimal.

REFERENCES

Books containing chapters on chart recorders include:

"Basic industrial electronic controls" J. H. Ruiter and R. G. Murphy, Rolf Rinehart and Winston, 1962.

"Measurement systems" E. O. Doeblin, McGraw-Hill, 1966.

"Principles of instrumentation", J. T. Miller, United Trade Press, 1968.

A brief survey of recorders is given in:

"Instruments-electric recorders", Siemens, September, 1968. Liquid jet oscillographs are discussed in detail in Siemen's pamphlet MS7/200e, 1967 of that title.

A review of the merits of various writing systems is given in:

"Graphic recorder writing systems", D. R. Davis and C. K. Michener, Hewlett-Packard Jnl. October, 1968.

Chart inaccuracies are discussed in "Recording charts", L. Briggs Dunn, Instruments and Control Systems, July, 1969.

When using optical oscillographs the correct choice of galvanometer head and source impedance usually requires calculations to be made. (There is a trend toward elimination of this by providing suitable amplifiers). Manufacturers usually provide such detail. A paper "The Theory of recording galvanometers" by M. A. Le Gette, Consolidated Electrodynamics, Pasadena, California provides in depth detail.

36 The control of power

IN PART 10, WHEN DISCUSSING the types of amplifiers, we briefly mentioned the power stage found at the output end of electronic systems. Typical devices requiring amplifiers to drive them are loudspeakers, electric motors, and heaters.

The power handling capability of the various designs of these special amplifiers can range from one watt to many kilowatts. In this final part we introduce the special semiconductors and techniques used in electronic power control.

HEATSINKS

As some power is lost as heat in power transistors they may usually be recognized by the large heatsinks on which they are mounted. A rectifier stage using flat-plate heatsinks is shown in Fig. 1. These metal structures are needed to rapidly conduct away and dissipate to the air the heat generated at the junction of the device — this is a critical design requirement. The approach to designing heatsinks is common to all power components.

All semiconductors used in analogue control will have heat losses (the power lost as heat equals the current through the device multiplied by the voltage drop across it) which will cause the junction temperature to rise above the case outer temperature. For example, a transistor power amplifier stage may have at half output power (say) 10 V drop and 10 amp collector current. The heat loss is, therefore, 100 W and this must be liberated in order to keep the transistor temperature lower than its recommended maximum value.

All materials resist the conduction of heat to some extent — this property is called 'thermal resistance' and its value depends upon the material (copper is less resistive to heat flow than iron) and the cross-sectional area (increasing the area decreases the resistance). In practice catalogues for power components usually quote the thermal resistivity θ (which has units ^{o}C/W) between two points on the device. For example, typical measured temperatures for a certain power transistor mounted on a heatsink are as shown in Fig. 2. From these temperatures we can see that:—

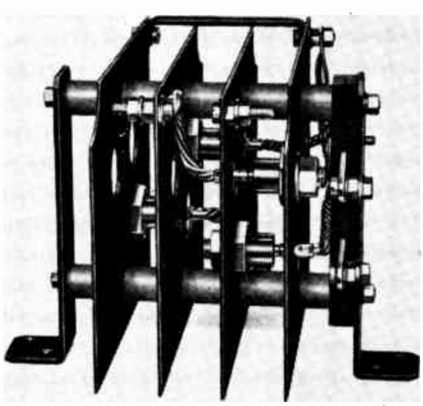


Fig. 1. Power handling stages are easily identified by the heatsink assemblies on which they are mounted — a 35 A, 60 VDC rectifier stack is shown here.

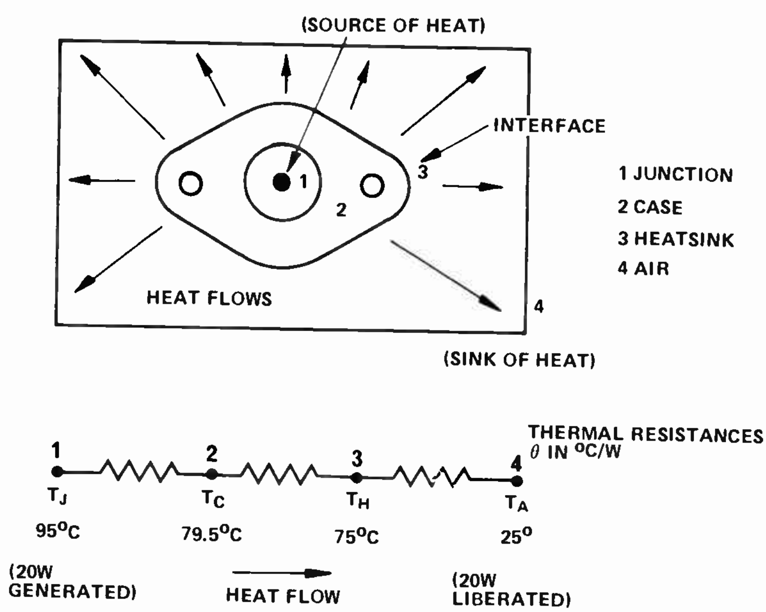


Fig. 2. Thermal resistances for a 20 W power transistor.

TABLE 1

$\theta_{J\text{-}C} = (95 - 79.5)/20 = 0.77\ ^{o}C/W$
$\theta_{C\text{-}H} = (79.5 - 75)/20 = 0.23\ ^{o}C/W$
$\theta_{H\text{-}A} = (75 - 25)/20 = 2.5\ ^{o}C/W$
$\theta_{J\text{-}A} = (\theta_{J\text{-}C} + \theta_{C\text{-}H} + \theta_{H\text{-}A}) = 3.5\ ^{o}C/W$

Where J = junction, C = case of device, H = heatsink and A = air.

Material used between device and heat sink (for insulation)	Thermal Resistance Dry	$\theta_{C\text{-}H}$ in $^{o}C/W$ with heat conducting grease
Direct contact (TO3)	0.20	0.10
Teflon insulator shim (TO3)	1.45	0.80
Mica shim (TO3)	0.80	0.40
Anodized aluminium (TO3)	0.40	0.35
0.25in stud mount (direct)	0.40	0.25
0.50in stud mount (direct)	0.12	0.07
0.75in stud mount (direct)	0.07	0.04

Fig. 3. Table of thermal resistances $\theta_{C\text{-}H}$ for typical mounting methods. Values can vary widely.

From this example we can see that the thermal resistance within the device — the parameter the user has no control over — is larger than the case-to-the-heatsink value. This means it is not worth improving the contact and heatsink material. The important thermal resistance is that between the junction and the air (presumed to be at constant ambient value); in many cases a different shape heatsink, one that transfers heat better to the air (finned for example) would make an improvement. The thermal resistivity (heatsink to air) can also be reduced by forcing air past the heatsink and/or by increasing the heatsink surface area. The latter measure, however, also has its limits because the thermal resistance between the device connection point and extremities of larger plates rises with increasing dimensions (reducing the effectiveness of outer areas).

The above example illustrates how a heatsink stage can be designed using the concept of series thermal resistances. In practice the design procedure must be worked in reverse. The aim is to ensure that the junction temperature remains less than a specified maximum limit. Beyond this quoted value the junction will be destroyed. A practical difficulty is that the junction temperature cannot be measured to ensure that the design is adequate so selection of mounting and heatsink type must be made with care using manufacturers' quoted thermal resistance values as the basis of a design. The following steps are given as a guide but full detail should be sought from more detailed accounts — see reading list.

Step 1: Assess the maximum power (W max) to be dissipated by the device. This will be the worst case of V.I product remembering to allow for temperature effects and maximum values. In switching designs the base to emitter junction voltage of a transistor is significant.

Step 2: Establish $T_J$max, $T_A$max from data sheets and expected ambient conditions. This enables the minimum required value of $T_{J\text{-}A}$ to be calculated.

Step 3: Calculate the overall thermal resistivity needed from $\theta_{J\text{-}A} = T_{J\text{-}A}/W_{max}$.

Step 4: Establish $\theta_{J\text{-}C}$ and $\theta_{C\text{-}H}$ from device table charts and the mount thermal resistivity for the device clamping method. Fig. 3 lists typical θ values for various clamping methods.

Step 5: Calculate $\theta_{H\text{-}A}$ required
$\theta_{H\text{-}A} = \theta_{J\text{-}A} - (\theta_{J\text{-}C} + \theta_{C\text{-}H})$

Step 6: Use heatsink tables to find

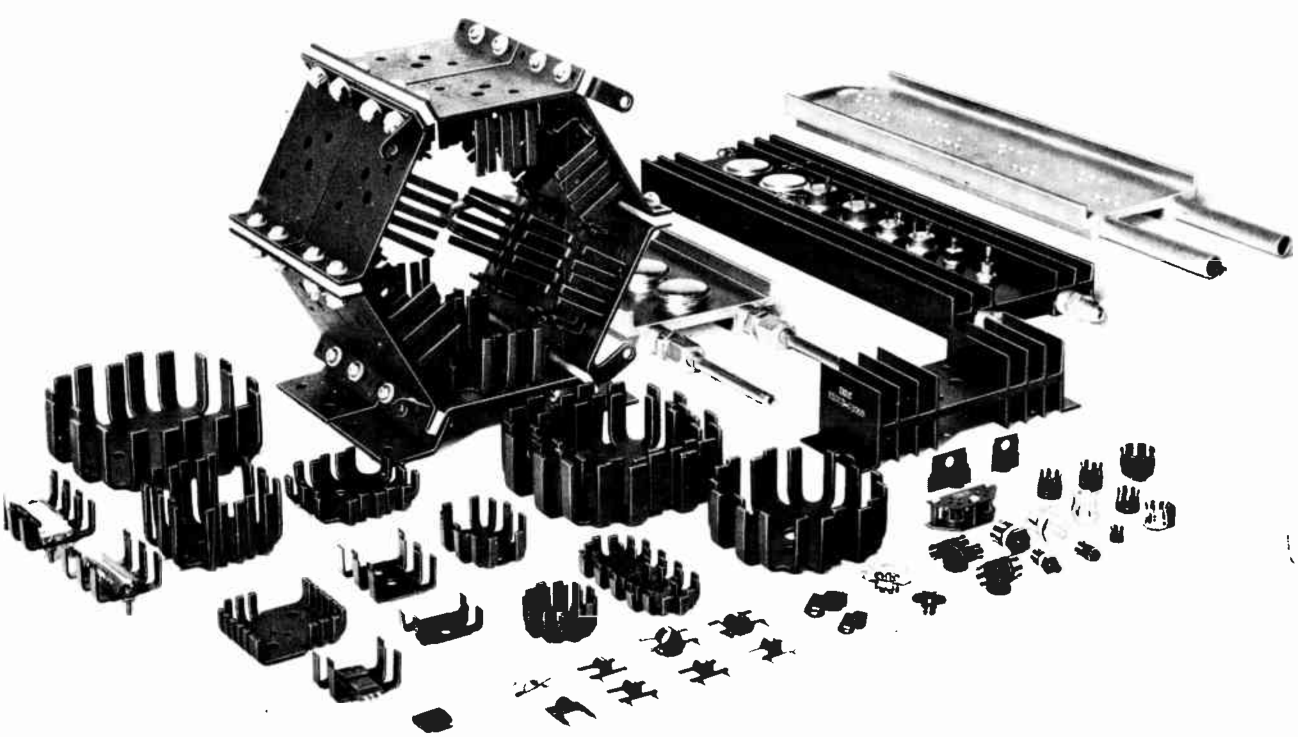


Fig. 4. Range of heatsinks for dissipating excess heat in semiconductors over a range from milliwatts to kilowatts.

suitable design having θ_{H-A} value or smaller.

In general if θ_{H-A} needs to be less than 2 to 5 °C/W the heatsink becomes prohibitively bulky. Design of the whole system is usually limited by the manufacturer's value of θ_{J-C}, which cannot be reduced. The interface coefficient θ_{C-H} is usually around 0.15–0.20 °C/W for direct contact using the recommended heat conducting silicon grease. Mica insulation degrades this value a little, poor heat conducting insulators should be avoided as they contribute a quite high value of θ_{C-H}.

Heatsinks for analogue control power units will need to be much larger than those of switching designs such as the switching regulator and normal rectifier stacks. This is because the latter need only dissipate the V.I product of the two extremes of V and I. The voltage drop across a power diode running at many amperes is around one volt: when reverse biassed the voltage is high but the current negligible.

Figure 4 shows a wide selection of heatsinks including units for fluid cooling applications. Fins should always be positioned to assist the vertical convective flow of air over the surfaces. Total immersion of the electronic circuit in cooling liquid is not used.

POWER TRANSISTORS

Power transistors are little different to small-signal devices in their basic semiconductor principle of operation: the distinguishing factors are the heavy-duty design which enables high collector currents and voltages to be controlled. The junction areas are much larger and the case design is made to keep the thermal resistivity as low as possible (around 0.8 °C/W) in order that the losses can be removed. Collector currents being higher and the gains being lower than small-power transistors means the base currents are also large. Thus, high power stages have to have lesser power stages driving them. They are available for several hundred volts operation and current levels exceeding a 1000 A. Cut-off frequencies into the gigahertz region are available (with less gain than that of lower frequencies). At RF frequencies gains range from 4-13 dB for powers in the range 0.1–80 W. There are few power applications that transistor devices cannot handle. In practice, however, certain other semiconductor devices are often a better choice.

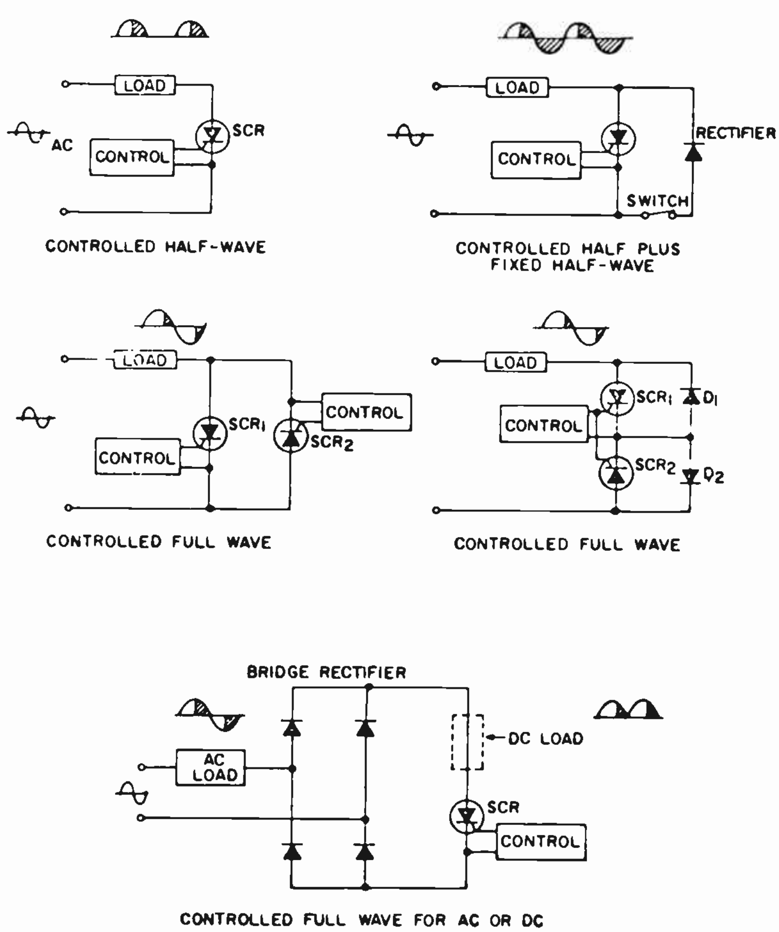


Fig. 5. Five arrangements by which a load can be fed with power flow controlled by SCR devices.

SCRs, THYRISTORS AND TRIACS

Semiconductors and diodes have one p-n junction and transistors have two junctions, p-n-p or n-p-n. A logical progression is the three-junction device, p-n-p-n. This family contains such devices as the silicon-controlled rectifier SCR, the silicon-controlled switch SCS, the gate-turn-off switch GTO, the light-activated, silicon-controlled switch LASCS, and the Shockley diode. Of these, the SCR (also called a thyristor) mainly concerns us as it is able to control high-power levels (they were introduced in Part 16). The SCR has an anode and cathode and a gate lead (which when held positive prevents the unit from conducting).

By controlling the gate voltage it is possible to control when power begins to flow during an ac cycle. Once the SCR is triggered (or fired) it remains on until the anode-cathode voltage falls to zero again. SCRs are, therefore, extremely useful when an alternating current source is available as this automatically provides the necessary switch-off conditions at each half cycle.

TRIACS are special SCRs that can be switched on to allow both positive and negative half cycles to pass. This action can also be arranged by using two SCRs.

This class of device cannot control the flow of dc power from a dc source, because once turned on they remain on, acting like an adequately low-resistance contact. They are, however, invaluable for controlling loads which can be energised by ac power – heating coils, motors, lighting and furnaces.

The operating circuitry for an SCR is designed to provide the appropriate gate on-voltage level at the correct time during the half cycle. Fig. 5 shows five basic forms of phase control. A typical trigger circuit is given in Fig. 6. One difficulty in this kind of control is that large line transients are generated, along with RF interference, when the power begins to flow during each cycle.

A more refined type of control derives the required average output power as the mean of a series of complete whole-cycles rather than in the mean of many partial cycles. This method generates substantially reduced line transients and RF interference because switching always occurs at the zero voltage condition: Figure 7a shows one form of proportional zero-voltage-switching controller using a TRIAC to control the heat produced in the element. Figure 7b is a typical output signal burst of gradually increasing power.

Capabilities of SCR devices range to hundreds of amperes, reverse voltages to as much as 2000 V. The maximum voltage drop across the turned-on SCR lies in the range 1.3–2.5 V, with leakage currents being in the region of 40 mA in the turned-off state.

These characteristics may make SCR devices appear extremely robust. Design of reliable, high-power, units, however, is a matter for a specialist. Many pitfalls can occur if their operation is not understood in detail. Designed properly they will, however, give utmost reliability.

Fuses for SCR circuitry also need special consideration because semiconductor junctions when overloaded will blow more rapidly than simple wire fuses or electromagnetic circuit breakers. The criterion is that the I^2t rating of the SCR must be greater than that of the fuse. I^2t values are usually provided in maker's data sheets. During the turn-on period of the SCR this value may drop significantly. Selection of adequate protection fuses is a matter that must be studied in some depth. Care must be taken to mend blown fuses in SCR units with the correct replacement — this invariably means carrying the correct spare ready to use.

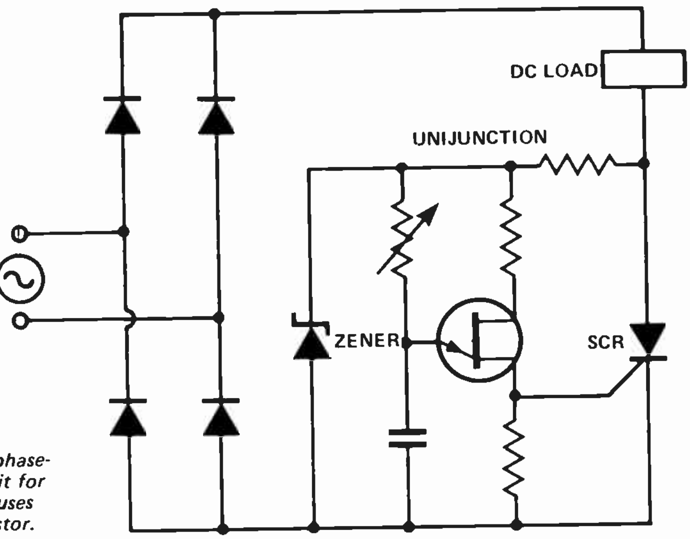


Fig. 6. This typical phase-control trigger circuit for controlling an SCR uses a unijunction transistor.

SWITCHING REGULATORS AND CONTROLLERS

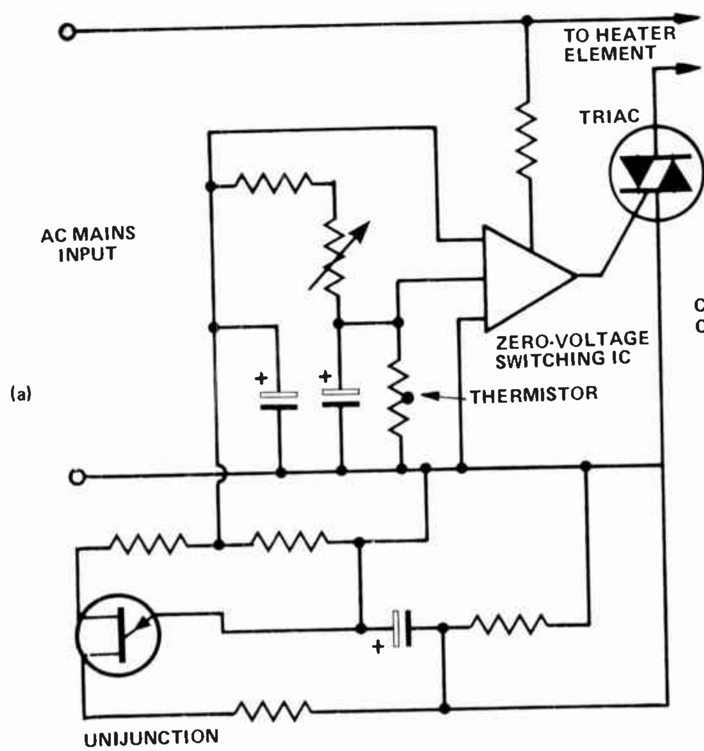
Parts 15 and 16 discussed methods used to regulate dc power supply output voltages or currents. For small power levels — a few watts — the series regulator and zener diode arrangements are acceptable because the power they dissipate is not an economic factor. The controlling transistor (as is shown diagrammatically in Fig. 8) can, instead, be used as a switch varying the on-to-off time ratio (mark-space is the term used) in order to vary the average dc power obtained after smoothing.

The switching method has the significant advantage of very small losses in the regulator stage. The transistor is either fully-on (high current but very low voltage) or fully-off (highest voltage but minimal current). As well as reducing the losses the method also can use a smaller capacity transistor. The price paid is the need for a filter stage and for a pulse generator to drive the switch.

Switching regulators are especially necessary when the voltage drop between the source and the load requirements becomes large.

Modern designs often make use of an integrated circuit as the basic control unit adding an additional switching transistor to cope with the output current needed. Fig. 9 is a high-current switching regulator which can supply 3 A continuously at 30 V input with losses sufficiently small to allow the use of quite small heat-sinks.

Switching is also a suitable method to efficiently control output loads — the difference between this and regulator design is that the feedback loop (dotted in Fig. 8) is not used; the mark-space ratio of the generator being controlled instead by the input



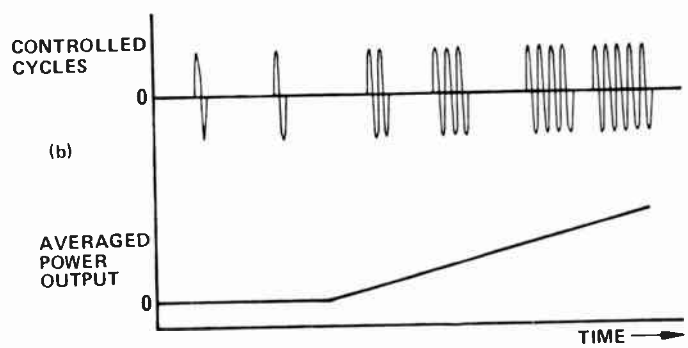


Fig. 7.
(a) Zero-voltage-switching temperature controller using a zero voltage switching IC driving a TRIAC from a thermistor sensor.
(b) Output signal with gradually increasing power.

signal to be amplified. This principle is used in high-current dc motor control and in advanced forms of audio amplifier.

INVERTERS AND CONVERTERS

A converter, in the electrical power engineering sense, is a machine (or a circuit) that changes current from one kind to another, or from one frequency to another. An inverter, in the same sense, is a machine that specifically converts dc to ac – being one kind of converter. Originally rotating machines were used but today the trend is to use static solid-state equipment.

There are many instances where these are required – providing a 240 V ac 50 Hz supply when only 12 V batteries exist, providing a 200 V dc supply from 12 V dc and to change frequency such as where a 240 V ac 50 Hz mains might be needed to drive aircraft equipment operating at 400 Hz.

The basic principles used in each are based on the technology discussed before in this part. These are now summarised with examples of the procedures used.

AC to DC: This conversion path has been discussed when we dealt with rectification in Part 14. A transformer is used to obtain the required ac voltage; this is then rectified with diodes and smoothed to provide dc.

DC to AC: This path first changes the dc into a suitable ac signal which can then be transformed to the desired signal level. The frequency of the ac signal is decided by the output load requirement for once produced it must remain at that frequency. (In some cases it is preferable to make use of a higher frequency than 50 Hz).

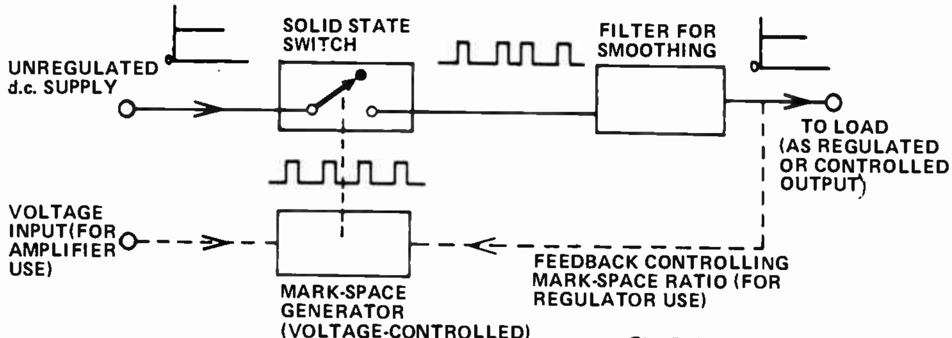


Fig. 8. Schematic of switching regulator or controller (see text).

Figure 10 shows a number of configurations used to produce ac power from a dc supply.

Switching produces square-wave energy after inversion and in many instances this roughly square-wave output waveform is satisfactory. Where the output must be sinusoidal more complex circuitry is required to obtain an undistorted wave shape. When choosing a commercially made inverter it is important to verify if the output waveshape is suitable for the task.

Crystal oscillators can be incorporated into an inverter design where the output frequency must be kept within exacting limits.

DC to DC: The procedure here is to first form the dc to ac conversion. After transformation to the correct voltage (usually the need is a voltage increase) with a double-wound transformer the output is full-wave rectified and smoothed. The transformers used use special iron laminations material to get the best out of the square-wave input waveforms. Figure 11 is a typical 200 W dc to dc up-converter. The transistors, in conjunction with the transformer primary, form a square-wave oscillator circuit causing flux changes in the transformer which induce the higher output voltage needed for rectification back to dc current.

AC to AC: Some mains equipment can run on either 50 or 60 Hz frequency with little change in performance. Occasionally, however, it is necessary to use the correct frequency specified. To change frequencies the simplest procedure is to convert the original ac supply to a suitable dc value inverting this back to ac at the other frequency. This procedure is easiest to implement because it makes use of standard rectification and inverter packages.

The cost of semiconductor converters has fallen rapidly over the 1970 decade. This has brought about new philosophies in power electrical engineering. In the future there will be more use made of dc electrical transmission. Speed-changing motors are becoming easier to implement using frequency-varied supplies to drive conventional ac machines. Large dc motors are also becoming useful again because regenerative braking of large units – using them as a generator driving into a load – can be put to use

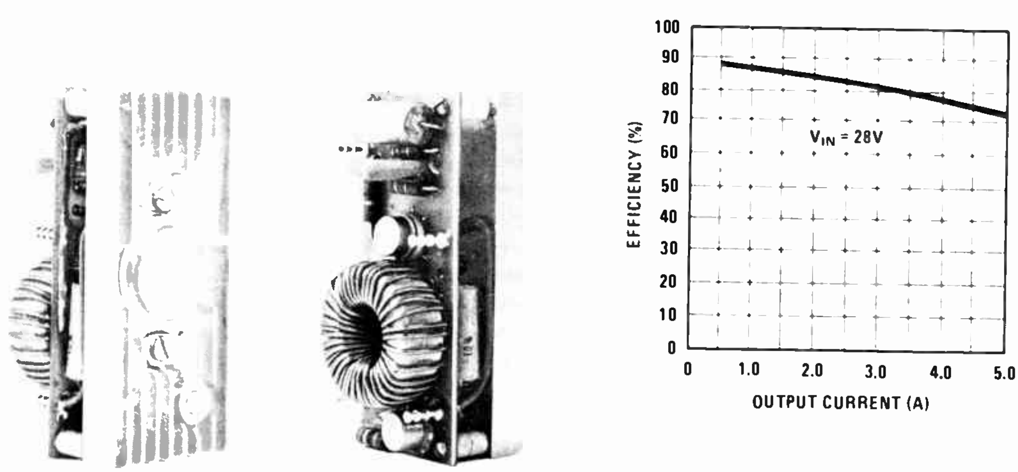


Fig. 9. Switching regulators provide the smallest overall package due to the greatly reduced heatsink needed. They are also the most efficient form of regulator.

to charge power into the ac mains by the use of dc to ac inverters.

Revolutions have occurred in both power and signal electronics. Attitudes to problem solving are now quite different to just a decade ago. No doubt this trend will continue.

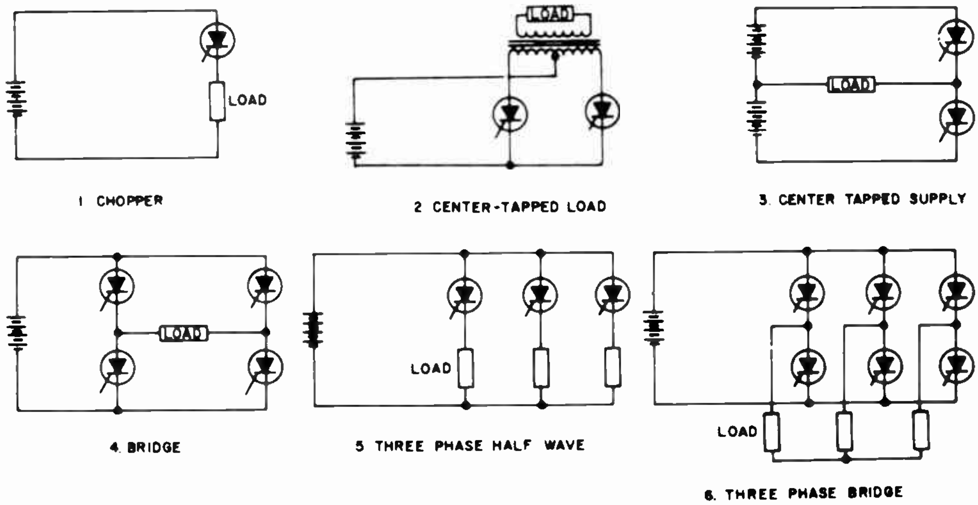


Fig. 10. Various inverter configurations using SCR switches. Triggering methods have been omitted for clarity.

REFERENCES

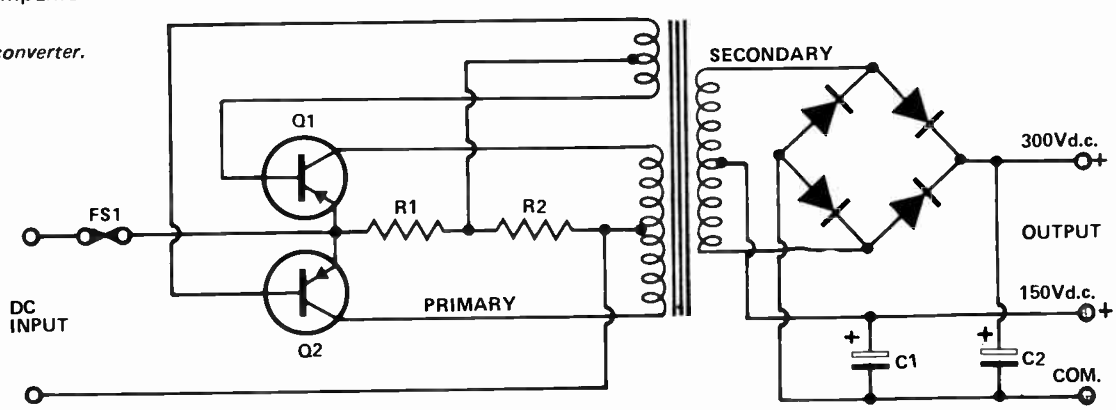
Heatsink Design: Several articles were published in **Electronics** October – December, 1973. Most power semiconductor device manufacturers have design sheets available. Heatsink makers usually provide graphs and design procedures. Forced-air cooling is discussed in "Cooling high-power equipment by forced-air convection", L. Katz, **Electronics**, September 21, 1964. A good basic treatment of design is "Heatsinks for power transistors" **Miniwatt Digest**, December, 1962. A deeper discussion is "Design of cooling fins for silicon power rectifiers" **Mullard Tech. Comm. Vol.5**, June, 1960. "Silicon Rectifier Handbook" Motorola, 1966, devotes a chapter to heatsink design.

Power Devices The above Motorola handbook is most useful on power diode design. A companion volume "Power transistor Handbook 1961" deals with transistors. SCRs are covered in "SCR Manual", General Electric, 1964.

Converters and Inverters: An early paper "Self-excited inverters" W.J.R. Farmer and R. J. Spreadbury, **AEI Engng.**, March, 1965 provides a tutorial summary. Many power engineering companies can provide papers and articles.

Switching regulators: Low to medium power units based on an IC are discussed in "Designing switching regulators" R.J. Widlar, **National Semiconductor AN-2**, 1969. See also "Switching regulators: the efficient way to power", R.S. Olla, **Electronics**, August 16, 1973, and application notes of Motorola, Delco Radio, RCA, Kepco, and others. ●

Fig.11 Circuit of dc up-converter.



CONCLUSION

In selecting the material for this series, it would have been helpful to have had powers of prediction. Even the greatest minds have been wildly wrong about the near future. Lee de Forest who invented the triode valve in 1906, had this to say in 1937:-

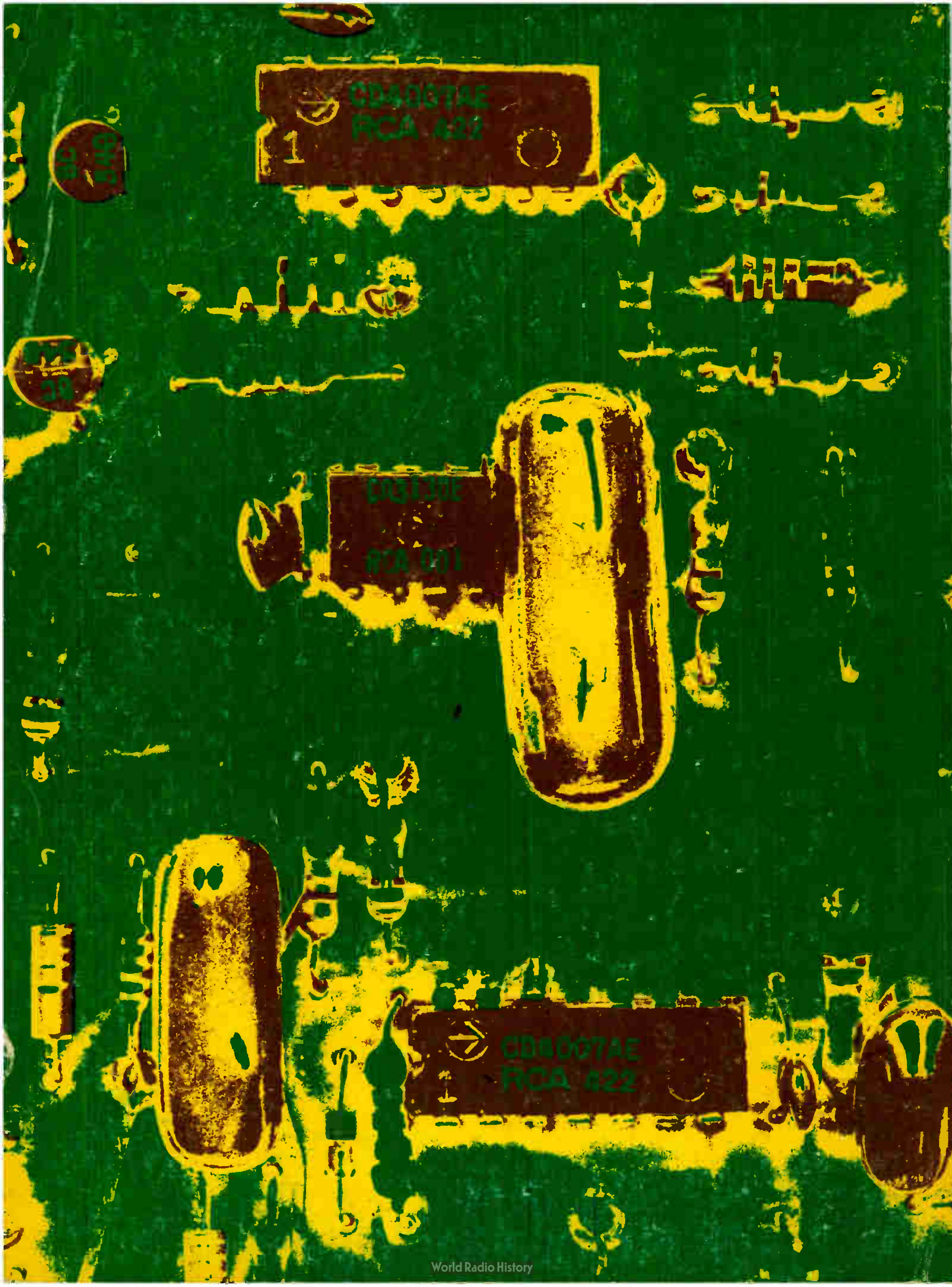
"While theoretically and technically television may be feasible, commercially and financially I consider it an impossibility, a development of which we need waste little time dreaming".

Having knowledge of materials does not ensure that they will find immediate use. Germanium was predicted well before it was isolated in 1886. Silicon was discovered in 1823. Both were available at the time of the emergence of the valve in the 1900's; so was the idea of the transistor. History shows time and time again that ideas seem to sit and wait. Few people in the 1950s foresaw the great impact that the transistor would have on everyday life. In 1955, Sir George Thomson, a great physicist, said this about the germanium transistor:-

"It is possible that a short-range 'walkie-talkie', light enough to be carried regularly, may in the course of a few decades replace a good deal of telephony over wires".

Who could have forecast the great changes to electronics over the following two decades. One wonders what is to come in the next two!

Index


CD4007AE
RCA 422
CD4007AE
RCA 422