FOUNDED 1925
INCORPORATED BY
ROYAL CHARTER 1961

"To promote the advancement of radio, electronics and kindred subjects by the exchange of information in these branches of engineering."

THE
RADIO AND ELECTRONIC
ENGINEER

The Journal of the Institution of Electronic and Radio Engineers

VOLUME 37 No. 6 JUNE 1969

Digital Methods of Measurement

PULSE techniques in radio engineering first became important with the development of radar; after the Second World War the digital computer made use of many of the techniques developed for radar and since then there has been what might be called a 'digital revolution'. Many instruments today employ a digital form of presentation while there is a growing number of measurement applications in which there are advantages in using digital rather than analogue techniques. The digital frequency meter and the digital voltmeter are two well-known examples—considerations of 'human engineering' are significant reasons for digital display to be adopted in many situations. Where measurement and testing is required to be automatic, programmable sources under the control of computers or similar apparatus which produce digital command signals are necessary.

To an ever-increasing extent communication systems are also 'going digital': modulation of pulse width, frequency or amplitude was introduced during the War and now with the great flexibility for signal processing which integrated circuits provide, the very versatile technique of pulse code modulation is much in evidence throughout the whole field of communications. There is in addition the expansion of data transmission networks, both nationally and internationally, and these, and other applications, call for special measurement techniques.

All these developments have made it appropriate for there to be held, at this particular stage, a Conference which will present some of the common factors and provide a forum for engineers to discuss common problems. The Institution of Electronic and Radio Engineers with the association of the Electronics Division of the Institution of Electrical Engineers and the United Kingdom and Republic of Ireland Section of the Institute of Electrical and Electronics Engineers, has therefore organized a three-day meeting which will be held in the University of Kent at Canterbury from Wednesday 23rd to Friday 25th July next. The provisional programme of the Conference listing the 35 papers to be presented is given on pages 337 and 338 of this issue and it will be seen that the subject has been dealt with under three main sessions, on the first day 'Measurement of Physical Quantities', on the second day 'Digital Instruments', and on the third day 'Measurements on Communications Systems'. The authors of the papers are representative of the three major sectors of national endeavour, namely industry, the universities, and public corporations. In addition there will also be three papers from European organizations in each of these categories. The interest which the first announcement of the Conference aroused among engineers in other countries generally indicates the likelihood of a sizeable group of participants from overseas taking part in the discussions.

The advantages of holding a Conference outside London at a University where many of those taking part can be in residence were pointed out in the May issue of *The Radio and Electronic Engineer*, and there will be the usual opportunities for informal discussion in Eliot College, one of the three Colleges which comprise the new University of Kent. This University has not been visited for a conference previously by any of the sponsoring Institutions—its excellent facilities as well as its fine situation overlooking one of the most famous and interesting of English mediaeval cathedral cities will surely add to the attractions to the engineer of the subject of the Conference!

Overall, the Conference on 'Digital Methods of Measurement' promises to prove a well-worthwhile survey of work which is in progress in an area of measurement techniques that will involve virtually every electronic and radio engineer sooner or later.

K. J. DEAN

INSTITUTION NOTICES

U.S. Honour for I.E.R.E. President-Elect

On 29th May last the Pioneer Award for 1969 of the Aerospace Electronic Systems Group of the Institute of Electrical and Electronics Engineers was presented to Mr. Harvey F. Schwarz (Fellow) and Mr. William J. O'Brien (Fellow). This annual award, which takes the form of a plaque, was made on the occasion of the National Aerospace Electronics Conference in Dayton, Ohio, and recognized the lasting importance of the pioneer work of the two recipients in the invention, development and application of the Decca Navigator System.

Mr. Schwarz and Mr. O'Brien are respectively Managing Director and Research Director of the Decca Navigator Company. Mr. Schwarz, at present a Vice-President of the I.E.R.E., has been nominated for election as President for the year 1969–70. In a speech at the Conference in Dayton he conveyed greetings from the Institution to its American counterpart and spoke in some detail of the early history of the Decca Navigator, notably its adoption for use by the Royal Navy during the landings in Normandy in June 1944. The definitive paper on this revolutionary c.w. phase-comparison, hyperbolic navigation system was read before the Institution by Mr. O'Brien at the Convention in Bournemouth in 1947, and following its publication in the *Journal* (October 1947), the author received the Heinrich Hertz Premium.

Institution Premiums and Awards

The Council of the Institution announces that the following awards are to be made for outstanding papers published in *The Radio and Electronic Engineer* during 1968.

REDIFFUSION TELEVISION PREMIUM

'Colour Television Studio Equipment and Problems' by F. G. Parker (October).

P. PERRING THOMS PREMIUM

'Television by Satellite' by L. H. Bedford (November).

J. LANGHAM THOMPSON PREMIUM

'A Positioning System using Three-Valued Codes' by A. Pugh (October).

A. F. BULGIN PREMIUM

'F.M. Deviation Measurements' by P. Broderick (May).

LESLIE MCMICHAEL PREMIUM

'Wideband Circular Array for H.F. Communications' by I. D. Longstaff and D. E. N. Davies (June).

DR. NORMAN PARTRIDGE MEMORIAL PREMIUM

'The Design of Low-Noise Audio-Frequency Amplifiers' by E. A. Faulkner (July).

ARTHUR GAY PREMIUM

'High Density Soldered Interconnections' by A. G. Cozens and J. E. Tomlin (August).

LORD BRABAZON AWARD

'Radar Pulse-Compression by Random Phase-Coding' by G. C. Bagley (July).

CHARLES BABBAGE AWARD

'High Speed Computer Logic with Gunn-Effect Devices' by H. L. Hartnagel and S. H. Izadpanah (October).

MARCONI AWARD

'A C.W. Comparator for Precision R.F. Attenuators' by R. W. A. Siddle and I. A. Harris (March).

The following premiums and awards are being withheld as papers published during the year falling within their respective terms of reference were not of a sufficiently high standard: Clerk Maxwell Premium, Heinrich Hertz Premium, Vladimir K. Zworykin Premium, Hugh Brennan Premium, Lord Rutherford Award.

Leeds University Electronics Exhibition

The President of the I.E.R.E., Sir Leonard Atkinson, will officially open the Electronics Exhibition to be held at the Department of Electrical and Electronic Engineering at the University of Leeds at 10.30 a.m. on 1st July. In the afternoon following the opening two lectures sponsored by the Yorkshire Section of the I.E.R.E. will be given: 'The British Calibration Service' by F. L. N. Samuels, and 'Measurement for Process Control Computers' by S. L. H. Clarke. (Mr. Clarke is currently vice-chairman of the Institution's Instrumentation and Control Group Committee.)

The Exhibition, which has been organized by the Department, will have 100 exhibiting firms and will include a section of 12 stands showing electronic equipment and devices for medical applications.

Tickets for the Exhibition, which will cover attendance at the lectures, as well as any further information, may be obtained from Mr. C. S. Petch, Exhibition Organizer, Department of Electrical and Electronic Engineering, University of Leeds, Leeds LS2 9JT (Telephone: 0532-31751).

Index to Volume 37

This issue of *The Radio and Electronic Engineer* is the last in the present volume, No. 37. An index to this volume will be circulated with the August *Journal*.

Members who wish to have their *Journals* bound are asked *not* to send them in before the Index is published.

U.D.C. 621.397

Future Trends in Television Broadcast Engineering

By

A. V. LORD,
M.Sc., C.Eng., F.I.E.E.†

Presented at the Institution's Convention on 'Electronics in the 1970s', held in Cambridge on 2nd to 5th July 1968.

Summary: Future engineering objectives for each of the parts of the transmission chain are forecast. They include smaller cameras with higher sensitivity (possibly using solid-state sensors), flying-spot film scanners operating on both the 625/50 and 525/60 standards, automated signal switching systems, improved recording systems using high-density magnetic and laser techniques, the use of p.c.m. for television links (which may employ optical fibres), and solid-state receiver displays.

1. Introduction

At the present time the situation in most of those countries which have operated television services for an appreciable period is such that the number of monochrome receivers per hundred of the population exceeds 20 and, broadly speaking, these countries have either already introduced colour to a significant extent or are planning to do so very soon. Presumably those countries of the world which have only started television relatively recently will follow a similar pattern of development.

Colour is already playing a very significant role in the development of television broadcast engineering and in some countries considerations of colour are predominant. Recent figures quoted for the number of colour receivers in various countries include 15 million in the U.S.A., over 1½ million in Japan, over 100,000 in West Germany and some 80–90,000 in the U.K. In the U.K. and West Germany the rate of growth in the number of colour receivers will no doubt gain further impetus when the total number of programme hours in colour is increased. At the present time in the U.K. the average number of colour programme hours per week is about 30 and this figure will be considerably increased when the BBC-1 and Independent Television programmes are transmitted in colour on 625 lines and u.h.f. in late 1969.

Perhaps the most important engineering task facing broadcasters in this country at the moment is that of providing adequate coverage of the country with u.h.f. transmissions, so that, as far as is feasible, the colour programmes will be available to the whole population.

Programmes are becoming ever wider in scope and present-day engineering objectives include the provision of technical facilities that match up to the requirements of programme organizers and producers. For example, the exchange of colour programmes between countries operating with radically different scanning standards (i.e. 525 lines, 60 fields and 625 lines, 50 fields) has been greatly facilitated by the development of special standards-conversion equipment;[1] there is no doubt that specialized engineering requirements will continue to arise and will doubtless be similarly satisfied.

However, the main purpose of this paper is to attempt to forecast some of the engineering objectives which as yet are not so clearly defined and to indicate some personal views as to some of the technological developments which may make significant contributions to fulfilling these objectives.

2. Future Engineering Objectives

There is no doubt that the future will see a continuation of efforts to provide better technical performance, and considerable effort will be made in this country to ensure that television signals conform closely to fairly rigid specifications; the need for this will be very apparent when several services are operating in colour simultaneously (e.g. BBC-2, BBC-1 and ITV). Although the PAL system does permit relatively wide tolerances with regard to some aspects of the performance of the signal chain, it is considered that this ruggedness should be exploited mainly by the viewer's receiver.

Increasing emphasis will be placed on reliability and efficient operation; although colour television is considerably more complex, in the engineering sense, than monochrome television, it is expected that improvements in equipment will permit colour television equipment to be operated as efficiently as monochrome equipment. Indeed it is possible that some fairly radical changes in television engineering will take place during the seventies and these, if exploited, might well lead to further increases in operating efficiency. In this connection, the time is

† British Broadcasting Corporation Research Department, Kingswood Warren, Tadworth, Surrey.

rapidly approaching when the introduction, where relevant, of digital methods should be seriously considered;[2] these are being introduced into other communication systems and this must lead to their increasing use in television. The conversion of analogue television signals into digital form and vice versa is already quite practicable. The use of digital methods could lead, ideally, to a situation where the equipment in the broadcasting chain is of a form such that when it is switched on it immediately operates correctly, requires neither adjustment nor careful attention and is unusable when faulty.

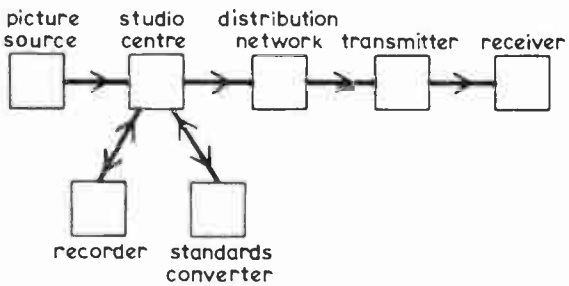


Fig. 1. Basic television transmission chain.

Figure 1 shows the basic parts of a typical transmission chain and the paper outlines in general terms the way in which each of these parts is likely to develop within the next ten years or so and speculates about some of the developments which might prove significant in the future.

3. Cameras

At present colour cameras are characterized by an average performance which permits good-quality colour pictures to be provided to the viewer. However, they require fairly careful adjustment, they can deal with only a fairly limited scene-contrast range and the majority are somewhat restricted in manoeuvreability owing to their size, shape and weight. In general all these limitations are due either to the size or to the sensitivity of the camera tube now used.

In future it is likely that new and small forms of camera tube[3] will be developed and it is hoped that there will also be a significant increase in sensitivity. Given both these developments it should be possible for future cameras to be considerably smaller and far more manoeuvrable, and it is conceivable that their performance may be such as to provide a performance margin which could lead to greater operating efficiency and greater freedom with regard to scene contrast.

Cameras of the future may use solid-state sensors in addition to, or in place of, conventional camera tubes.[4] Figure 2 illustrates the principles of one form of solid-state image sensor in which each basic element consists of a photo-sensitive resistor shunted by a suitable capacitance. Each element of the sensor corresponds to an elemental area of camera-tube target and the action of the scanning beam is replaced in the sensor by a series of switches, operated by a shift register, which cyclically re-charges the elements to a fixed potential. Between scans the action of incident light is to reduce the resistance of an element, thus discharging the associated capacitor to an extent depending upon the intensity of the light. When the element is scanned the recharging of the capacitance causes a signal current to flow whose magnitude is determined by the total amount of light that has

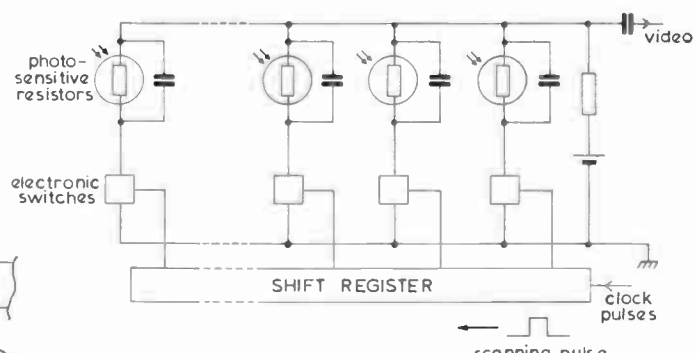
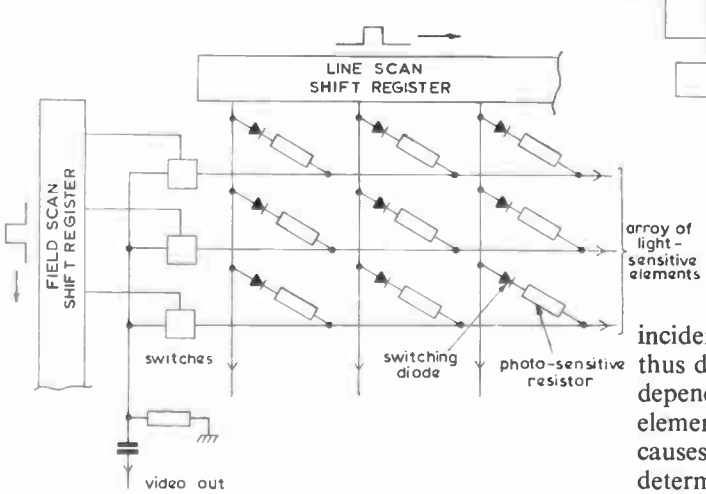


Fig. 2. Principles of solid-state sensor.

Fig. 3 (*below*). Diode switched sensor array.



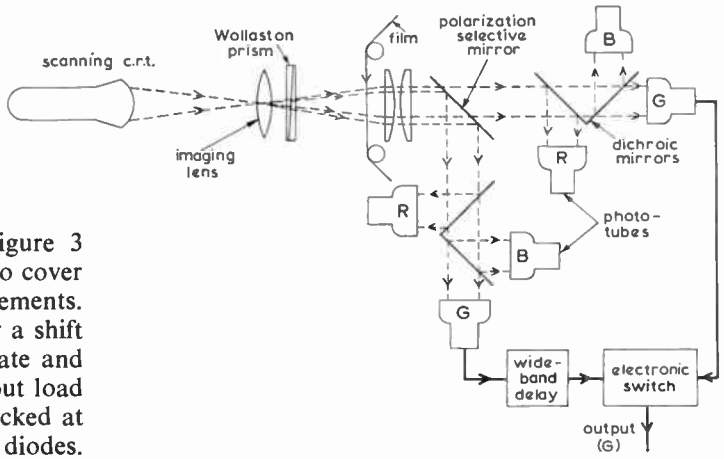


Fig. 4. Elements of flying-spot intermittent-motion film scanner.

fallen upon the element between scans. Figure 3 shows how these principles may be extended to cover a two-dimensional array of photo-sensitive elements. Scanning in the line direction is controlled by a shift register which is clocked at picture-element rate and each line of sensors is connected to the output load in turn by a further shift register which is clocked at line rate and element switching is performed by diodes. It will be seen that by these means the array of photo-sensitive elements is scanned in a manner which is closely analogous to the scanning of a camera-tube target.

4. Film Scanners

The two forms of colour film-scanner, one using camera tubes and the other using the flying-spot principle, should undergo improvement in the next few years.

The camera-tube film-scanner will of course benefit by developments in the camera field, but unless the camera tubes (or other forms of sensor) either become appreciably cheaper or have much longer lives than at present, the running costs of a camera-tube colour scanner may place this form at a disadvantage. With regard to the flying-spot scanner, the next few years might see further improvements in the phosphors used for scanning cathode-ray tubes; in addition photo-multiplier tubes may also improve. In general the flying-spot technique is used, at present, only in those countries using 50 fields-per-second television standards. However, it is possible that flying-spot scanners may be introduced which can be used with both 50 and 60 fields-per-second scanning standards. Figure 4 illustrates the elements of such a scanner. The basic principle consists of arranging that the odd and even field scans of a particular film frame are carried out simultaneously, the film remaining stationary during this interval. The signals corresponding to one of the scans (say the even) are delayed so as to contribute to the output signal at the correct time. Scanning is carried out by two spots which respectively trace out simultaneously, on each film frame, paths corresponding to the odd and even fields. By arranging that the quality of the light from one spot is different from that from the other, the light collected after scanning of the film may be separated into two components and fed to separate sets of photo-multipliers. In Fig. 4 the light from the single scanning spot at the screen of the cathode-ray tube is split into two orthogonally polarized components by the bi-refringent Wollaston prism†; this prism is also used to space the two images of the spot correctly at the film. The polarized components of the light passing through the film are separated by a suitable mirror and fed to separate groups of photo-multipliers.

The outputs from one group of photo-multipliers are fed to wideband delays (of the order of one field period) and then combined with the signals from the other group by means of electronic switches. Suitable operation of the switches results in continuous signals of the correct form.

It can be shown that these arrangements allow the film to be moved intermittently at 25 frames per second (for 50 field television) and 24 frames per second (for 60 field television) whilst permitting adequate time for pull-down in each case.

5. Studio Centres

It is likely that the technical facilities at studio centres will be improved in ways which may involve fairly radical changes. Significant developments are likely in signal switching systems, recording and standards conversion and, in the main, these developments are likely to exploit advances in digital signal processing and storage.[2]

5.1. *Signal Switching*

Figure 5 shows one possible way in which signal switching systems may evolve. In such an arrangement coded instructions are added to the signal at or near its source and these instructions, which may well be digital in form, are read at various switching points along the intended route of the signal where the

†The use of a Wollaston prism for this purpose was suggested by G. D. Monteath.

required operations are carried out. On reaching the intended destination the coded instructions are deleted and the signal reverts to its standard form. Such a system, of course, will involve other arrangements which, for example, ensure that the optimum route is chosen and that no difficulties arise due to more than one signal competing for the same circuit at the same time. A further development of such a

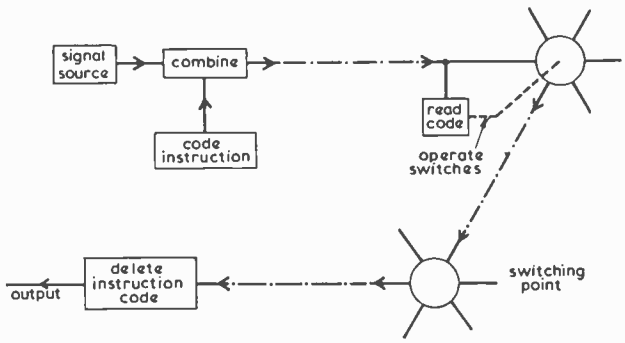


Fig. 5. Simple automatic signal switching system.

system might occur if and when the video signals are available in digital form; in such circumstances the problems of maintaining correct timing between signals from different sources would be simplified and the need for equalizing the transmission characteristics of individual signal paths within the studio centre might be eliminated. In general these exploitations of digital signal transmission and control would be well matched to any computerized system of programme administrative and operational control.

5.2. *Recording*

A very considerable amount of work is being carried out in various parts of the world in order to provide digital backing stores of very large capacity for general-purpose computers and, if successful, the results of this work might provide means for recording television signals in digital form in the future. The digital recording of television would provide very significant advantages; in principle, recordings would be replayed and copied without loss of signal quality and timing error problems could be dealt with quite easily.

Several media are potentially suitable for digital recording and these include magnetic tape and photographic film. The present-day capabilities of high-density magnetic recording systems are such that for television, 2-inch (5·1 cm) tape would be used at about 200 cm/s (80 in/s). However, it is possible that future developments will reduce the rate of tape consumption to a value approximately equal to that characterizing present-day analogue recorders.

Photographic systems are being developed using electron-beam and laser recording techniques. However considerable work remains to be done in this field, particularly in connection with reading high-density recordings of this type.

Recent proposals for the recording of digital information have included the use of a laser, modulated by digital signals, whose output is focused to an extremely fine spot which punches minute holes in the opaque coating of an otherwise transparent tape.[5] The same basic arrangements may be used for both recording and replay. In the recording mode the laser is operated at high peak output and a photo-detector may be used for monitoring purposes. In the replay mode the laser is operated continuously at relatively low power, so as not to impair the surface of the tape, and the light passed by the holes in the opaque tape-surface is converted, by the photo-detector, into the required digital output signals. The tape can be scanned as in a helical magnetic recorder but the recording head merely consists of an optical arrangement focusing the laser beam on to the surface of the tape.

5.3. *Standards Conversion*

Standards converters basically consist of an assembly of stores together with suitable switching arrangements, and in current equipment the signals are stored in their conventional or analogue form. In present-day line-store converters information corresponding to individual picture elements is stored in lumped networks or capacitances (equal in number to the number of elements per line) and impairment of the output picture results from small differences in the characteristics of these stores. In a current field-store converter storage is provided by means of delay lines and, in the converter, a group of these lines is suitably switched so as to subject the signal to a cyclically varying delay. A small distortion of the signal, together with the addition of noise and other forms of interference, occurs during each passage of the signal through a delay line, with the result that the output picture may be subject to cyclically varying impairments.

However, with television signals available in digital form, standards converters may be developed which will be free of impairments of this nature. In a digital line-store converter the signal, in digital form, would be written into a store having a minimum capacity corresponding to the information contained in one input line; new information would be continuously written into the store and would replace old information by 'overwriting'. The information would

be read out at a rate appropriate to the output standard. As in a present-day line-store converter, interpolation would be required to give satisfactory portrayal of sloping features in the picture. This could be achieved by increasing the capacity of the store so as to make available, simultaneously, information describing corresponding picture elements on a few adjacent lines of the same input field; a store with a much larger capacity would permit access to corresponding points on adjacent picture lines with a consequent improvement in interpolation. The process of interpolation would be performed on groups of digits representing the vertically adjacent picture elements by means of digital multiplication and addition. Figure 6 shows, in essence, how these principles could be exploited; this figure however would also be applicable to one possible form of

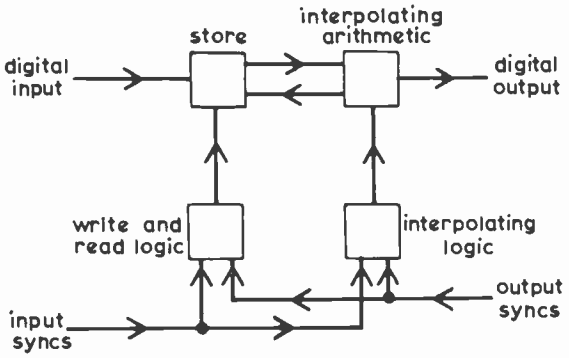


Fig. 6. Digital standards converter.

digital field-store standards converter. In such a field-store converter the store might have a total capacity corresponding to two or three fields of the input standard. Input digital signals would be written into the store, overwriting previous signals, and the store would be organized so that easy access could be simultaneously obtained to corresponding elements on adjacent picture lines. This could enable interpolation to be carried out, not only for the purpose of substantially eliminating distortions at sloping picture features, as in a line-store converter, but also to minimize the impairments due to movement in the picture; the fact that field-store conversion increases or decreases the number of fields per second must, in principle, lead to discontinuities in the portrayal of movement, and the effects of this may be reduced to a very considerable extent if some form of movement interpolation is incorporated in the converter. Groups of digits representing vertically adjacent picture elements on successive picture lines could be fed to interpolating arithmetic which would then return interpolated data back to the store. Reading of the store would then be carried out at a rate appropriate to the outgoing standard. Logical control circuits would be needed for both the store and the interpolation arrangements; these would derive their input information from the synchronizing pulse trains of both the input and output standards.

6. Links

Digital transmission systems have now been introduced to an increasing extent by communications authorities. In this country p.c.m. telephony systems have already been installed by the G.P.O. and it seems likely that these will gradually replace the existing systems based on frequency-division multiplex. It is possible that high-capacity digital networks will be installed throughout the country and these will carry telephony, provide access to large central computers etc. and they may well be capable of handling television signals in digital form. If television signals are distributed throughout the country by means of p.c.m. one may look forward to the situation where the signals fed to transmitters in remote parts of the country are of a quality which is indistinguishable from that at the source.

Developments broadly along these lines are fairly well advanced in the U.S.A. where extensive tests have been carried out on a 224 megabit-per-second system in which digital signals are carried in a relatively inexpensive cable fitted with closely spaced repeaters. Further in the future, higher bit rates may be possible using the transmission of light pulses through glass-fibres. Figure 7 illustrates how a laser may be modulated by p.c.m. signals so as to provide a modulated light beam for such purposes.[6] The laser is pulsed at

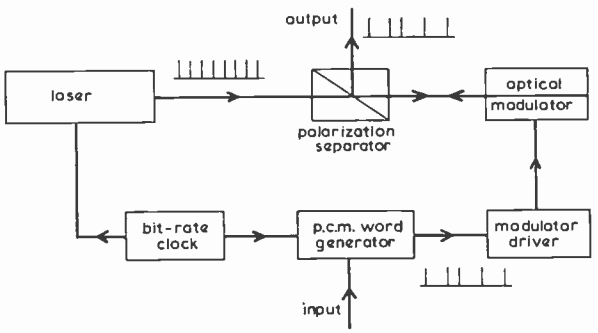


Fig. 7. Optical modulator for p.c.m.

the regular bit rate of the system and is modulated by a passage through a crystal whose optical properties can be changed by varying an applied electric field. Thus by using a polarization separator it is possible to provide light output which is modulated according to the electric field applied to the crystal.

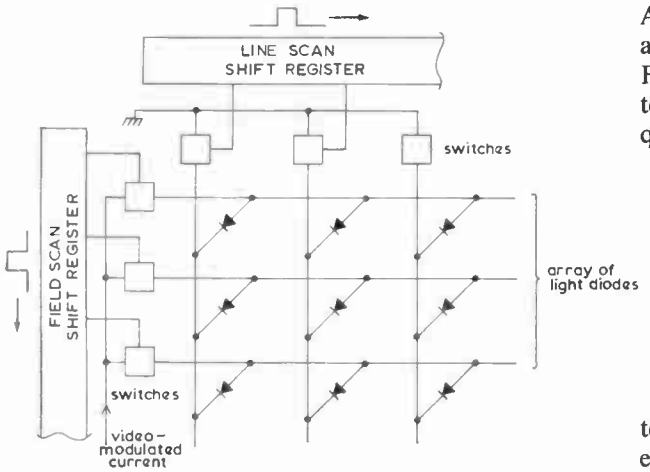


Fig. 8. Elements of solid-state display panel.

Satellite links in the future may well use digital transmission systems as these offer the possibility of relatively reduced satellite transmitter power.

7. Transmitters and Receivers

Radical changes in the nature of terrestrial broadcast transmitters are rather unlikely for a considerable time to come. There are several reasons for this, which include the need to make the most effective use of restricted channel space and, perhaps most importantly, the difficulty of making a significant change in the nature of television broadcast transmissions when many millions of receivers are in the hands of the public.

The advent of direct broadcasting from satellites, however, raises the possiblity of introducing a form of digital broadcast transmission. Receivers for such transmission will probably require new aerial systems and modified tuners; it might be possible to introduce cheap digital-to-analogue converters at the same time.

One significant change in the domestic television receiver may take place with the development of suitable solid-state display panels. Figure 8 shows the elements of such a panel suitable for monochrome operation. It consists of a large two-dimensional array of light-emitting diodes together with a suitable arrangement of switches and shift registers which scan the array. Each line of diodes is switched by the field-scan shift-register and during each line-period diodes corresponding to individual picture elements are switched by the line-scan shift-register. The amount of light emitted by each diode is determined, in this rather hypothetical arrangement, by the current flowing through it.

8. Conclusions

The next ten years or so will see the introduction of new technologies which must lead to very radical changes in a large part of the television signal chain. As a result picture quality will become very consistent and probably of an even higher standard than today. For example, one can look forward to world-wide television interconnexions with negligible loss of quality in spite of differences in basic scanning standards.

One principal consequence of the improved techniques must be a more efficient use of television engineering resources. Initially the necessary equipment may well be appreciably higher in capital cost than the equipment it is replacing; this is quite a common state of affairs when innovations occur. However, these high costs are likely to be more than compensated by great reductions in running costs.

9. Acknowledgments

The author wishes to express his gratitude to his colleagues who helped him in the preparation of this paper and to acknowledge helpful discussions with Messrs. G. D. Monteath and E. R. Rout. He is indebted to the Director of Engineering of the B.B.C. for permission to deliver this lecture and to publish it.

10. References

1. Rout, E. R. and Davies, R. F., 'Electronic standards conversion for transatlantic color television', *J. Soc. Mot. Pict. Televis. Engrs*, 77, p. 12, 1968.
2. Lord, A. V., 'Digital methods applied to television', *J. Roy. Telev. Soc.*, 12, No. 2, pp. 27–36, Summer 1968.
3. The Norelco 'Mighty Mite', discussed in 'New Products', *Electronics*, 41, No. 10, p. 138, 13th May 1968.
4. Weimer, P. K., *et al.*, 'A self-contained solid-state image sensor', *Proc. Inst. Elect. Electronics Engrs*, 55, No. 9, pp. 1591–1602, September 1967.
5. Becker, C. H., 'Unicon computer mass memory system', Proceedings of the Fall Joint Computer Conference, San Francisco, November 1966.
6. Denton, R. T. and Kinsel, T. S., 'Terminals for a high-speed optical pulse code modulation communication system. I. 224-Mbit/s single channel', *Proc. Inst. Elect. Electronics Engrs*, 56, No. 2, pp. 140–5, February 1968.

Manuscript received by the Institution on 15th October 1968. (Paper No. 1258/Com. 16).

U.D.C. 621.397.621

Television Receivers—A Survey of Current Design Trends

By

PETER L. MOTHERSOLE,
C.Eng., F.I.E.E., F.I.E.R.E.†

Based on a survey paper presented at the International Broadcasting Convention held in London on 9th to 13th September 1968.

Summary: The main features of domestic monochrome and colour receivers are surveyed, particularly with respect to the larger screen size. Circuit technology has reached the stage at which integrated circuits are starting to challenge the transistor while in certain applications the valve still has a major advantage, mainly due to the scanning power required by the picture tube.

1. Introduction

The modern television receiver is capable of producing a very satisfactory picture in a wide variety of reception conditions. Furthermore the controls have been simplified to the point where operation requires little more than switching on, selecting the required station and adjusting the sound volume level and possibly the brightness. In addition to the channel selector most receivers have front-mounted controls for brightness and/or contrast and saturation in the case of colour receivers (plus a hue control with N.T.S.C.) together with a volume control and an on/off switch. Reception can be at u.h.f. and v.h.f. and the only noticeable difference from the viewer's point of view is in the aerial complication. Wired systems are employed with conventional or simplified receivers operated at h.f. as well as v.h.f. Switching from v.h.f. to u.h.f. automatically changes standards from 405 to 625 lines in the case of the most recently-produced U.K. receivers.

Monochrome receivers are produced with a wide variety of picture tube sizes ranging from a few inches to 25 inches or so. Small-screen all-transistor receivers are mainly confined to portable types where battery operation is required. The most popular sizes of mains-operated receivers, particularly in Europe and the U.S.A., are in the range 16 to 23 inches for monochrome. For colour receivers the most popular tube size at the moment appears to be the 25-inch with slightly smaller tubes finding increased favour. The brightness available from modern picture tubes means that operation is possible with reasonably high ambient illumination levels and little concession has to be made in the room lighting requirements for television reception.

Receiver performance has reached a high standard both in terms of reliability and displayed picture. After price, however, receiver styling and ease of use are probably the most significant factors from the viewer's point of view. For these reasons, particularly in the case of colour, automatic frequency control is becoming increasingly popular together with pre-set tuning.

2. Active Devices—Thermionic versus Semiconductor

In the signal amplifier circuits transistors can give improved noise performance, notably at u.h.f., and varicap diodes simplify the mechanical aspects of the tuner. The video and signal processing circuits are again those that lend themselves to transistor circuitry, particularly in the case of colour receivers.

In Europe virtually all receivers now being made are hybrid, transistors being used for the small signal circuits and valves retained for the deflection and output circuits only. The combination of transistors and thermionic devices poses many problems to the receiver designer especially in the video section where picture tube flashover is a hazard.

The power required by a 110° monochrome line time-base and 20 kV e.h.t. supply is some 35 watts for a valved circuit and it falls to about 24 watts when transistors are used. For field deflection circuits the corresponding figures are 12 and 9 watts respectively.

The volt-ampere product for the line circuit is about 3500 VA. To minimize the circulating current, high voltage operation is normally employed with the valve operated in a boosted h.t. circuit with a peak potential of some 6 kV applied to its anode. The voltage capability for transistors is much lower and hence such a circuit configuration is not possible. The highest collector potential that is currently envisaged for a line transistor such as the BU105 is 1·5 kV and hence operation is limited to an h.t. line of about 150 V.

In most European countries the mains potential is about 220 V and a simple half-wave rectifier type of power supply, as currently used in television receivers,

† Central Application Laboratory, Mullard Limited, Mitcham, Surrey.

produces a smoothed h.t. potential of about 250 V. Since the line time-base circuit has the major power requirement, the generation of a special 150 V supply imposes a major cost penalty. This is to some extent offset since the BU105 transistor collector/base junction is used as the parallel efficiency diode. However, the cost of the high voltage drive transistor with its associated circuit imposes a further cost increase compared to a valved circuit. It is apparent therefore that valves are likely to be retained in the deflection circuits for some time to come. When a line transistor with a peak potential rating of 3 kV is available this will change the situation since a simple power supply can be retained. However, even then stabilization against mains potential variations with a simple feedback circuit is not possible due to the limited dissipation of transistors.

Similar considerations apply to the audio output stage and field time-base where in general the valved circuit has a slight advantage. Until the complete receiver can be economically transistorized the tendency to retain valves for all the power output functions is therefore likely to continue.

The power required for the transistor signal amplifying and processing circuits in a hybrid receiver amounts to only a few watts. Operation is therefore possible by resistive dropping from the main h.t. line or from a potential developed by rectification of the line scanning waveform.

In countries employing a mains potential of about 117 V a simple half-wave rectifier type of h.t. supply produces a smoothed h.t. line of about 135 V, hence similar considerations may not apply. However, the overall economics still appear to make a hybrid receiver a viable proposition.

The above figures for the line time-base power requirements are about doubled for a 90° colour receiver and hence an all-transistor receiver is likely at first sight to be even less economic. However, other factors influence the decision. Firstly, the cost of semiconductors relative to the complete receiver cost is lower for colour than for monochrome and furthermore colour television is only just starting in Europe and hence production quantities are small. Secondly, the further reduction of heat resulting from the use of semiconductors eases layout problems and leads to greater overall stability and reliability. Hence transistors are used to a greater degree in colour receivers and indeed two all-transistor receivers are in production, one in the U.K. and the other in the U.S.A.

In an attempt to make the valve even more attractive American manufacturers have produced a range of triple and multiple-function valves called compactrons. These valves combine such functions as: line oscillator and phase detector; a.g.c., video output and sound i.f. amplifier; horizontal output and booster diode etc.

Although the compactron appears to have been reasonably successful in that hybrid receivers have only been produced in small quantities in America, indications are that the situation is changing. The situation is likely to be further complicated by the introduction of integrated circuits. Power and frequency limitations, however, confine their use to the signal amplifying or processing circuits and hence hybrid receivers have been produced employing valves, transistors and integrated circuits.

3. Tuners and I.F. Amplifiers

European multi-band receivers often use two tuners. The u.h.f. tuner, which consists of an r.f. amplifier and self-oscillating mixer, feeds into the mixer input of the v.h.f. tuner to equalize v.h.f. and u.h.f. receiver gains. The overall gain of such a transistor arrangement is typically 30 dB with a noise figure of 6 dB at v.h.f. and some 8 dB at u.h.f. When an integrated or all-band tuner is used a similar performance is obtained but with fewer devices.

In the U.S.A. valves are normally used in the v.h.f. tuner and the u.h.f. tuner consists of a diode mixer and transistor oscillator preceded by a tuned bandpass circuit. Since the gain of this tuner is very low its output feeds into the v.h.f. tuner. The noise performance of such a front-end is poor but satisfactory reception is provided in reasonably strong signal strength areas. The main advantage of such an arrangement is, of course, the very low cost of the u.h.f. tuning facilities.

Considerable effort is now being concentrated on electronically-tuned tuners employing varicap diodes. Many large setmakers employ a standard chassis which is used in a wide variety of cabinet constructions and varicap tuning enables the tuner to be made integral with the i.f. amplifier with only d.c. connections to the station selector on the cabinet.

At u.h.f. adequate coverage of Bands IV and V can be obtained without switching but the varicap losses (series resistance) reduce the circuit Q-factors and modified circuit arrangements are often necessary. The image rejection of a typical varicap tuner, as shown in Fig. 1, is about 30 to 40 dB. Such tuners can be used in countries where there is only a single u.h.f. programme, such as Germany, but for use in the U.K. where it will be required to receive four main programmes in an 80 MHz channel space, higher image rejection and station selectivity is essential, the recommended figure for image rejection being 53 dB.

The main problem in a design of a v.h.f. varicap tuner is that of providing coverage of Bands I and III

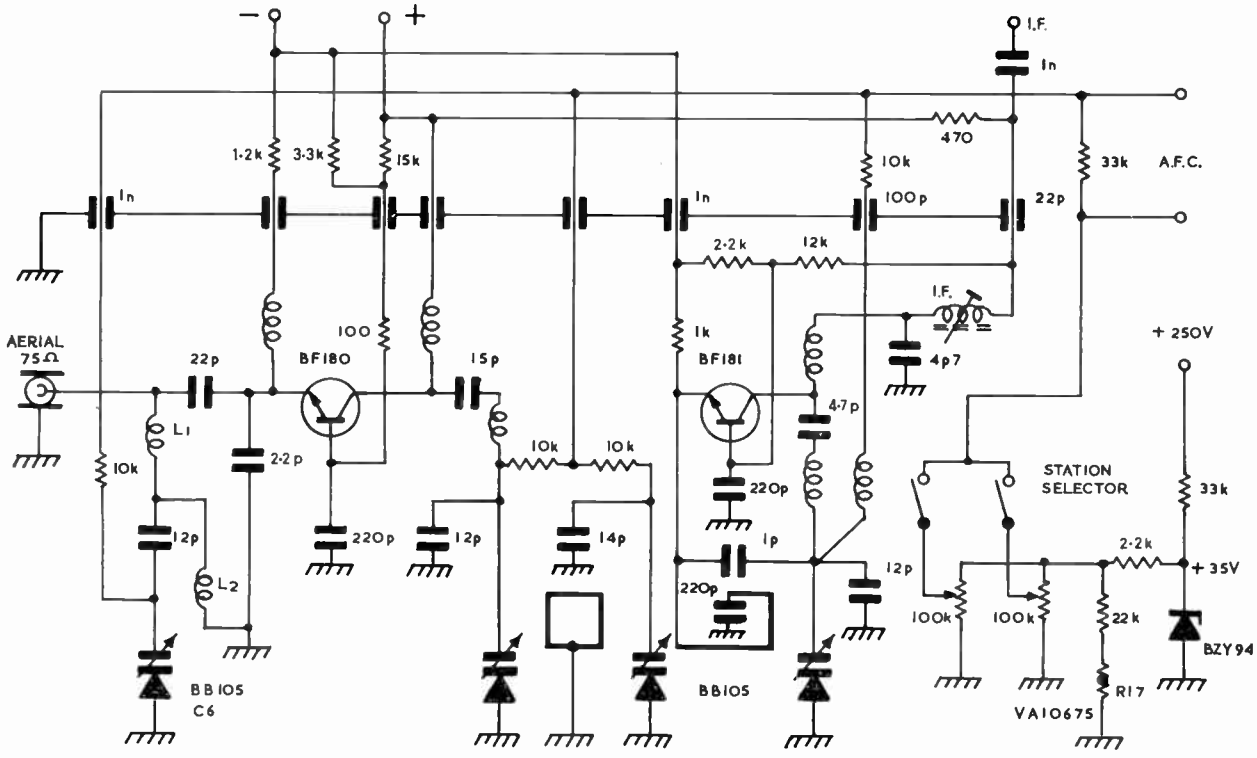


Fig. 1. U.h.f. varicap tuner.

without the use of mechanical switching. Switching diodes can be used but the overall cost of the varicap tuner is then relatively high since three or four varicaps are required in addition to some five or six switching diodes. Electrical performance of the diodes, however, does not impose any limitation to the tuner performance.

The d.c. supply for the varicaps must be stabilized and normally a Zener diode is used with further temperature compensation on u.h.f. (see Fig. 1). With such a combination the frequency stability is typically 250 kHz for a 10% mains potential variation and 0·5 MHz for 20°C overall temperature variation of the tuner and stabilizing circuit.

The frequency stability for colour receivers requires to be higher and a.f.c. is required. This can be very simply applied to a varicap tuner by connecting the output of the frequency discriminator in series with the tuning potential. In this way all the tuned circuits are corrected by the a.f.c. potential.

Vision i.f. amplifiers normally employ three stages with a.g.c. applied to the first stage. Typically 40 dB of control is obtained in the i.f. amplifier with a further 40 dB in the r.f. amplifier.

To minimize beat patterns and ease the fine tuning requirements, particularly in colour and better quality monochrome receivers, two detectors are often used. The luminance signal is produced in one detector, chrominance and sound signals in the case of colour receivers from the other. When a.f.c. is incorporated a further additional limiting amplifier is used to drive the discriminator which is tuned to the vision carrier frequency.

Apart from receivers designed to operate with an a.m. sound system (e.g. U.K. 405-lines, France and Belgium), all domestic receivers use the intercarrier sound system. It is in the sound i.f. amplifier where one of the major differences occurs between the expensive and cheap receiver. In the cheap receiver a single i.f. amplifier stage is normally used with the intercarrier signal being taken from the anode of the video output valve. In the more expensive receivers two i.f. stages are used, the input being taken from a separate detector diode. In the U.S.A. the locked oscillator type of demodulator is popular, the high-level output driving the output pentode directly.

Integrated circuits first made their appearance in consumer equipment as the intercarrier sound i.f. amplifier. The integrated circuit enables many cascaded circuits to be used and hence a high degree of a.m. rejection is obtained combined with a high input sensitivity.

4. Automatic Gain Control and Synchronization Circuits

The circuits performing these functions are often colloquially described as the 'jungle'. This is because a wide variety of various circuit techniques have been evolved over the years and probably the greatest variation between receivers is to be found in this area. The reason for this is that whereas the tuner, i.f. amplifier, and time-bases fulfil fairly specific and well-defined functions, the signal processing and control circuits are very dependent upon the system of transmission used and the degree of noise immunity required and they give scope for circuit innovation.

In recent years the variety of circuits has reduced as optimum performance has been evolved, often due to the special devices developed for a specific circuit function. Before reviewing the more popular receiver techniques it is probably profitable briefly to examine some of the differences incurred due to the polarity of vision modulation used.

Before the war, when the basic parameters of the British 405-line system were being evolved, positive modulation was adopted principally because of the ease of synchronization that resulted since interference pulses were in the direction of the video information leaving the synchronizing waveform relatively clean. Automatic gain control, if required, was possible by utilizing the residual line blanking or back porch following the line synchronizing pulses. Some form of gate circuit, however, would be required in the receiver. In the U.S.A. negative modulation was adopted principally because of the ease with which a.g.c. could be incorporated. The synchronizing pulse tips, corresponding to 100% transmitter output, provide a reference signal with each line which can be utilized by a simple a.g.c. system. With negative modulation noise pulses often exceed the amplitude of the synchronizing pulses and the difficulties that were encountered with synchronizing were overcome by circuit techniques. Other parameters also influenced the decisions but in principle the main criteria appear to be the ease of synchronization versus the simplicity of a.g.c.

In the event automatic gain control was not fitted into U.K. 405-line receivers until about 1953 when the second programme in Band III was introduced. The early a.g.c. systems utilized the back porch of the video waveform for the a.g.c. reference and hence the a.g.c. circuit was fairly complex in that a gating pulse was required to enable the black level of the video signal to be measured line by line. These gated a.g.c. circuits gave rise to many operational difficulties, principally due to the difficulty of servicing and when mean level a.g.c. was devised gating systems fell into disfavour.

In a valve television receiver the video signal is negatively d.c.-restored at the control grid of the sync. pulse separator. The mean level of this signal can be utilized for a.g.c. purposes, the only additional components being one or two resistors and a capacitor. The major disadvantage of mean level a.g.c. is that it suppresses the d.c. component of the video signal and hence receivers employing mean level a.g.c. are characterized by the absence of the d.c. component in the displayed picture.

With negative modulation on the other hand a true a.g.c. potential can be obtained from the video signal with a simple peak detector measuring the tips of the synchronizing pulses. Unfortunately such a simple system runs into difficulties with interference pulses since the a.g.c. potential tends to ride up on the tips of the interference and the video signal tends to become suppressed. Hence receivers utilizing negative modulation normally have some form of gate circuit which operates during the line synchronizing period.

At first sight it would appear that mean level a.g.c. would be equally applicable to a negative modulation receiver and result in some component saving. However, difficulties can occur in that with positive modulation the video signal applied to the video amplifier tends to drive the valve into conduction producing a negative-going swing at its anode which is applied to the a.g.c. system and is in the correct sense to restore control. With negative modulation, on the other hand, the video valve is driven to cut off and no signal is produced at its anode or at the grid of the synchronizing pulse separator and hence no control potential is produced and a condition of lock-out results. Furthermore, with negative modulation, in the event of the video signal varying appreciably with the mean level of the transmitted signal, the tendency exists for crushing of the sync. pulses to occur and poor synchronization results. Hence negative modulation receivers tend to employ an a.g.c. circuit that operates on the tip of the synchronizing pulse and therefore the video signal at the video amplifier is proportional to the true value of transmitted signal.

In dual-standard receivers mean level a.g.c. is almost invariably used to minimize the switching problem. To prevent lock-out in the case of valve receivers at least part of the a.g.c. potential is derived from the video detector circuit. In the case of hybrid receivers, the a.g.c. system is low level and d.c.-coupled to the detector circuit. In colour receivers mean level a.g.c. is less satisfactory since if the picture tube is driven into grid current on low key scenes the grey scale tracking may be disturbed producing colouration after peak white signals and hence true sync. tip a.g.c. tends to be used.

In a domestic television receiver the e.h.t. potential is generally obtained by rectifying the flyback pulse generated in the line-scan circuit. Normally such e.h.t. systems are relatively high impedance and therefore the d.c. component of the video signal can cause appreciable picture breathing, i.e. variation of size due to variation of e.h.t. potential. For these reasons it became common practice to attenuate the d.c. component in the displayed picture by means of an a.c. coupling to the picture tube.

Operationally, receivers that employed a.c.-coupled video stages found favour in that the picture appeared to require less adjustment, particularly when changing channels. It must be appreciated that when the d.c. component is fully retained in the displayed picture any variation of the black level at the transmitter or between transmitters results in a significant shift of the receiver brightness. Although it may be argued that this can be corrected by a simple adjustment of the brightness control, in practice viewers generally find the picture is more acceptable when it has a constant average brightness level. Measurements of the mean value of transmitted waveforms over long periods indicate low values of d.c. component only occur for some 10% of the time and much of this time the picture is non-critical, i.e. the displayed picture may be just a caption. This topic has been the subject of much discussion and circuits have been devised to retain a measure of d.c. component in the picture, at least for low key scenes.

In a colour receiver maintenance of the d.c. component is much more important, particularly in case of colour difference drive, since here an a.c. coupling will result in complementary colours being produced under certain conditions. Hence even if the d.c. component of the luminance channel is reduced almost invariably the d.c. component of the colour difference channel is fairly accurately maintained.

The sync. separator circuit associated with a positive modulation receiver almost invariably consists of a valve (or transistor) with the video signal a.c.-coupled to its control grid and negatively d.c.-restored by grid current. The d.c. operating conditions are so arranged that the valve can only conduct on the synchronizing pulses and hence separated synchronizing pulses are produced in its anode circuit. These are then separated and applied to the appropriate time-bases.

With negative modulation noise protection of the synchronizing channel is essential and the most commonly-used circuit consists of a heptode valve. The video signal is a.c.-coupled to the second control grid and negative-going noise pulses are applied to the first control grid. The pulses at the first grid cut the valve off and hence prevent current flowing and noise outputs from appearing in the anode circuit. This function is usefully performed in many American receivers by means of a dual heptode with a common first control grid. One half of the valve is used as the a.g.c. gate, the other as the synchronizing pulse separator.

A wide range of vertical synchronizing pulse separators exist, their main function being to separate the vertical pulse so that the field time-base is triggered accurately at the end of each scanning stroke.

Line synchronizing is almost always effected by means of a flywheel system. In such a system the synchronizing pulses together with a reference waveform from the output stage are applied to a phase detector. A d.c. potential is produced, the value of which depends on the phase of the two waveforms and this potential, after smoothing, is used to control the frequency and phase of the horizontal oscillator. With such a system good frequency stability of the horizontal oscillator is essential and this is normally obtained by including a stable tuned circuit as part of the oscillator timing mechanism. In the U.S.A. multivibrators with sine-wave stabilization tend to find favour but in Europe the Class C sine-wave oscillator is more popular.

Continental receivers are often required to receive signals from very distant transmitters and hence the line and field synchronizing circuits are required to produce steady pictures under extremely adverse operating conditions. 'Automatic' or two-mode synchronization systems have become fairly popular. With such systems the synchronizing system has a fairly wide pull-in range but once the time-base has become synchronized the range is reduced to produce a corresponding improvement in the noise performance.

5. Video Output Stage

The video output stage fulfils the difficult function of linking the low-level transistor circuits to the thermionic picture tube which requires a drive of some 80 to 120 V. In the case of a valve receiver the synchronizing and a.g.c. circuits are normally driven from the anode of the video amplifier together with the cathode of the picture tube and hence the video amplifier has a relatively high value of capacitive load. In the case of a hybrid receiver the a.g.c. and synchronizing circuits are invariably driven at low level from the video pre-stage and hence the video output stage has only the picture tube cathode as its load and the collector load resistor can be quite high in value, typically 4·7 kΩ, and simple emitter compensation can be used. A typical circuit is shown in Fig. 2.

A major hazard to the video output transistor is transient energy that is produced at the cathode of the picture tube in the event of an internal flashover.

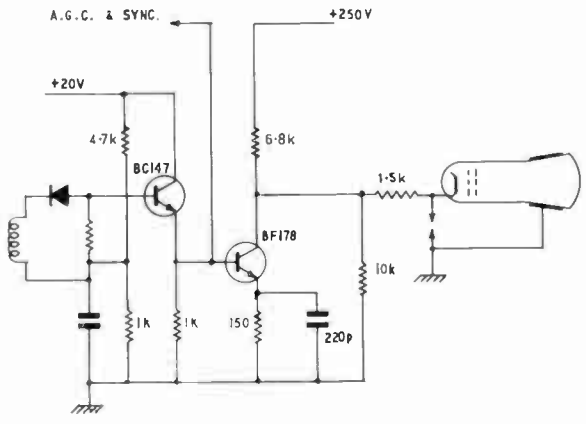


Fig. 2. Typical video amplifier.

In fact the discharged energy can prove a hazard to other semiconductor devices unless precautions are taken to prevent the associated circulating currents flowing through the receiver circuits. The normal protective measures consist of spark-gaps mounted around the tube electrode pins with stand off resistors to the driving circuits. A low inductance earth return is then used between the cold-end of the spark-gaps and the external conductive coating of the picture tube. With such an arrangement the transient current flows in the low inductance loop and virtually no energy is applied to the video or receiver circuits. This is illustrated in Fig. 3.

6. Colour Signal Processing Circuits

In order to display a colour picture it is necessary to decode the chrominance information to obtain the colour difference signals and these must be combined with the luminance signal to produce the individual RGB signals. One of the fundamental principles embodied in all broadcast colour systems is that of transmitting a wideband luminance signal with narrowband colour difference information. Furthermore, on monochrome scenes the R, G and B signals are equal and any departure represents colouration of the picture. Differential drifts between the video amplifiers in early colour receivers resulted in a serious instability of the grey scale and a significant step forward was made with the innovation of colour difference drive. With this system a common wideband luminance signal is applied to the three cathodes of the picture tube and the individual colour difference signals are applied to the grids. Provided that the black level of the colour difference signals at the control grid of the picture tube is clamped, a stable grey scale results and the individual linearities and bandwidths of the colour difference channels are relatively unimportant. The amplitudes of the colour difference signals require to be approximately twice that of the luminance signal and hence typically over 200 V peak to peak (B-Y) signal is required.

When a transistor colour receiver is considered RGB drive finds favour in that video transistors capable of producing a video signal of over 200 V have not been readily available in commercial quantities. Furthermore transistor amplifiers do not have the difficulties encountered in their valve counterparts due to variation of heater potential and cathode emission. Hence, in Europe, where colour television has only just started and hybrid receivers are popular, the use of RGB drive has found favour.

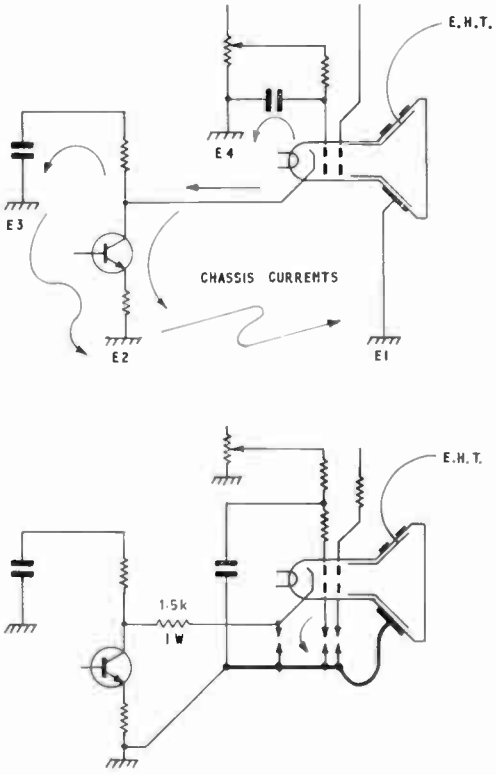


Fig. 3. Transient current paths in picture tube circuits.

To demodulate the colour difference signals a reconstituted sub-carrier is required and normally a crystal oscillator is used for this purpose, phase-locked to the transmitted colour burst by means of an a.p.c. system. Alternatively a crystal ringing circuit is employed with the colour burst applied directly to the crystal. To minimize variations in saturation with fine tuning or propagation variations, automatic chrominance control is almost always employed. The amplitude of the colour burst is used for reference so that its amplitude and hence that of the chrominance signals is held constant at the output of the gain controlled chrominance amplifier.

In a receiver for the PAL system there are additional circuit functions to those found in an N.T.S.C. receiver. These are the delay line driving circuit and the bistable switch that is driven at line frequency and must be synchronized to the corresponding switch at the transmitter. This synchronization is effected by alternating the phase of the transmitted colour burst by $\pm 45^{\circ}$ on alternate lines and utilizing the 7·8 kHz signal developed in the receiver a.p.c. circuit to synchronize the PAL switch.

Many of the circuit functions in the colour demodulator area are eminently suitable for integrated circuit techniques since the voltage and power levels are relatively low. This new technique permits the economic use of advanced circuits and furthermore simplification in assembly is possible since the number of components is reduced. In an RGB system, for example, there is a considerable advantage in using a common integrated circuit for the early stages of the amplifiers in that the three channels are thermally linked together and accurately matched. Hence fully d.c.-coupled colour demodulators and video drive systems are practicable employing integrated circuit techniques in conjunction with transistor video output circuits. Such a system is shown in Fig. 4.

7. Scanning and E.H.T. Generating Circuits

The vertical time-base in a valve or hybrid receiver has changed very little over the years. The most commonly used circuit uses a triode oscillator driving a pentode output valve that is transformer-coupled to the scanning coil. Of the wide variety of vertical oscillator circuits that have been evolved the most popular appears to be the back-coupled arrangement in which the flyback pulse produced at the anode of the output valve is used to trigger the oscillator or discharge valve.

A similar configuration has been used with all-transistor receivers but the single-ended push-pull circuit has been evolved in an effort to eliminate the bulky output transformer.

The line output circuit with a series-connected energy recovery diode is almost universally employed. E.h.t. for the picture tube is generated by means of an overwind on the line output transformer and potentials of 18 to 20 kV are currently used for 23-inch monochrome receivers. In the U.K. voltage doubling, using semiconductor diodes, has been used in an effort to simplify the line output transformer.

In Europe the stabilized line output circuit is invariably used. With this circuit a pulse obtained from the scanning transformer is rectified by means of a non-linear resistor and the resulting direct voltage used as bias for the output valve. The feedback system is so arranged that the amplitude of the flyback pulse is maintained constant in spite of variations in the h.t. potential and output valve ageing. A further advantage of this circuit technique is that the output valve operates above the knee in its anode characteristic and hence generation of Barkhausen oscillation is prevented.

Similar techniques have been used for colour receivers where a mean load current of 1·5 mA at 25 kV is required together with a low source resistance (less that 2 MΩ). The conventional method used to obtain it is with a shunt regulator valve. The x-radiation hazard in addition to the high power dissipation and size of the e.h.t. box are the principal disadvantages of the system. Many continental receivers employ a separate e.h.t. system to avoid the use of a shunt regulator. A recent technique evolved in the U.K. to overcome these problems is to employ a voltage tripler circuit with semiconductor diodes.

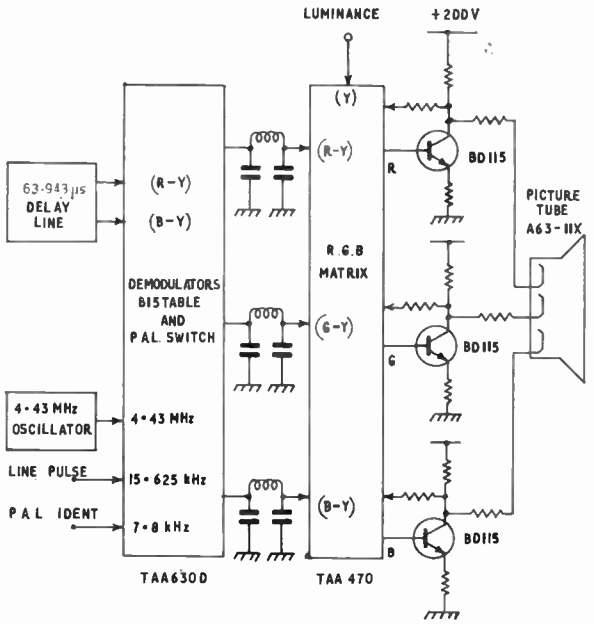


Fig. 4. RGB drive system using integrated circuits for colour signal demodulation and feedback amplifier matrix.

The basic circuit is shown in Fig. 5. The transformer is only required to produce 8·5 kV pulses and hence only a small overwind is necessary. This reduces the leakage inductance (and also ringing) and with the simple voltage feedback circuit, as is used in monochrome receivers to maintain a constant amplitude of flyback pulse, a low source resistance is obtained. A further advantage of this circuit configuration is that the line output valve only operates at its peak current rating at maximum e.h.t. load due to the feedback circuit.

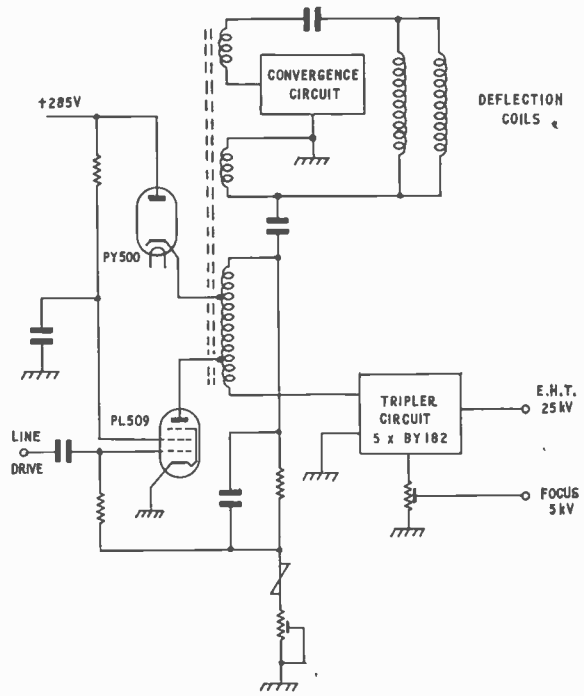


Fig. 5. Basic tripler line output circuit.

In colour receivers using smaller picture tubes than the normal 25-inch reasonably satisfactory e.h.t. performance is obtained by retaining the full overwind and relying on the voltage feedback circuit to provide a relatively low output impedance.

In the case of semiconductor circuits peak voltage limitations of the line scan device preclude the use of an h.t. line derived directly from the mains with a simple half-wave rectifier. A mains transformer is therefore required together with stabilizing circuit to produce an h.t. line of about 60 V to 70 V. A parallel efficiency diode is normally employed with a peak voltage of some 600 V applied to the output devices. The major disadvantage with such a system is, of course, the bulk of the mains transformer and the expense of the special stabilizing circuit. For colour receivers two output transistors are normally used and the e.h.t. is generated by means of a separate e.h.t. generating circuit which includes its own feedback system to maintain a constant e.h.t. potential.

The convergence and raster or pin-cushion correction circuits of the colour receiver are driven from both the horizontal and vertical time-bases. These circuits are characterized by a large number of pre-set controls, many of which in the case of a dual-standard receiver require switching between standards. Although the circuits do not appear to have changed very markedly over the years the quality and stability of convergence is the result of many small innovations in both the circuitry and the picture tube itself.

8. Future Trends

Although it is outside the scope of a survey paper to indicate future trends, many have in fact been indicated. With the introduction of colour television, for which the receiver cost is some three times that of a monochrome receiver, it is likely that monochrome receivers will become smaller and hence more portable and be used as second receivers. Unfortunately, however, portable receivers are not well suited to u.h.f. reception and hence in the U.K. dual-standard portable receivers may be required for some time until the v.h.f. bands are re-allocated.

Mechanical simplification of a receiver is greatly assisted by the use of semiconductors. In particular, varicap diodes enable the tuner to be constructed as part of the chassis with only d.c. connections to the controls. Hence, greater flexibility in styling will become possible and servicing simplified by modular construction. This latter aspect is most important in the case of colour receivers. Indications are, however, that hybrid receivers will be produced for some time to come with valves being retained for the power output stages. Circuits in the signal processing area lend themselves to integrated techniques. Only when the integrated circuits can show a cost saving will their use be justified but indications are that this will occur before the phase of hybrid receivers is over.

9. Acknowledgments

The author would like to acknowledge the contributions made by his colleagues to the preparation of this paper and to thank the Directors of Mullard Limited for permission to publish it.

Manuscript first received by the Institution on 16th December 1968, and in final form on 19th February 1969. (Paper No. 1259/Com. 17.)

Joint Conference on 'Digital Methods of Measurement'

University of Kent at Canterbury—23rd to 25th July 1969

PROVISIONAL PROGRAMME

The exact order and time of presentation of papers has not yet been finalized

Wednesday, 23rd July. 10 a.m. to 5.30 p.m.

Session on 'MEASUREMENT OF PHYSICAL QUANTITIES'

Chairman: K. J. DEAN

'Spatial Measurement using Digitized Vidicon Cameras,' J. E. CRAWFORD, P. E. OSMON and J. A. STRONG, *Westfield College, University of London.*

'A Commercial Laser Interferometer for Length Measurement by Fringe Counting,' A. J. BONNER and B. W. BARRINGER, *Rank Research Laboratories.*

'Computerized Microdensitometer for Mass Spectro Analysis,' L. MOLYNEUX and G. MIDGELEY, *University of Newcastle upon Tyne* and G. H. SCOTT, *Joyce, Loebl & Co.*

Chairman: Squadron-Leader D. R. McCALL

'The Digital Measurement of Angular Acceleration,' G. DE VISME, *University of Manchester Institute of Science and Technology.*

'Extracting Angular Values in Digital Form from Vectors expressed in Rectangular Co-ordinates,' P. DIEDERICH, *Decca Navigator Company.*

'Direct Digital Transduction of Environmental Properties,' P. B. FELLGETT, *University of Reading.*

'Pure Digital Electrical Power and Energy Measurement—Instruments and their Integration in On/Off-Line Systems,' C. PENESCU and T. IONESCU, *University of Bucharest, Roumania.*

'A Novel Digital Method for Linearizing and Displaying Strongly Non-Linear Frequency-Signal of Transducers,' L. SIMONFAY, G. MAYER and P. POTZY, *Central Measurement Research Laboratory, Budapest, Hungary.*

6 p.m. to 7 p.m. **Conference Cocktail Party in Eliot College.**

Thursday, 24th July. 9.15 a.m. to 5.45 p.m.

Session on 'DIGITAL INSTRUMENTS'

Chairman: D. R. OLLINGTON

'The Design of Logic and Counting Circuits Capable of Operating at 1 gigahertz Pulse Repetition Rate,' J. S. T. CHARTERS, *Heriot-Watt University.*

'A Flat, Dot Matrix, Glow Discharge Alpha-numeric Tube,' R. F. HALL, *Mullard Research Laboratories.*

'Decision Making in Digital Voltmeter Design,' B. MANN, *College of Aeronautics, Cranfield.*

'The Isolation of Digital Voltmeter Input Circuits,' J. R. PEARCE, *Solartron Electronic Group.*

'Digital Carry Applied to Successive Approximation Digital Voltmeters,' J. R. PEARCE, J. BLOOMFIELD and U. QURESHI, *Solartron Electronic Group.*

'Optimization of the Dual Ramp Voltmeter,' D. WHEABLE, *Solartron Electronic Group.*

'The Principles and Design of Digital Averaging Instruments,' H. E. HANRAHAN, *University of Sussex.*

'A Digital Trace Reader for the Evaluation of the Areas of Immuno-electrophoresis Curves,' L. MOLYNEUX, *University of Newcastle upon Tyne* and K. HERRON, *Joyce, Loebl & Co.*

Joint Conference on 'Digital Methods of Measurement'—continued

Chairman: R. B. KNIGHT

'Pulse-Rate to Digital-Code Tracking Converters,' J. D. MARTIN, *Bath University of Technology.*

'An Analogue to Digital Conversion System for Use in a Low Cost Digital Multimeter,' C. A. SPARKES, *University of Surrey* and R. G. CLARK, *Avo.*

'A Battery Powered Digital Recorder for Field Data Logging of Meteorological Parameters,' R. V. WALL, *The Plessey Company.*

'A U.H.F. Field Strength Measuring Receiver with Digital Display,' D. G. BEADLE, J. A. FOX and D. E. SUSANS, *B.B.C. Research Department.*

'Automatic Test Equipment for Industrial Electronics,' K. PFEIFFER, *Siemens, Munich.*

'Programming and Auto-ranging Digital Voltmeters,' D. WHEABLE, *Solartron Electronic Group.*

7 p.m. for 7.30 p.m. **Conference Dinner in Eliot College.**

Friday, 25th July. 9.15 a.m. to 4.30 p.m.

Session on 'MEASUREMENTS ON DIGITAL COMMUNICATION SYSTEMS'

Chairman: A. H. APPLEYARD

'Measurement Techniques Associated with P.C.M. Systems,' G. H. BENNETT, *G.P.O. Telecommunications Headquarters.*

'The Measurement of the Performance Parameter of P.C.M. Links,' A. N. RAMSDEN, *Marconi Instruments.*

'A Technique for the Evaluation of Data Communication Networks,' M. B. ASHDOWN, *Trend Electronics.*

'Assessment of Error Rates on Digital Communication Channels,' A. J. PERNA, *Marconi Instruments.*

'A Method of Recording and Analysing Errors Resulting from Digital Transmission over Telephone Lines,' P. NOAKES, *I.C.L.*

'Error Measurements on a Laser Digital Data Link,' P. G. FARRELL, *University of Kent.*

Chairman: A. G. WRAY

'The Correction of Distortion in Digital Instrumentation Systems,' R. V. LEEDHAM, *University of Bradford.*

'Polarity Coincidence Techniques for Correlation Function and System Response Evaluation,' J. A. M. MCDONNELL and J. FORRESTER, *University of Kent.*

'Phase Measurement Using an On-Line Computer,' J. S. BRIDLE, *Standard Telecommunication Laboratories.*

'Signal Processing of I.L.S. Information by Means of Delta-Modulation Techniques,' R. THOMASON and M. MUKUTMONI, *College of Aeronautics, Cranfield.*

'Standard L.F. Noise Sources using Digital Techniques and their Application to the Measurement of Noise Spectra,' H. SUTCLIFFE and K. F. KNOTT, *University of Salford.*

The registration charges are £11 for members of the sponsoring Institutions and £15 for non-members; these include the preprint volume, and lunch and refreshments on the three days of the Conference as well as a ticket for the Conference Cocktail party. The charge for residential accommodation in Eliot College is £6 10s. 0d. for two nights and includes a ticket for the Conference Dinner. (Non-residents may purchase tickets for the Dinner, the cost being £2 5s. 0d.) Registration forms may be obtained from the Conference Registrar, 8–9 Bedford Square, London, W.C.1 (Telephone 01-636 1901, Ext. 3). These should be returned as quickly as possible and *if accommodation is required, not later than 7th July.*

U.D.C. 681.38:621.374

An Introduction to Threshold Logic: A Survey of Present Theory and Practice

By

S. L. HURST, M.Sc.(Eng.), C.Eng., M.I.E.E., M.I.E.R.E.†

Presented (in abridged form) at the Conference on 'Electronic Switching and Logic Circuit Design', organized by the College of Technology, Letchworth, with the support of the I.E.R.E., and held at Letchworth on 24th October 1968.

Summary: The familiar logic building blocks of AND, OR, NAND, and NOR functions are practical realizations of well-known basic Boolean algebraic connectives, and as such enable any two-state logic system to be constructed. However, a more powerful basic logic building block than Boolean gates is the 'threshold gate'. Such a gate has binary inputs A, B, C, . . ., with 'weights' a, b, c, . . . associated with these respective inputs, and a binary output Z. The output from such a gate is:

$$Z = 1 \text{ if } \langle a.A + b.B + c.C + \ldots \rangle \geqslant \text{some value } t_1$$

$$Z = 0 \text{ if } \langle a.A + b.B + c.C + \ldots \rangle \leqslant \text{some value } t_2$$

where normal arithmetic rules are involved in the above summations.

Such gates can be used to realize all the normal basic binary logic functions, and in addition can realize more complex switching functions which would otherwise require the use of several normal logic gates. Thus the number of gates in a switching system may be drastically reduced by using threshold gates in place of the normal types.

List of Symbols

A, B, etc. = independent binary variables, taking the normal binary values of 0 and 1.

a, b, etc. = independent real-number coefficients or 'weights'.

t_1 = real-number upper gate threshold.

t_2 = real-number lower gate threshold, $t_1 \geqslant t_2$.

$f(A, B, \ldots)$ = Boolean function of independent variables A, B, etc.

To differentiate between Boolean expressions and threshold functions, the following notations are employed:

Boolean expressions employ [ ] for outer brackets, and ( ) for inner brackets. Within these defining brackets + and . take the normal Boolean meaning of OR and AND respectively. The latter may be omitted where no ambiguity results.

Threshold expressions employ $\langle\ \rangle$ for outer brackets, and { } for inner brackets. Within these defining brackets, the normal arithmetic rules of addition and multiplication hold.

† School of Electrical Engineering, Bath University of Technology, Bath, Somerset.

1. Introduction

Most digital systems in current use employ the familiar building blocks of AND, NAND, OR, and NOR functions. These functions or 'gates' are all direct practical realizations of fundamental Boolean algebra connectives, and as such are basic natural choices for a digital system.

However, when more complex switching functions than simple AND or OR connectives are required, then increasingly complex interconnexions of large numbers of such gates become necessary, the design of such interconnexions being aided by the various algebraic and graphical techniques now widely known.

Now if a range of digital logic gates was available which were logically more powerful than the above basic types, it might prove possible to use a reduced number of such gates to realize a given complex switching system, and an overall economy in design may become possible. Two more powerful types of logic gate have been proposed to supplement the basic Boolean gates, namely:

(a) majority gates, and

(b) threshold gates.

The former will be shown to be a particular case of the latter, with the basic Boolean gates being particular cases of both.

2. Majority Gates

As may be inferred from its designation, a majority gate is a multi-input function whose output is '0' unless a majority of its inputs are '1'. In order to avoid ambiguities, the number of inputs i per gate is normally considered to be an odd number, giving:

$$\text{output } Z = 1 \text{ if } \frac{i}{2} + \tfrac{1}{2} \text{ inputs are 1,}$$

otherwise 0.

Suppose $i = 7$. Then such a gate would give an output $Z = 1$ if any four or more of its inputs were 1. However if three inputs were permanently connected to 1, then the four remaining inputs could be used to realize a normal Boolean OR gate (see Fig. 1).

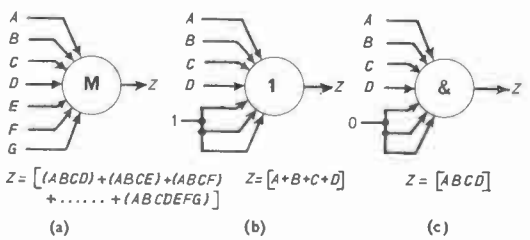


Fig. 1. 7-input majority gate, with different possible uses.
(a) Basic gate. (b) Used as a 4-input inclusive OR.
(c) Used as a 4-input AND.

Conversely, if three inputs of the gate were permanently connected to 0, then a four-input AND function would be realized. By permanently connecting various numbers of inputs to either 0 or 1, a large variety of logic functions may be realized by this one design of gate, with also the very powerful provision of changing its duty from one arrangement to another if the 'preset' 0 and 1 connections are made variable by some further 'master control' part of the logic system.

It will therefore be appreciated that if realizable in practice, a majority gate would prove a more powerful logic device than a simple Boolean gate. This increase in logic power however is obtained at the expense of circuit complexity of the gate, and also possibly at the expense of allowable tolerances on signal voltage levels of the system.

3. Threshold Gates

Now in the majority gate introduced above, each input has the same 'importance' or 'weight' in determining the gate output state, and also the 'threshold' of the summation of the input states is by definition equal to one-half the number of inputs. The previous expression for the gate output Z may therefore be re-phrased as follows:

$$Z = 1 \text{ if } \langle A+B+C+\ldots\rangle \text{ is more than } i/2$$
$$= 0 \text{ if } \langle A+B+C+\ldots\rangle \text{ is less than } i/2,$$

where A, B, C, etc., are the respective values 0 or 1 of the various binary input variables, and where + now signifies normal arithmetic addition.

However, if the restriction that each input of the gate has the same importance in determining the gate output state is now modified, and also the 'threshold' below which the output is 0 is not restricted to $i/2$, then a perfectly general binary 'threshold gate' is realized.

Let 'weighting' a, b, c, . . . be given to each gate input A, B, C, . . . respectively, where a, b, c, etc., are any decimal numbers, positive or negative. (In practice, such numbers are usual positive integers, but this restriction is for convenience only in many cases.) Then the output of such threshold gate may be expressed as:

$$Z = 1 \text{ if } \langle a.A+b.B+\ldots\rangle \geqslant t_1,$$
$$= 0 \text{ if } \langle a.A+\mathrm{b}.B+\ldots\rangle \leqslant t_2$$

where t_1 = the upper threshold level of the gate,

t_2 = the lower threshold level of the gate,

and where + again signifies normal arithmetic addition, and . signifies normal arithmetic multiplication (The . symbol will normally be omitted.)

The introduction of two threshold values t_1 and t_2, where $t_1 \geqslant t_2$ and where neither need be an integer, is desirable in order to define tolerances that may be present in a practical threshold gate. The 'threshold gap' t_1 to t_2 defines the inadmissible weighted input summation which may give rise to an ambiguous output from the gate, and which should therefore be avoided in practice. The sharper the circuit action of the gate, then the closer t_2 may be made to t_1.

At this stage, therefore, the majority gate will be seen to be a particular type of threshold gate in which all input weights a, b, etc., are equal to $+1{\cdot}0$, and where t_1 and $t_2 = i/2 \pm$ appropriate circuit threshold tolerance. Ideally $t_1 = t_2 = i/2$. Hence the general threshold gate will be appreciated as being a more powerful logic device than the majority gate, both in their turn being more powerful than conventional Boolean logic gates.

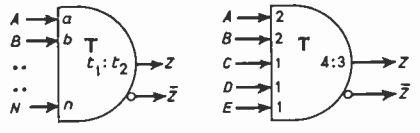
In order to simplify the above two separate equations that specify the action of any threshold gate, from henceforth the following combined notation

$$Z = \langle a.A+b.B+\ldots\rangle_{t_1:t_2}$$

will be used, where t_1 and t_2 are the upper and lower gate threshold values as defined above.

Also, the general symbol that will be adopted for a threshold gate will be as shown in Fig. 2(a), the

identification on this symbol being as previously defined. Figure 2(b) is the symbol for a specific threshold gate with five inputs of dissimilar positive weighting, and upper and lower thresholds of 4 and 3 respectively. If complementary outputs are also available from a gate, these may be included as shown in Fig. 2.



(a) General symbol (b) Specific gate,

$$Z = \langle 2A + 2B + C + D + E \rangle_{4:3}$$

Fig. 2. Threshold logic gate symbols.

4. The Use of Threshold Gates

Consider first a relatively simple logic function:

$$f(A, B, C) = [A + BC]$$

To realize this function by normal Boolean gates would necessitate the use of two gates, an AND gate to generate the function BC, plus an OR gate to combine BC with A. However it is clear that in terms of threshold working the input variable A has twice the 'importance' of variables B and C. Hence the function may be realized with a single threshold gate as follows.

$$Z = \langle 2A + B + C \rangle_{2:1} \qquad \text{.....(1)}$$

In a similar manner, the following further examples may be realized by single threshold gates. It will be appreciated that any given logic function should first be reduced to its minimum irredundant form by a suitable minimization technique, before 'translation' of the function into a threshold realization is attempted. Failure to use a minimum sum-of-products expression may involve much wasted effort in the 'translation' procedure.

Boolean Function	Threshold Realization
$f(A, B, C) = [A(B + C)]$	$Z = \langle 2A + B + C \rangle_{3:2}$ .....(2)
$f(A, B, C) = [AB + BC + AC]$	$Z = \langle A + B + C \rangle_{2:1}$ .....(3) (= simple majority gate)
$f(A, B, C, D) = [AB + BCD]$	$Z = \langle 2A + 3B + C + D \rangle_{5:4}$ (4)

The threshold realization of the above simple examples may be formulated by a heuristic approach rather than by any more formal analytical analysis of the given Boolean function. With such simple examples, it is usually clear which is the 'most important' variable in the function, and which are the 'least important'. The most important variable must then be given the highest value weighting, with the least important variables the lowest weighting (normally unity), the gate thresholds then being adjusted accordingly.

Often factorizing the given Boolean function will reveal more clearly the most important variable(s), and possibly the co-equal variables. For example, if function (4) above is re-written as

$$f(A, B, C, D) = [B(A + CD)]$$

then A will be seen as being 'equal' to CD, with B as the dominating variable. This distinction is then realized by the weighting of the equivalent threshold function.

There are however many possible Boolean functions which cannot be realized by one single threshold gate of any weighting or threshold. Such functions are stated to be 'not linearly separable', as distinct from functions which have a single threshold realization, and which therefore are 'linearly separable'. (This feature of linear separability will be discussed more fully later.) Consider the simple function

$$f(A, B, C, D) = [AB + CD]$$

If a solution such as $Z = \langle A + B + C + D \rangle_{2:1}$ is tried, such solution is clearly incorrect as it does not preclude terms such as AC, AD, etc., from giving an output $Z = 1$. No weightings or thresholds can be formulated to realize this particular function with one threshold gate.

To realize such a function with threshold gates, more than one gate is required. Now function AB may be realized by $\langle A + B \rangle_{2:1}$, which gate has an output 0 unless A *and* B are present. Thus using this gate output as a single variable, the complete function $f(A, B, C, D) = [AB + CD]$ may be realized using two threshold gates as follows:

$$Z = \langle 2\{\langle A + B \rangle_{2:1}\} + C + D \rangle_{2:1} \qquad \text{.....(5)}$$

This realization is illustrated in Fig. 3 below. It may be noted that this realization still uses one gate less than a normal Boolean gate realization.

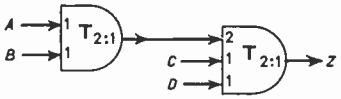


Fig. 3. Threshold realization of $Z = [AB + CD]$.

In cases where the threshold level of the gate is fixed by a particular gate design, then unused gate inputs may be used to realize the desired function. Suppose function (2) above was to be realized, but the threshold gate design available was, say, a 5-input gate with thresholds 4 : 3. In order to produce effective thresholds 3 : 2 for the three input variables

A, B, and C, then an unused gate with unity weighting must be connected to a 1 logic level voltage. This will effectively permanently bias the gate with a 1 in its input summation, which is equivalent to reducing the thresholds against which the summated variable inputs A, B and C are equated by one.

The threshold gate illustrated in Fig. 2(b) is a specific design that has been produced in small quantities in integrated circuit form. If all its five inputs are used as separate variable inputs, it realizes the Boolean function:

$$f(A,B,C,D,E) = [AB+(A+B)(CD+CE+DE)] \quad (6)$$

However, when one or more of its inputs are preset to 0 or 1, and/or when two or more inputs leads are tied together, a large range of logic functions may be realized. Amongst them are the following:

(i) $$f(P,Q,R) = [P+QR] \quad \text{.....(7)}$$

where $P = B$,
$Q = C$,
$R = D$,
with $A = 1$ $E = 0$,

(ii) $$f(P,Q,R) = [PQ+PR+QR] \quad \text{.....(8)}$$

where $P = C$, D and E tied together,
$Q = A$,
$R = B$,

(iii) $$f(P,Q,R,S) = [P+QR+RS+QS] \quad \text{.....(9)}$$

where $P = B$,
$Q = C$,
$R = D$,
$S = E$,
with $A = 1$,

plus several further variations. It may be seen that the effective thresholds of this gate can be modified from 4 : 3 maximum down to 1 : 0 minimum by permanently connecting one or more inputs to 1, whilst an input variable may be weighted up to the useful maximum of 4 by tying two or more gate inputs together.

5. Complemented Variables: Complemented Functions, Negative Weights

If the given logic function contains complemented variables, these may be treated in exactly the same way as functions with all non-complemented variables.

For example, if function (2) in the preceeding section had been given as $f(A, B, C) = [A(B+\bar{C})]$, then its threshold realization would be very similar, namely:

$$Z = \langle 2A+B+\bar{C}\rangle_{3:2} \quad \text{.....(10)}$$

However, where a variable appears in both its complemented and uncomplemented form in a given logic function, the given function must first be rechecked to ensure that it is in its minimum irredundant sum-of-products form, as all functions should be before translation into threshold realization is attempted. For example, suppose the function $f(A, B, C, D) = [A+\bar{A}CD]$ is given. This may be realized as:

$$Z = \langle 3A+\bar{A}+C+D\rangle_{3:2} \quad \text{.....(11a)}$$

but if the given function is re-examined, it will be seen that its minimum form is $[A+CD]$, which therefore has a preferable threshold realization of:

$$Z = \langle 2A+C+D\rangle_{2:1} \quad \text{.....(11b)}$$

If the true minimum sum-of-products function does contain one or more variable present in both its complemented and uncomplemented form, then it will be found to be impossible to realize such a function with one single-threshold gate. (Note that eqn. (11a) was not in true minimum form.) Thus only functions that do not contain any variable in both complemented and uncomplemented form in a minimum sum-of-products expression are 'linearly separable', and may thus be realized with one threshold gate of appropriate weights and thresholds. Such functions are termed 'unate functions'. Now although all threshold functions are unate, not all unate functions are threshold functions. For functions of three or fewer variables, all unate functions are however threshold functions, but for four and more variables an increasing number of functions, even though unate, becomes unrealizable by a single threshold gate.

The following examples illustrate the above features.

Consider the function $f(A, B, C, D) = [AB+\bar{B}CD]$. This function is in minimum irredundant form, but is not unate. It cannot therefore be realized with one threshold gate. Two gates are necessary in this case giving a realization of:

$$Z = \langle 3\{\langle A+B\rangle_{2:1}\}+\bar{B}+C+D\rangle_{3:2} \quad \text{...(12)}$$

The function $f(A, B, C) = [\bar{A}+B\bar{C}]$ is unate, is in minimum irredundant form, contains only three variables, and hence is realizable with one threshold gate. The simplest realization is:

$$Z = \langle 2\bar{A}+B+\bar{C}\rangle_{2:1} \quad \text{.....(13)}$$

The function $f(A, B, C, D) = [AB+\bar{C}\bar{D}]$ is also unate and in minimum irredundant form, but is one of the unate functions not realizable by one threshold gate. Two gates are necessary, giving a realization of:

$$Z = \langle 2\{\langle A+B\rangle_{2:1}\}+\bar{C}+\bar{D}\rangle_{2:1} \quad \text{.....(14)}$$

If any logic function has a single threshold gate realization, then the complement of that function also has a single-gate realization. The conversion from a

given threshold function to its complement is obtained by:

(i) changing the signs of all the weights of the variables, and

(ii) changing the signs of t_1 and t_2 and interchanging their positions.

For example, given the threshold function

$$Z = \langle \bar{A}+2B+C+D \rangle_{4:3}$$

the complement of this function is then given by:

$$\bar{Z} = \langle -\bar{A}-2B-C-D \rangle_{-3:-4}$$

However, negative weights and negative thresholds may not be practical propositions, actual threshold gate circuits generally requiring all positive-weight realization. Expressions involving negative weights such as above, may be transformed into positive weighting by applying the identity $x = 1-\bar{x}$, where x is any binary variable, complemented or uncomplemented, following which the thresholds may be adjusted to make them also positive.

Substituting therefore $\{1-A\}$ for $\bar{A}$, $\{1-\bar{B}\}$ for B, etc., the above threshold function becomes:

$$\begin{aligned}\bar{Z} &= \langle -\{1-A\}-2\{1-\bar{B}\}-\{1-\bar{C}\}-\{1-\bar{D}\} \rangle_{-3:-4} \\ &= \langle A+2\bar{B}+\bar{C}+\bar{D}-5 \rangle_{-3:-4} \\ &= \langle A+2\bar{B}+\bar{C}+\bar{D} \rangle_{2:1} \qquad \text{.....(15)}\end{aligned}$$

It will be noted that in the final positive-weighting form, the threshold function for $\bar{Z}$ is very similar to the function for Z, involving the complements of each of the original terms, but the upper and lower threshold values may be dissimilar in the two cases.

This technique for eliminating a negative weighting may be applied in all cases where a negative weight arises. For example a trial approach to realize a particular logic function may produce:

$$Z = \langle 2A+B+2C-D \rangle_{3:2}$$

which can then be transformed to give a more practical result as follows:

$$\begin{aligned}Z &= \langle 2A+B+2C-\{1-\bar{D}\} \rangle_{3:2} \\ &= \langle 2A+B+2C+\bar{D} \rangle_{4:3} \qquad \text{.....(16)}\end{aligned}$$

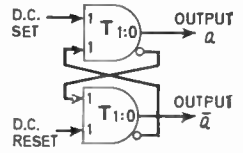
Conversely, if negative weights are allowed by the particular gate design, then positive weightings may be converted into negative weights, with a possible beneficial reduction in the resultant gate upper and lower threshold levels.

6. Bistable (Storage) Threshold Arrangements

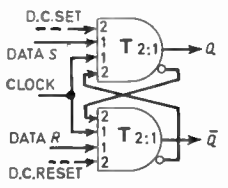
Threshold gates lend themselves to the building of storage circuit arrangements with very great power and flexibility. It is often possible to achieve with an extra input on a threshold gate a storage action that would necessitate additional separate gates in normal Boolean realization.

The simplest form of an unclocked d.c.-coupled bistable assembly is shown in Fig. 4(a). Its circuit action should be clear, as it very closely parallels the circuit action of the usual simple cross-coupled symmetrical transistor bistable circuit.

A clocked bistable circuit with 'set' and 'reset' data inputs S and R respectively is shown in Fig. 4(b). This arrangement produces type R-S circuit action, as the data input condition $R = 1$ and $S = 1$ will give rise to ambiguous operation.



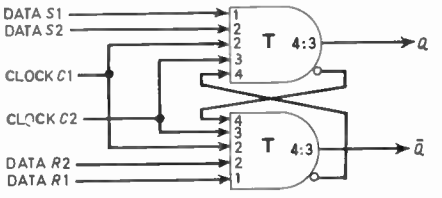
(a) Unclocked bistable circuit



State table

R	S	Q_{n+1}
0	0	Q_n
0	1	1
1	0	0
1	1	?

(b) Clocked 'set–reset' circuit



(c) Double-clocked 'set–reset' circuit

Fig. 4. Possible bistable (storage) circuit arrangements.

Additional inputs with further weightings can provide other bistable circuit actions, including selective response to more than one clock line. Figure 4(c) illustrates a possible arrangement wherein clock line $C1$ is unable to feed in data held on $S1$ and $R1$, but can feed in to store data held on $S2$ and $R2$. It will be seen that in all these circuit arrangements, the cross-coupling feedback to hold the circuit in each of its two stable conditions has a weighting greater than or equal to the upper threshold of each threshold gate.

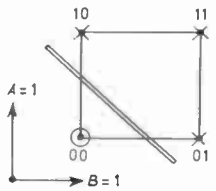
Assemblies of bistable circuits as above into ring counters, shift registers and all the normal type of binary coding assemblies are possible. However in

some of these applications, the great potential of the threshold gate is not always fully exploited, and hence the more conventional binary circuits often remain valid and economical solutions in this particular sphere of application.

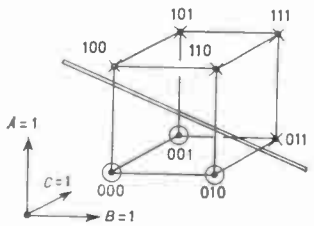
7. Linear Separability

In the preceding sections, examples of Boolean functions have been given which are realizable with one single threshold gate of appropriate weightings and thresholds, and also other functions which did not have single-gate realization.

Functions which have a single-gate realization are the class of Boolean functions that are 'linearly separable'. Now if such linearly-separable functions are plotted as a multi-dimensional hypercube representation, then for all such functions there exists a plane cutting the hypercube, dividing the vertices of the hypercube representing output equal to 1 from all the vertices representing output equal to 0. If no such plane exists, then the given function is not linearly separable, and thus not realizable by one threshold gate.



(a) $f(A, B) = [A + B]$



(b) $f(A, B, C) = [A + BC]$

Fig. 5. Hypercube representation of two- and three-variable functions.

As a very simple case, consider the function $f(A, B) = [A+B]$. The hypercube representation of this two-variable expression is simple, as shown in Fig. 5(a), the vertices being marked O and × to represent the $f(A, B) = 0$ and $f(A, B) = 1$ function values, respectively. It will be clear that a plane dividing the 0 from the 1 vertices is possible in this case, and thus the function is linearly separable. Its threshold realization is of course $Z = \langle A+B\rangle_{1:0}$.

However if the function $f(A, B) = [\bar{A}\bar{B}+AB]$ is given, then no one plane can be drawn through the hypercube to divide the 0 and the 1 vertices. Such function is not therefore linearly separable.

Extending the hypercube construction to functions of three variables gives the construction shown in Fig. 5(b), the previous function $[A+BC]$ of eqn. (1) being plotted on this hypercube. The plane which divides the 0 from the 1 vertices is apparent. If, however, the function of say $f(A, B, C) = [\bar{A}\bar{B}\bar{C}+AC]$ was plotted, then no one plane can be found to divide these vertices, and therefore this function is not linearly separable.

Extension of this hypercube representation beyond three variables becomes increasingly impractical to draw and visualize clearly. Methods have been suggested, which involve drawing two separate three-variable hypercubes for four-variable functions, and four separate three-variable hypercubes for five-variable functions, and then applying certain search techniques to prove linear separability, but such techniques are tedious.

Alternative methods are thus desirable in order firstly to decide whether given Boolean functions of three or more variables are linearly separable or not, and secondly to determine the weightings and thresholds necessary to realize such functions that are linearly separable.

A great deal of algebraic and associated work has been done on these closely related problems, but fundamentally they appear to be problems which cannot have a quick and simple general method of solution. For up to three variables, no trouble exists, as all possible threshold functions may be easily tabulated (see Table 1). To tabulate all possible threshold functions for four and more variables however, becomes increasingly tedious, as the number of possible single-threshold functions of n variables increases exponentially with n.

Table 1

All possible threshold functions for 1-, 2-, and 3-variable functions

Function	Threshold realization
x	$\langle x\rangle_{1:0}$
$x + y$	$\langle x + y\rangle_{1:0}$
xy	$\langle x + y\rangle_{2:1}$
$x + y + z$	$\langle x + y + z\rangle_{1:0}$
$xy + yz + xz$	$\langle x + y + z\rangle_{2:1}$
xyz	$\langle x + y + z\rangle_{3:2}$
$x + yz$	$\langle 2x + y + z\rangle_{2:1}$
$x(y + z)$	$\langle 2x + y + z\rangle_{3:2}$

Note: x, y, z = any variable, uncomplemented or complemented.

To tackle functions of up to seven variables, alternative tabulations to list all possible threshold functions have been devised and published. As it proves impractical to list precisely all the actual functions that are linearly separable, these further tabulations are given in terms of further parameters of the switching functions, often termed the 'Chow parameters'. If the Chow parameters for any given Boolean switching function are computed, then these published tabulations may be consulted to see if the computed parameter values are listed, or not. If they are listed, then the given function is linearly separable, and appropriate weightings may be read off the table; if the computed parameters are not listed, then the function is not linearly separable, and no single-gate realization is possible.

The Chow parameters for a function of n variable are labelled b_0, b_A, b_B, . . ., b_N, and are computed as follows:

(i) Compile the complete truth-table of 2^n entries for the function, together with the function value 0 or 1 against each entry.

(ii) Let m_1 be the number of truth-table entries for which the function value is 1, and m_0 the number of entries for which the function value is 0.

Then $b_0 = m_1 - m_0$.

(This and all following parameters may be a positive or a negative number. The sign as well as the magnitude must be noted.)

(iii) Let m_A be the number of entries for which $A = 1$ and the function value is 1, and $m_{\bar{A}}$ the number of entries for which $A = 0$ and the function value is 1.

Then

$$b_A = 2\{m_A - m_{\bar{A}}\}.$$

Similarly,

$$b_B = 2\{m_B - m_{\bar{B}}\}$$

etc., down to

$$b_N = 2\{m_N - m_{\bar{N}}\}$$

(iv) Finally list all above parameter values in magnitude order, largest magnitude on the left. This final sequence of numbers is the sequence that is compared against the prepared tables of Chow parameters for single-gate threshold functions.

As an illustration of the use of these parameters to determine linear separability, consider the following examples.

Example 1:

Take the simple example

$$f(A, B, C, D) = [AB + CD]$$

which has already been shown by a heuristic approach to be not linearly separable.

The complete truth table of this function is given as Table 2. (For functions of not more than four variables, a Karnaugh map layout rather than a truth table layout may be preferred.)

Table 2

A	B	C	D	$f(A, B, C, D)$
0	0	0	0	0
0	0	0	1	0
0	0	1	0	0
0	0	1	1	1
0	1	0	0	0
0	1	0	1	0
0	1	1	0	0
0	1	1	1	1
1	0	0	0	0
1	0	0	1	0
1	0	1	0	0
1	0	1	1	1
1	1	0	0	1
1	1	0	1	1
1	1	1	0	1
1	1	1	1	1

From the Table, the Chow parameters may be computed as follows:

$$b_0 = 7 - 9 = -2$$
$$b_A = 2\{5 - 2\} = +6$$
$$b_B = 2\{5 - 2\} = +6$$
$$b_C = 2\{5 - 2\} = +6$$
$$b_D = 2\{5 - 2\} = +6$$

Therefore parameters are:

b_A	b_B	b_C	b_D	b_0
+6	+6	+6	+6	−2

or taking magnitudes only:

6	6	6	6	2

The Chow parameters for not exceeding five variables are given in Table 3. Looking through this tabulation it will be seen that the above pattern of numbers does not appear. Thus the given function is not linearly separable, and cannot therefore be realized by one threshold gate, as has previously been reasoned by a less formal approach.

Example 2:

$$f(A, B, C, D) = [\bar{A}B + BC + BD + \bar{A}CD]$$

Evaluating the Chow parameters as previously yields:

b_B	b_A	b_C	b_D	b_0
+12	−4	+4	+4	0

or magnitudes:

12 4 4 4 0

Now this sequence of parameter values *is* listed in the linearly-separable Chow parameter tabulation. Reading off from this table, the magnitude of the required weighting $|w|$ for the respective variables is:

2 1 1 1 0

Thus the final respective weights a, b, etc., for the individual variables A, B, etc., are given by the above values for $|w|$, but with the sign of the originally computed Chow parameter, i.e.

$$a = -1$$
$$b = +2$$
$$c = +1$$
$$d = +1$$

It now remains to compute the required threshold to 'match' the above weightings. The rule for obtaining the threshold values is:

$$t_1 = \tfrac{1}{2}\{\alpha - b_0 + 1\}$$

and

$$t_2 = \tfrac{1}{2}\{\alpha - b_0 - 1\}$$

where

$$\alpha = \sum_{a}^{n} [\text{weights}]$$

Therefore, for the above function,

$$t_1 = \tfrac{1}{2}\{\{-1+2+1+1\}-0+1\} = 2$$
$$t_2 = \tfrac{1}{2}\{\{-1+2+1+1\}-0-1\} = 1$$

Hence

$$f(A, B, C, D) = \langle -A+2B+C+D \rangle_{2:1}$$

or, if the negative weighting of A is eliminated by the technique previously covered,

$$f(A, B, C, D) = \langle \bar{A}+2B+C+D \rangle_{3:2}$$

The above technique, with the aid of the available Chow parameter tables, is possibly the easiest method yet published for synthesizing threshold functions from given Boolean functions. It has however two drawbacks, namely:

(a) It cannot be used for functions of more than seven binary variables, due to the non-availability to date of the tables for greater than seven variables. Also, in view of the exponential increase of possible threshold functions with the number of variables, it seems improbable that further tables will be compiled.

(b) It cannot efficiently handle incompletely specified Boolean functions, that is functions that contain 'don't care' output conditions in their truth table compilation. With such incompletely specified functions, it is very much trial and error to find an optimum threshold realization using Chow parameters, as several solutions may be pursued by arbitrarily allocating 0 or 1 outputs to the 'don't care' conditions.

Table 3

Chow parameters and resultant weightings for all linearly-separable functions of not exceeding 5 variables

No. of variables	Chow parameter magnitudes $\|b\|$						Required weighting magnitudes $\|w\|$					
⩽3	8	0	0	0			1	0	0	0		
	6	2	2	2			2	1	1	1		
	4	4	4	0			1	1	1	0		
⩽4	16	0	0	0	0		1	0	0	0	0	
	14	2	2	2	2		3	1	1	1	1	
	12	4	4	4	0		2	1	1	1	0	
	10	6	6	2	2		3	2	2	1	1	
	8	8	8	0	0		1	1	1	0	0	
	8	8	4	4	4		2	2	1	1	1	
	6	6	6	6	6		1	1	1	1	1	
⩽5	32	0	0	0	0	0	1	0	0	0	0	0
	30	2	2	2	2	2	4	1	1	1	1	1
	28	4	4	4	4	0	3	1	1	1	1	0
	26	6	6	6	2	2	5	2	2	2	1	1
	24	8	8	4	4	4	4	2	2	1	1	1
	24	8	8	8	0	0	2	1	1	1	0	0
	22	10	10	6	2	2	5	3	3	2	1	1
	22	10	6	6	6	6	3	2	1	1	1	1
	20	12	12	4	4	0	3	2	2	1	1	0
	20	12	8	8	4	4	4	3	2	2	1	1
	20	8	8	8	8	8	2	1	1	1	1	1
	18	14	14	2	2	2	4	3	3	1	1	1
	18	14	10	6	6	2	5	4	3	2	2	1
	18	10	10	10	6	6	3	2	2	2	1	1
	16	16	16	0	0	0	1	1	1	0	0	0
	16	16	12	4	4	4	3	3	2	1	1	1
	16	16	8	8	8	0	2	2	1	1	1	0
	16	12	12	8	8	4	4	3	3	2	2	1
	14	14	14	6	6	6	2	2	2	1	1	1
	14	14	10	10	10	2	3	3	2	2	2	1
	12	12	12	12	12	0	1	1	1	1	1	0

As alternatives therefore to the foregoing synthesizing technique, certain graphical and allied techniques have been proposed. Possibly the two most well-known ones are the graphical 'threshold tree' construction of Lewis and Coates, and the 'threshold map' construction of Torng. Both techniques allow 'don't care' conditions to be built into their construction, albeit with some complication to the working, and both theoretically are not restricted by the number of binary variables involved. In practice, however, both constructions become a little tedious with a large number of variables, and the possibility of human error in the working unfortunately increases proportionally.

In the Lewis and Coates 'threshold tree' method of synthesis, which is similar to an earlier majority-gate synthesis of Miyata, the given Boolean function is 'decomposed' or factorized into smaller and smaller

functions of fewer variables at each step. For example, a given function $f(A, B, C, D)$ may first be broken down into two parts namely:

$$f(A, B, C, D) = [A.f_1(B, C, D) + \bar{A}.f_0(B, C, D)]$$

The right-hand terms are termed the 'descendants' of the left-hand 'ancestor' term.

Each of these 1-level descendant terms of the 0-level ancestor may now be broken down into two further descendants, to produce the 2-level terms, namely:

$$A.f_1(B, C, D) = [AB.f_{11}(C, D) + A\bar{B}.f_{10}(C, D)]$$

and

$$\bar{A}.f_0(B, C, D) = [\bar{A}B.f_{01}(C, D) + \bar{A}\bar{B}.f_{00}(C, D)]$$

These four 2-level descendants may be further broken down in the same manner, and so on until all descendant terms which cannot be further factorized are reached. For n variables, there may be a maximum of n levels of decomposition below the given 0-level Boolean function. The decomposition may not of course extend down to the nth level, as terms of one variable only may be reached at a higher level of decomposition, depending upon the given function.

Finally, a pyramid or 'tree' of terms will be obtained, with the given function at the top, and each branch downwards finally terminating in a single constant term.

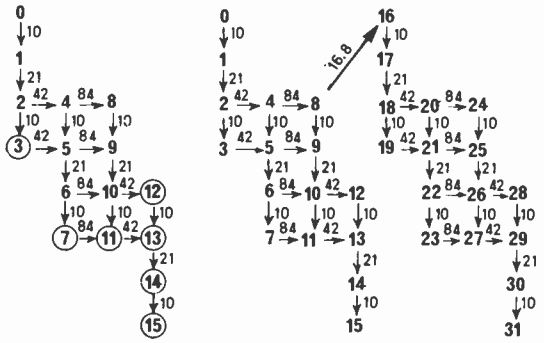
Rules may now be applied to this tree in order to determine the weights and thresholds, starting at the bottom and working upwards. If the given function is not linearly separable, then at some higher level of the tree a confliction of results between two branches of the tree will be encountered, which indicates that single-gate realization of the given function is not possible. If, however, the given function is linearly separable, then the rules for 'climbing' the tree may be carried out without confliction right to the top, producing a final single-gate realization for the given function. Full details of this technique with numerical examples are given by Lewis and Coates.[3]

In Torng's 'threshold map' construction, the 2^n standard products or 'minterms' of a switching function of n variables are first laid out in a two-dimensional lattice type of arrangement, with minterm $000\cdots$ at the top left, and minterm $111\cdots$ at the bottom right (see Fig. 6). Each node of this lattice thus represents a particular minterm, in exactly the same manner as each vertex in a hypercube representation (see Fig. 5).

Between each pair of adjacent nodes, an arrow is drawn, which is the 'characteristic vector' of the function between these two minterms. These characteristic vectors are each appropriately labelled to indicate which one or more variable changes from 0 to 1 between the two adjacent nodes. For n variables, there will be n values of characteristic vector linking the 2^n nodes.

Finally, in the construction of this map or grid, the nodes for which the function value is 1 are circled; nodes left uncircled thus represent value 0. 'Don't care' conditions, if present, may be represented by a broken circle around appropriate nodes.

Having completed this construction, it will be clearly seen how traversing each of the characteristic vectors between all nodes causes the function value to change from 0 to 1 or vice versa. Functions which are not linearly separable will be revealed by the inconsistency that in traversing one particular vector between two nodes, the function value changes from say 0 to 1, whereas traversing the same valued vector in the *same direction* between two other nodes, the function value changes from 1 to 0. This is clearly an inconsistent result, and means that no single-gate threshold realization is possible.



(a) 4-variable map, with $f(A, B, C, D) = [AB + CD]$ plotted thereon.

(b) 5-variable map.

Fig. 6. Torng maps for 4- and 5-variable Boolean functions. (Nodes 0, 1, etc., correspond to minterms 0000, 0001, etc.)

If, however, no inconsistent results as above can be found, then the function is linearly separable, and appropriate weights and thresholds exist for its single-gate realization.

The example plotted in Fig. 6(a) is the function previously encountered in Section 4, eqn. (5), and which was then shown to be not a linearly separable function. When Fig. 6(a) is examined, the inconsistency which precludes single-gate realization is readily observed, namely that in traversing the '42' vector from minterm 3 to minterm 5 the function value changes from 1 to 0, but in traversing the same vector from minterm 10 to minterm 12 the function value changes from 0 to 1. Thus no single-gate realization of the function is possible.

However, if no inconsistencies as above can be found on a Torng map, then a subsequent tabular procedure is available which involves the simple listing of inequalities that must be satisfied by the various characteristic vectors. The respective weights required by the individual variables can then be generated. As the number of inequalities involved is relatively small, a rapid solution can normally be found. Full details with worked examples may be found in papers by Torng[6] and Blomgren.[16]

The three different techniques introduced above for the synthesis of threshold gates still leave two main questions unanswered, namely:

(i) Is the gate realization produced by synthesis of the given function a realistic solution, that is, can an actual circuit be made to have the weighting and threshold performance demanded by the solution?

(ii) In the case of a given function that has been shown to be not linearly separable, what is the optimum way in which the function may be divided for realization by the fewest number of threshold gates?

As far as the latter problem is concerned, no generally accepted solution to the way multi-threshold gate synthesis should be approached is yet available. This is one area of threshold logic in which a great deal of work is still open to investigation.

Regarding the performance available from threshold gate circuits, this also is subject to much current development, particularly in the field of microelectronic realization. It can be visualized that a threshold gate circuit is a combination of an analogue form of input summation and a digital output stage, and thus some of the problems associated with analogue computational accuracy become involved in the design of high-performance threshold gates.

Further details and consideration of actual gate circuits and their performance will be given in the following Section.

8. Threshold Gate Circuits: Minimum Gap, Relative Gap, Normalized Fan-In

It is intuitively apparent that to produce a threshold gate circuit with thresholds of say 7 : 6 is more difficult and demanding on component tolerances, etc., than to produce a gate with thresholds of say 3 : 2. In the former case there is only a 13% increase in total input summation between the upper limit of the 0 output and the lower limit of the 1 output, whilst in the latter case there is a 50% increase between the same two limits.

In order to formalize the specification of the gate performance, the following terms have been proposed:

8.1. *Minimum Gap, or Gap Length*

The minimum gap of a gate is given by $t_1 - t_2$.

In all the foregoing sections, it has been assumed that t_1 and t_2 are integers, with $t_1 - t_2 =$ unity. This, however, was for convenience only, as no account was being taken in these sections of possible tolerances on input signal levels and gate weights. When such tolerances are taken into account, then to ensure the gate output remains 0 under certain maximum tolerances of signals, etc., and becomes 1 under certain minimum tolerances, t_1 and t_2 may not remain integers, but become fractional numbers as required by the worst-case design considerations.

8.2. *Relative Gap*

The value of the minimum gap $t_1 - t_2$ by itself does not indicate the selectivity of the gate circuit. The examples quoted above of thresholds of 7 : 6 and 3 : 2 have the same minimum gap, but it has already been argued that the 7 : 6 gate has to be more selective in operation than the 3 : 2 gate.

The relative gap is therefore defined as $\left\{\frac{t_1 - t_2}{t_1 + t_2}\right\}$.

With this more critical definition, it is obvious that the smaller the relative gap the more selective and difficult to build will be the gate circuit. The actual values of the threshold t_1 and t_2 are of secondary difficulty in the gate circuit design.

8.3. *Normalized Fan-In*

The fan-in of a threshold gate is simply the sum of the weights of all its inputs. The normalized fan-in is this summation divided by the minimum gap, i.e.

$$\text{normalized fan-in} = \left\{\frac{\sum_{a}^{n} (\text{weights})}{t_1 - t_2}\right\}$$

Again, intuitively, it will be recognized that the larger the value of the normalized fan-in, the more difficult will it be to cater for the practical tolerances on input signal levels and weights, etc.

Applying the above definitions to the threshold gate realization of eqn. (4), namely

$$Z = \langle 2A + 3B + C + D \rangle_{5:4}$$

this gate is required to have a relative gap of 0·11, with a normalized fan-in of 7. If these parameters are not feasible in the circuit design, then the function cannot be realized in practice with one gate. The threshold equation may, however, be broken down into two parts, which will give easier gate realizations; for example, a two-gate realization could be:

$$Z = \langle B + \langle 2A + C + D \rangle_{2:1} \rangle_{2:1}$$

which employs two gates, each of relative gap 0·33 and normalized fan-in of 2 and 4 respectively.

In most practical threshold gate circuit designs, resistors are used to 'weight' the different inputs. Square-loop magnetic core threshold logic circuits have been proposed, in which the turns and hence ampere-turns associated with each input are the controlling influences in the gate, but such realizations do not lend themselves to large-scale adoption, particularly as such techniques are not in line with current microelectronic device technology. One great advantage that such magnetic logic devices do possess, however, which resistor weighting cannot easily produce, is that negative weighting is perfectly possible by merely reversing the connexions to an input winding.

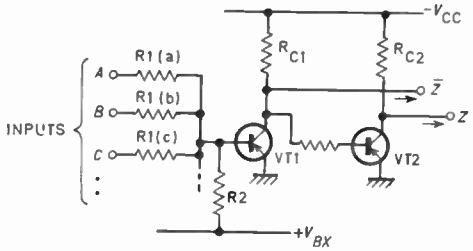


Fig. 7. Simple potential-divider type resistor–transistor threshold gate.

However to concentrate here solely on resistor-weighted gate circuits, Fig. 7 illustrates a resistor-transistor threshold logic gate originally developed for digital computer applications. With this circuit arrangement, until a certain number of inputs are energized with a '1' negative voltage input signal, the effective potential divider R1 : R2 holds the base of the transistor VT1 reverse-biased. When sufficient inputs are energized, then the base input of this transistor becomes forward biased, and provided the transistor gain is sufficiently high, VT1 will saturate, switching 'off' VT2.

Different input weights and thresholds can be provided by varying the values of the various input resistors R1, and the reverse-bias resistor R2 or supply V_{BX}. The circuit, however, is very critical of resistance and supply voltage tolerances, and very demanding on current gain of VT1 to produce sharp discrimination between the upper and lower thresholds t_1 and t_2.

More satisfactory threshold gates, particularly with regard to permitted tolerance of '1' input signal voltage levels, are produced by using long-tailed pair transistor assemblies as the input weighting circuits. Figure 8 illustrates the basic circuit arrangement of one such gate. The emitter resistors R1, R2, . . . R5 control the 'weight' of each respective input, the resistor R6 being the summing resistor through which the units of current produced by the various inputs are summed.

Each input voltage is compared against a preset reference voltage V_{REF}, which prevents noise and other small voltage excursions on '0' input lines from affecting the correct summation. A further long-tailed-pair circuit acts as the threshold detector, comparing the volt-drop developed across the summing resistor R6 with the volt-drop across resistor R7. Emitter-follower output stages for Z and $\bar{Z}$, with diode

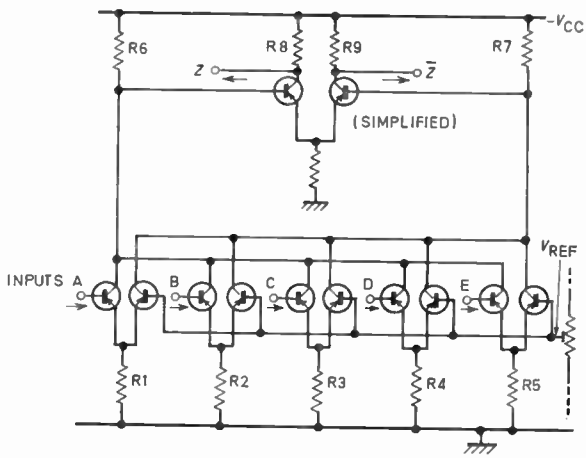


Fig. 8. Basic circuit arrangement of long-tailed-pair type threshold gate.

clamping to provide well-defined output voltage levels, are also often incorporated, rather than employing directly the output points shown in the simplified circuit of Fig. 8.

Various versions of this type of circuit are possible, particularly in the threshold detector part of the gate. The long-tailed-pair input circuits, with their generation of units of current that are then summed in one series resistor, however, appear to be becoming the standard and most satisfactory way to produce the weighting for the different inputs.

Using the latter types of circuit, threshold gates can be constructed with a relative gap of about 0·1 minimum, and a normalized fan-in of about 15 maximum. This represents upper and lower thresholds of say 6 : 5, and inputs weighted say 5; 4; 2; 2; 1; 1. The noise immunity of such circuits is as good as, if not superior to, normal Boolean gates, as two decision stages are involved per gate, namely the input circuits and the threshold detector. These two stages of gain result in a very steep transfer curve between the $Z = 0$ and the $Z = 1$ output states, steeper than is

normally encountered in conventional Boolean circuits and hence less prone to spurious output variations.

Possibly the first published use of solid-state threshold logic gates was in the design of a small general-purpose computer built by the General Electric Research Laboratories, Schenectady, N.Y., in 1964 for evaluation purposes. The computer, termed the DONUT computer ('Digitally-Operated-Network-Using-Thresholds'), employed discrete component resistor-transistor gates as discussed in Fig. 7, with thresholds ranging from 1 : 0 minimum to 5 : 4 maximum for the general-purpose logic functions. For certain accumulator functions, increased thresholds of 9 : 8 were used with a maximum fan-in of 32, but these exceptional and few cases were realized by tightening allowable tolerances on signal levels, etc., and by incorporating additional emitter-followers within the gate circuitry.

The general-purposes gates had the following supply voltage, etc., specification:

Main collector power supply V_{cc}
= 28 V ± 0·1%

Diode clamp and reference power supplies
= 13·6 V ± 0·1% and 17·1 V + 0·1%

Gate resistors = ± 2% tolerance

Logical '1' voltage = 14·0 V ± 0·15 V

Logical '0' voltage = 0·05 V min., 0·75 V max.

Nominal input resistance per unit weight
= 15 kΩ

Some 3500 gates were employed in the design, this number being about one-quarter the number of NOR gates that would be needed for an all-NOR realization of the same capacity computer. Exact comparison of relative numbers of gates of different types is difficult, as allowable tolerances and fan-ins, etc., affect the total number of gates in all types of realization. However, the DONUT project did establish the fact that threshold gates were practical and economic propositions for digital computer designs.

A more recently announced development in the field of gate realization, is the development of an integrated circuit threshold logic module, developed by the R.C.A. Laboratories, Princetown, N.J. This development consists of two separate threshold gates in one 14-pin flat-pack, one gate being a simple 3-input majority gate, with unity-weighted inputs and thresholds of 2 : 1, the other being the 5-input gate previously illustrated in Fig. 2(b).

The circuit of these microelectronic gates is basically as shown in Fig. 8 but with clamped emitter-follower outputs, and thus has improved performance compared with the earlier types of circuit used in the DONUT computer. The supply voltages, speed of response and signal levels of these threshold gates are compatible with conventional microelectronic logic gates; a single d.c. supply of −5 V is employed, with signal levels of −0·85 V maximum and −1·60 V minimum for 0 and 1 respectively. Both true and complemented outputs are available from each gate of the module.

Compared with conventional microelectronic logic modules, the chip size is about 25% greater in area. The expected yield in production is estimated to be about the same as conventional circuits, and hence the cost should not be appreciably greater than normal logic modules.

The very great advantage in microelectronic threshold gate realization is that the matching of resistor values and transistor characteristics on a chip is inherently good. This is of great significance in the input weighting and summation, as the absolute values of resistors, etc., are not critical, but rather the matching of values 'along the line' is far more important for correct summation. Resistor matching to within ±2% on any chip is easily realized with current microelectronic technology, and possibly ±1% can be maintained without noticeably increased rejection rates on test.

9. Future Prospects

In the foregoing sections, the subject of threshold logic, its synthesis, and its realization by electronic circuits of conventional circuit design, has been briefly introduced. Without doubt, the threshold gate offers the logic designer a new tool of very great potential, with possibilities of considerable economy in the overall design of a complex logic system such as a digital computer.

Not all the problem of synthesis have yet been solved by any means, particularly in cases where single-gate realization is not possible and some multi-gate solution has to be sought. Much work remains to be done, and undoubtedly will be, in this area.

Similarly, in the field of the actual circuit realization, it is likely that a considerable amount of work will be done in the future to establish a standard range of commercially available threshold gates, similar to the range of Boolean modules that are now so widely available. The question of how many inputs, what weighting, and what gate thresholds constitute the most versatile module, are factors that are involved in the consideration of what a 'standard' module should be. The circuit realization of such gates undoubtedly will be in microelectronic rather than discrete-component form.

Possible fields of application doubtless will embrace the complete digital computer field. Much of the

existing application work on threshold gates has been in this area, and it has been convincingly shown that all the standard computer circuit requirements, such as full-adders and shift registers, etc., can very conveniently be constructed using threshold gates. Industrial control systems, which at present employ conventional Boolean logic gates, are also fields where threshold gates could be used to advantage, limited only by the ingenuity of the circuit design engineer to utilize to the full the power and flexibility of the devices available.

Computer applications will undoubtedly go hand-in-hand with large-scale integration of the threshold devices. It has been estimated that threshold logic plus l.s.i. will cut the cost of computers by as much as 50% in comparison with present-day micro-electronic designs.

Perhaps in many years time, we may find that the numerous present-day Boolean logic modules of AND and OR, NAND and NOR, etc., will have become obsolete, being superseded entirely by universal threshold modules. Such universal modules will possess a large fan-in and high-value thresholds, and will enable all the usual Boolean functions as well as non-Boolean threshold functions to be realized by different patterns of connexion of the input points. The future only can say whether this prophecy is correct, and how soon it may be fulfilled.

10. References and Bibliography

1. Dertouzos, M. L., 'Threshold Logic: A Synthesis Approach' (M.I.T. Press, Cambridge, Mass., 1965).
2. Hu, S. T., 'Threshold Logic' (University of California Press, 1965).
3. Lewis, P. M. and Coates, C. L., 'Threshold Logic' (Wiley, New York, 1967).
4. Lewis, P. M., 'Practical guide to threshold logic', *Electronic Design*, **22**, pp. 66–68, 25th October 1967.
5. Winder, R. O., 'Threshold logic will cut costs', *Electronics*, **41**, No. 11, pp. 94–103, May 1968.
6. Torng, H. C., 'An approach for the realization of linearly separable functions', *I.E.E.E. Trans. on Electronic Computers*, **EC-15**, No. 1, pp. 14–20, February 1966.
7. Dertouzos, M. L., 'An approach to single-threshold-element synthesis', *ibid*, **EC-13**, No. 5, pp. 519–28, October 1964.
8. Coates, C. L. and Lewis, P. M., 'DONUT: a threshold gate computer', *ibid*, **EC-13**, No. 3, pp. 240–47, June 1964.
9. McNaughton, R., 'Unate truth functions', *ibid*, **EC-10**, No. 1, pp. 1–6, March 1961.
10. Amarel, S., Cooke, G. and Winder, R. O., 'Majority gate networks', *ibid*, **EC-13**, No. 1, pp. 4–13, February 1964.
11. Sheng, C. L., 'A method for testing and realization of threshold functions', *ibid*, **EC-13**, No. 3, pp. 232–39, June 1964.
12. Akers, S. B., 'A diagrammatic approach to multilevel logic synthesis', *ibid*, **EC-14**, No. 2, pp. 174–81, April 1965.
13. Winder, R. O., 'Enumeration of seven-argument threshold functions', *ibid*, **EC-14**, No. 3, pp. 315–25, June 1965.
14. Price, J. E., 'Counting with majority-logic networks', *ibid*, **EC-14**, No. 2, pp. 256–60, April 1965.
15. Amodi, J. J. *et al.*, 'An integrated threshold gate', 1967 International Solid-State Circuits Conference Digest of Technical Papers, University of Pennsylvania, pp. 114, 115.
16. Blomgren, G. H. and Torng, H. C., 'Single threshold device realization subject to sensitivity requirements', *J. Franklin Inst.*, **281**, No. 2, pp. 143–53, February 1966.
17. Winder, R. O., 'Fundamentals of Threshold Logic', Air Force Cambridge Research Laboratories Report AFCRL–68–0066, Mass., January 1968.

Manuscript first received by the Institution on 1st November 1968 and in final form on 20th February 1969 (Paper No. 1260/Comp. 117.)

Forthcoming Conferences

Systems Engineering

The Institution is co-sponsor with the I.E.E. of a vacation school on Systems Engineering, which is to be held at the University of Lancaster from 8th to 12th September 1969. Systems engineering has been defined as 'multi-disciplinary common sense' and the school, which is aimed at people with some years' experience of project or departmental management, as well as at engineers with specialist interests, will present a broad survey of the available techniques and indicate areas where these can best be applied. Lectures will cover the following subjects:

Role of systems engineering; investment decisions and industrial planning; models for plant design; design of experiments; inventory and production control; simulation techniques; real-time systems; data presentation for management and control; linear and non-linear programming; control system design; estimation of plant dynamics; analogue and hybrid computing and simulation languages; principles and applications of on-line computing.

Further information and registration are available from the Control and Automation Division, I.E.E., Savoy Place, London, W.C.2.

Joint Conference on Automatic Test Systems

Organized by the Institution of Electronic and Radio Engineers with the association of other Institutions concerned with the subject, a conference on Automatic Test Systems will be held at the University of Birmingham from 12th to 16th April 1970.

The conference will be concerned with the present state of the automatic testing field and the probable direction of future development. Papers will deal with industrial, aviation and military applications, the techniques available and under development, and the operational problems that are presented by the employment of automatic test systems.

Further information, including the 'Call for Papers' will be published in the next issue of the *Journal.*

Mechanized Information Storage and Retrieval Systems

The Second International Conference on Mechanized Information Storage and Retrieval Systems will be held at the College of Aeronautics, Cranfield, from 2nd to 5th September 1969; it is sponsored jointly by the College and the journal *Information Storage and Retrieval.* The theme of the conference will be the economic aspects of mechanized systems; this is not to say that all the papers will present detailed cost figures, but rather that where an operational system is being discussed, the speaker will include information about the true cost of the service. In those papers which consider future systems an attempt will be made to relate the possible costs and the economic benefits that might be expected.

Further information may be obtained from Mr. C. W. Cleverdon, College of Aeronautics, Cranfield, Bedfordshire.

Electrical Methods of Machining, Forming and Coating

The I.E.E., in association with the Institution of Mechanical Engineers and the Institution of Production Engineers, is organizing a conference on the above subject, to be held at the London Headquarters of the I.E.E. in Savoy Place, from 17th to 19th March 1970.

The conference will be concerned with machining and forming by novel means, including laser machining; electron beams; electrochemical processes; spark, plasma and glow discharge processes; electrohydraulic and electromagnetic methods. Recent progress in such methods and their industrial applications, as well as economic considerations, will be discussed. Progress in the coating of materials by electrical methods such as electrophoresis and electrostatic powder deposition will be included.

The organizing committee invites papers for inclusion in the programme. Intending authors should submit synopses without delay. Complete manuscripts will be required before 13th October. Further details and registration forms are obtainable from the Conference Department, I.E.E., Savoy Place, London, W.C.2.

Computer-Aided Design

Following the recent I.E.E.-I.E.R.E. Computer-Aided Design Conference and E.E.A. Exhibition at Southampton University, it has been decided to hold a similar event in the Spring of 1971. The conference will again be organized by the I.E.E. in association with other learned societies, and a supporting exhibition, 'CADEX '71', will also be organized by the Electronic Engineering Association. A call for papers and further details, including venues and dates, will be published later.

U.D.C. 621.372.632

Cross-modulation and Intermodulation Distortions in the Tuned Square-law Diode Frequency Converter

By

J. G. GARDINER
B.Sc., Ph.D. (*Graduate*)†

Summary: Diodes possessing very accurate square-law I/V characteristics have recently been developed. An analytical technique is set out in this paper for the evaluation of cross-modulation and intermodulation distortions in a tuned frequency-converter using such diodes. The theory is used to predict the distortions obtained under practical conditions and it is shown that, where limited local-oscillator power is available, such a converter potentially generates less non-linear distortion than circuits employing conventional diodes as resistive switches.

List of Principal Symbols

i	current in diode produced by voltage v across terminals
i_1, i_2	input signal current sources at angular frequencies ω_1 and ω_2
i_c	local-oscillator current source at ω_c
i_c'	local-oscillator current in diode produced by voltage v_c across terminals
i_m	i_2 amplitude-modulated by frequency ω_m
I	total of impressed current sources
i'	current through diode produced by voltage V across terminals
L	conversion loss of converter
m	depth of modulation
Q	ratio β/α
v_0	d.c. bias voltage
v_f	voltage across diode at frequency f
Y_f	admittance of circuitry at frequency f as seen from diode terminals
α	coefficient of linear-term in diode characteristic referred to bias point
β	coefficient of squared-term in diode characteristic referred to bias point
α'	coefficient of linear-term in diode characteristic as measured
β'	coefficient of squared-term in diode characteristic as measured
δ	ratio i_2/i_c

1. Introduction

Increasingly severe demands are now being made on the performance of frequency converters in h.f. receiving equipment, particularly with regard to distortion. Converters are needed which are capable of handling simultaneously signals differing in level by more than 100 dB without loss of the wanted low-level sidebands due to cross-modulation from the stronger signal or by obliteration by intermodulation products generated by several large out-of-band signals.

Modulators utilizing parametric effects have been reported but these devices appear to need large pump powers to achieve acceptable linearity[1] and are unsuitable for high ratio down-conversion.

The purpose of this paper is to examine the properties of a tuned frequency converter using the square-law diode recently developed by Wright[2] and to show that it is capable of good cross- and inter-modulation performance under conditions where large local-oscillator powers cannot be supplied. This advantage is achieved at the expense of increased conversion loss but, since the device relies on a space-charge-limited current mechanism and is therefore low-noise, this may not represent a significant drawback, particularly since the performance is maintained for high ratio down-conversion.

2. General Equations for the Narrow-band Frequency Converter

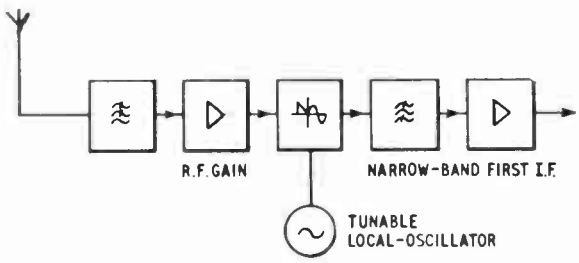
Figure 1(a) shows a typical receiver system requirement where the first i.f. may be higher or lower than the received band of signals. For simplicity, the mixer will be considered to be a single diode device as shown in Fig. 1(b) with selective circuits imposing constraints on the number of voltages appearing across the diode itself. The diode is assumed to possess the characteristic

$$i = \alpha v + \beta v^2 \qquad \ldots\ldots(1)$$

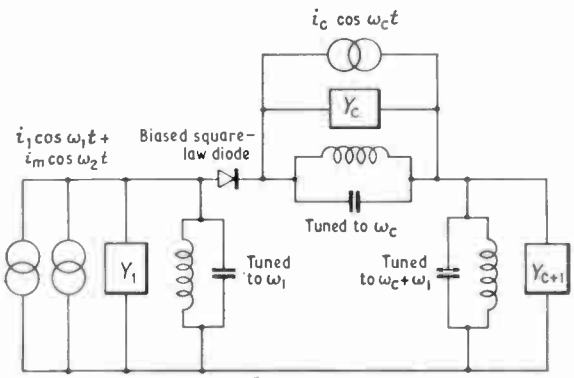
and an equivalent circuit for the converter takes the form of Fig. 2.

If the admittance Y is assumed to take finite values at a signal frequency ω_q and local-oscillator frequency ω_c and to be a short-circuit at all other frequencies

† Formerly with Racal Communications Ltd., Tewkesbury, Gloucestershire; now at the Postgraduate School of Electrical and Electronic Engineering, University of Bradford.



(a) First frequency conversion in typical receiver system.



(b) Schematic of tuned series converter.

Fig. 1.

then the current flowing in this short-circuit contains components at only the impressed frequencies, their second harmonics and the first modulation sidebands. If now Y takes a finite value at a sideband frequency in order to extract useful power then the sideband voltage appears across the diode terminals and a process of successive inter-modulation occurs. It is apparent that a very large number of product currents and voltages exist in the circuit which are comparable in magnitude to the distortion terms which it is desired to isolate. In consequence, analysis based on the assumption that the output spectrum can be described in terms of a restricted number of frequency components is inapplicable even though such an approach is adequate for calculation of losses to large signal components. A series solution for low-level product components may be obtained, however, by defining the voltage across the diode in terms of the coefficient of the squared-term in the diode characteristic. This may be demonstrated as follows:

The circuit equation for Fig. 2 may be written

$$\alpha v + \beta v^2 + Yv = I \qquad \ldots\ldots(2)$$

where I is the sum total of all impressed currents and is defined in general as

$$I = \sum_{n=0}^{\infty} \tfrac{1}{2}(i_n \, e^{j\omega_n t} + i_n^* \, e^{-j\omega_n t}) \qquad \ldots\ldots(3)$$

As suggested above, the voltage appearing across the diode and representing the combined voltage produced by all impressed currents and generated products may be defined in terms of β as

$$v = \sum_{k=0}^{\infty} \beta^k \lambda_k \qquad \ldots\ldots(4)$$

where λ_k are expressions containing components at a number of frequencies, this number increasing as k increases.

Equation (4) may now be substituted into eqn. (2) to give

$$\alpha \sum_{k=0}^{\infty} \beta^k \lambda_k + \beta \left\{ \sum_{k=0}^{\infty} \beta^k \lambda_k \right\}^2 + Y \sum_{k=0}^{\infty} \beta^k \lambda_k = \sum_{n=0}^{\infty} \tfrac{1}{2}(i_n \, e^{j\omega_n t} + i_n^* \, e^{-j\omega_n t}) \qquad \ldots\ldots(5)$$

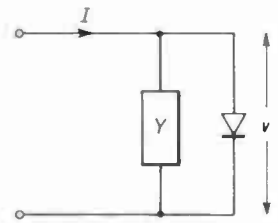


Fig. 2. Equivalent circuit of Fig. 1(b).

The object of defining v in terms of β is now apparent since by equating coefficients of powers of β on both sides of eqn. (5) (which is a valid step since α, λ, Y and i_n are all independent of β), it is possible to define λ_0 in terms of Y, α and i_n. λ_1 may then be defined in terms of λ_0, λ_n in terms of λ_{n-1}, $\lambda_{n-2} \ldots \lambda_0$ and so on. Provided that for any given frequency the series of coefficients is convergent, i.e. that the coefficient derived from λ_n of any frequency component is always less than the coefficient of the same frequency derived from any lower-order λ term, then the magnitude of the component at any required product frequency may be determined to any desired degree of accuracy by evaluating successive λ expressions and extracting the wanted component. It is seen by inspection that the series is rapidly convergent, evaluation of each successive λ term involves division of coefficients in lower-order expressions by admittances $(Y+\alpha)$. Therefore, equating coefficients of $(\beta)^0$ on both sides of eqn. (5) leads to

$$\lambda_0 = \frac{1}{Y+\alpha} \sum_{n=0}^{\infty} \tfrac{1}{2}(i_n \, e^{j\omega_n t} + i_n^* \, e^{-j\omega_n t}) \qquad \ldots\ldots(6)$$

Coefficients of β give

$$\lambda_1 (Y+\alpha) + \lambda_0^2 = 0 \qquad \ldots\ldots(7)$$

and so on, giving

$$\lambda_1 = \frac{-\lambda_0^2}{Y+\alpha} \qquad \text{.....(8)}$$

$$\lambda_2 = \frac{-2\lambda_1\lambda_0}{Y+\alpha} \qquad \text{.....(9)}$$

$$\lambda_3 = \frac{-(\lambda_1^2+2\lambda_0\lambda_2)}{Y+\alpha} \qquad \text{.....(10)}$$

$$\lambda_4 = \frac{-2(\lambda_2\lambda_1+\lambda_3\lambda_0)}{Y+\alpha} \qquad \text{.....(11)}$$

$$\lambda_5 = \frac{-(\lambda_2^2+2\lambda_1\lambda_3+2\lambda_0\lambda_4)}{Y+\alpha} \qquad \text{.....(12)}$$

etc.

In evaluating individual coefficients in the λ expressions Y takes the value of the total circuit admittance seen by the diode at the frequency of the component on which it operates. It will be seen that even if it is assumed that the tuned circuits supporting input signal, local-oscillator and wanted-sideband frequencies are sharply resonant, sufficient successive modulations occur to produce unwanted distortion products in low-order λ expressions.

3. Cross-modulation Performance

The spectrum assumed for analysis is shown in Fig. 3 and represents a typical first-mixer up-conversion at h.f. (e.g. 10 MHz to 46·5 MHz i.f.). The two input signals are designated ω_1 and ω_2 with corresponding currents i_1 and i_m. The signal at ω_2 is assumed to be a strong interfering input and to be modulated to a depth of $100\,m\%$ (usually 30% in practice), i.e.

$$i_m = i_2(1+m\cos\omega_m t) \qquad \text{.....(13)}$$

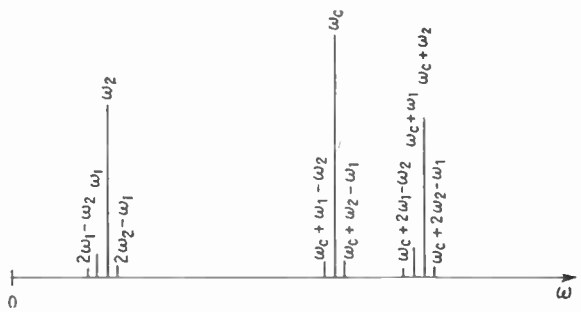


Fig. 3. Spectrum of frequencies for cross-modulation calculation.

The analysis is not restricted to up-conversion since the amplitudes of all generated products are independent of frequency. Thus in down-conversion with local-oscillator injection at a lower frequency than the received signal the wanted down-converted carrier becomes $(\omega_1-\omega_c)$, the high-level carrier $(\omega_2-\omega_c)$ and the distortion products close to the output sidebands become $(2\omega_1-\omega_2-\omega_c)$ and $(2\omega_2-\omega_1-\omega_c)$ instead of $(2\omega_2-\omega_1+\omega_c)$ and $(2\omega_1-\omega_2+\omega_c)$. But $(\omega_1-\omega_c)$ and $(\omega_c+\omega_1)$ have the same amplitude, as do all the above distortion products so that no difference appears in the final result. The same argument holds for down-conversion with injection at a higher frequency than the received signal.

A large number of possible intermodulations arise, the ones of interest being those which result in products of the form $(i_m K\cos\omega_c+\omega_1 t)$ where K is some coefficient and ω_c the local-oscillator frequency, since this represents transfer of the modulating frequency ω_m applied to i_2 to the wanted small signal at the converter output.

It is assumed that small but finite bandwidths exist for the tuned circuits so that products of the form $(\omega_c+\omega_2-\omega_1)$ etc., are terminated in the local-oscillator filter, $(\omega_c+2\omega_1-\omega_2)$ terminated in the upper-sideband filter and so on, while components separated from the filter midbands by appreciable percentages of the centre frequencies (e.g. $2\omega_1$, $2\omega_2$, etc.) are considered to be short circuited. The d.c. path to rectified current may be regarded as a short circuit since whatever d.c. is generated by the non-linear device would be compensated for by adjustment of the initial conditions of bias voltage.

Returning to eqn. (5) we may now define

$$\sum_{n=0}^{\infty} \tfrac{1}{2}(i_n\,e^{j\omega_n t}+i_n^*\,e^{-j\omega_n t}) = i_1\cos\omega_1 t+i_m\cos\omega_2 t+i_c\cos\omega_c t \qquad \text{.....(14)}$$

since no harmonic relationships exist between the modulating frequencies so that relative phases are immaterial.

In evaluating the λ terms the following subscript notation will be used:

Subscripts relating to admittances indicate the frequency at which the value of Y is considered, i.e. Y_{c+22-1} indicates the value of Y at $(\omega_c+2\omega_2-\omega_1)$.

The coefficients K, of the λ expressions are identified similarly:

$K_{p(c+22-1)}$ is the coefficient in λ_p of the component at $(\omega_c+2\omega_2-\omega_1)$.

The expressions for λ are developed in full in Appendix 1 and these expressions contain components at the various frequencies as follows:

λ_0 contains	ω_1	ω_2	ω_c
λ_1 contains	$\omega_c+\omega_1$		$\omega_c+\omega_2$
λ_2 contains	ω_c	ω_1	ω_2
	$\omega_c+\omega_1-\omega_2$		$\omega_c+\omega_2-\omega_1$
λ_3 contains	$\omega_c+\omega_1$		$\omega_c+\omega_2$
	$\omega_c+2\omega_1-\omega_2$		$\omega_c+2\omega_2-\omega_1$
λ_4 contains	ω_1	ω_2	ω_c
	$\omega_c+\omega_2-\omega_1$		$\omega_c+\omega_1-\omega_2$
	$\omega_c+2\omega_1-2\omega_2$		$\omega_c+2\omega_2-2\omega_1$
	$2\omega_1-\omega_2$		$2\omega_2-\omega_1$

It is immediately seen that only odd-order λ terms contain the wanted output frequency $\omega_c+\omega_1$ and the components of the λ expressions decrease in magnitude rapidly with increasing order of λ (by reason of repeated division by admittance terms) so that only components in λ_1 and λ_3 will be considered in evaluating the cross-modulation performance.

A 'cross-modulation factor', $\phi_{c.m.}$ may now be defined as the ratio of cross-modulated sideband voltage amplitude in the load termination to the total amplitude at the desired output frequency, i.e.

$$\phi_{c.m.} = |v_{c+1\pm m}/v_{c+1}| \qquad \text{.....(15)}$$

Now from equation (4)

$$v_{(c+1\pm m)}+v_{(c+1)} = \beta K_{1(c+1)}+\beta^3 K_{3(c+1)} \qquad \text{.....(16)}$$

Also, as shown in Appendix 1,

$$K_{1(c+1)} = \frac{-i_c i_1}{(Y_1+\alpha)(Y_c+\alpha)(Y_{c+1}+\alpha)} \qquad \text{.....(17)}$$

$$K_{3(c+1)} = \frac{-1}{(Y_{c+1}+\alpha)}\left[\frac{i_1 K_{2c}}{Y_1+\alpha}+\frac{i_m K_{2(c+1-2)}}{Y_1+\alpha}+\frac{i_c K_{21}}{Y_c+\alpha}\right] \qquad \text{.....(18)}$$

$$K_{2c} = \frac{i_c}{(Y_1+\alpha)^2(Y_c+\alpha)^2}\left[\frac{i_1^2}{Y_{c+1}+\alpha}+\frac{i_m^2}{Y_{c+2}+\alpha}\right] \qquad \text{.....(19)}$$

$$K_{2(c+1-2)} = \frac{i_1 i_c i_m}{(Y_1+\alpha)^2(Y_c+\alpha)(Y_{c+1}+\alpha)(Y_{c+1-2}+\alpha)} \qquad \text{.....(20)}$$

$$K_{21} = \frac{i_c^2 i_1}{(Y_1+\alpha)^2(Y_c+\alpha)^2(Y_{c+1}+\alpha)} \qquad \text{.....(21)}$$

From eqn. (12) we know that

$$i_m^2 = i_2^2\left[1+2m\cos\omega_m t+\frac{m^2}{2}(1+\cos 2\omega_m t)\right] \qquad \text{.....(22)}$$

and if m is small (e.g. 0·3) then the $m^2/2$ term may be neglected and eqns. (16)–(21) combined to evaluate the voltages at $(\omega_c+\omega_1)$ and $(\omega_c+\omega_1\pm\omega_m)$. This gives

$$v_{c+1} = \left[\frac{\beta i_c i_1}{(Y_1+\alpha)(Y_c+\alpha)(Y_{c+1}+\alpha)}+\frac{\beta^3 i_c^3 i_1}{(Y_c+\alpha)^3(Y_{c+1}+\alpha)^2(Y_1+\alpha)^2}+\frac{\beta^3 i_c i_1^3}{(Y_1+\alpha)^3(Y_c+\alpha)^2(Y_{c+1}+\alpha)^2}+\right.$$
$$\left.+\frac{\beta^3 i_1 i_c i_2^2}{(Y_{c+1-2}+\alpha)(Y_1+\alpha)^3(Y_{c+1}+\alpha)^2(Y_c+\alpha)}+\frac{\beta^3 i_c i_1 i_2^2}{(Y_1+\alpha)^3(Y_c+\alpha)^2(Y_{c+1}+\alpha)(Y_{c+2}+\alpha)}\right] \qquad \text{.....(23)}$$

and

$$v_{c+1\pm m} = \frac{m\beta^3 i_1 i_c i_2^2}{(Y_{c+1-2}+\alpha)(Y_1+\alpha)^3(Y_{c+1}+\alpha)^2(Y_c+\alpha)}+\frac{m\beta^3 i_1 i_c i_2^2}{(Y_1+\alpha)^3(Y_c+\alpha)^2(Y_{c+1}+\alpha)(Y_{c+2}+\alpha)} \qquad \text{.....(24)}$$

Thus, in general, the narrow-band cross-modulation factor reduces to

$$\phi_{c.m.} = \frac{\dfrac{mi_2^2\beta^2}{(Y_1+\alpha)^2}\left[\dfrac{1}{(Y_{c+1}+\alpha)(Y_{c+1-2}+\alpha)}+\dfrac{1}{(Y_{c+2}+\alpha)(Y_c+\alpha)}\right]}{\left\{1+\dfrac{\beta^2}{(Y_{c+1}+\alpha)(Y_1+\alpha)(Y_c+\alpha)}\left[\dfrac{i_c^2}{Y_c+\alpha}+\dfrac{i_1^2}{Y_1+\alpha}\right]+\dfrac{i_2^2\beta^2}{(Y_1+\alpha)^2}\times\right.}$$
$$\left.\times\left[\frac{1}{(Y_{c+1-2}+\alpha)(Y_{c+1}+\alpha)}+\frac{1}{(Y_c+\alpha)(Y_{c+2}+\alpha)}\right]\right\} \qquad \text{.....(25)}$$

Thus far, in order to preserve generality, the circuit admittances have been assumed to be different at all frequencies, but in any practical assessment of the cross-modulation performance the input signals are very close together in frequency (typically 20 kHz separation at 30 MHz) so that the $\omega_c+\omega_1$ sideband sees the same admittance as $\omega_c+\omega_2$; $\omega_c\pm\omega_1\mp\omega_2$, lying close to the local-oscillator frequency see the same admittance as ω_c. Thus

$$Y_{c+1} = Y_{c+2} \qquad \text{.....(26)}$$

$$Y_{c+1-2} = Y_c \qquad \text{.....(27)}$$

Also, it may readily be shown that the optimum termination for minimum conversion loss to the wanted sideband at $\omega_c+\omega_1$ is achieved by making

$$Y_{c+1} = Y_1. \qquad \text{.....(28)}$$

Applying these conditions to eqn. (24) the expression reduces to

$$\phi_{c.m.} = \frac{2mi_2^2\beta^2}{(Y_1+\alpha)^3(Y_c+\alpha)+\beta^2\left\{i_c^2\frac{Y_1+\alpha}{Y_c+\alpha}+i_1^2+2i_2^2\right\}} \qquad \text{.....(29)}$$

may be defined by the ratios

$$\phi_{i.m.} = \frac{v_{c+21-2}}{v_{c+1}} = \frac{v_{c+22-1}}{v_{c+2}} \qquad \text{.....(30)}$$

As in the case of cross-modulation, the nature of the conversion, i.e. up-conversion or down-conversion with high or low side injection, is immaterial to the final result.

The product voltages in question are readily deduced from the expressions already evaluated in Appendix 1 by putting

$$i_1 = i_2 = i_s \quad \text{and} \quad i_m = i_2$$

Then the level of the intermodulation product $(\omega_c+2\omega_1-\omega_2)$ is given by

$$v_{c+21-2} = \frac{\beta^3 i_s K_{2(c+1-2)}}{(Y_1+\alpha)(Y_{c+21-2}+\alpha)} \qquad \text{.....(31)}$$

and

$$K_{2(c+1-2)} = \frac{i_s^2 i_c'}{(Y_1+\alpha)^2(Y_c+\alpha)(Y_{c+1}+\alpha)(Y_{c+1-2}+\alpha)} \qquad \text{.....(32)}$$

this gives

$$\phi_{i.m.} = \frac{i_s^2\beta^2}{\left\{1+\beta^2\left[\frac{i_s^2}{(Y_1+\alpha)^2(Y_c+\alpha)(Y_{c+1}+\alpha)}+\frac{i_s^2}{(Y_1+\alpha)^2}\left(\frac{1}{(Y_{c+1-2}+\alpha)(Y_{c+1}+\alpha)}+\frac{1}{(Y_{c+2}+\alpha)(Y_c+\alpha)}\right)+\frac{i_c^2}{(Y_1+\alpha)^2(Y_c+\alpha)(Y_{c+1}+\alpha)}\right]\right\}[(Y_1+\alpha)^2(Y_{c+21-2}+\alpha)(Y_{c+1-2}+\alpha)]} \qquad \text{.....(33)}$$

Numerical evaluation of this expression will be considered in a later section.

4. Intermodulation Performance

4.1. *Third-order Distortion*

Again assuming a similar up-conversion to that of Fig. 3, suppose the receiver to be tuned to a wanted signal at frequency ω_s but that large out-of-band signals exist at frequencies ω_1 and ω_2. If ω_1 and ω_2 are separated by frequency $\delta\omega$ and this frequency also separates ω_1 from ω_s then ω_2 is $2\delta\omega$ from ω_s and after conversion the intermodulation product $(\omega_c+2\omega_1-\omega_2)$ is coincident with the wanted signal at $(\omega_s+\omega_c)$. A measure of this type of distortion is obtained as shown in Fig. 4 by applying signals at ω_1 and ω_2 of equal strength and measuring the difference in level at the converter output between the products $(\omega_c+\omega_1)$, $(\omega_c+\omega_2)$ and $(\omega_c+2\omega_1-\omega_2)$, $(\omega_c+2\omega_2-\omega_1)$, i.e. an intermodulation factor, $(\phi_{i.m.})$

Making the same assumptions as before concerning impedances over the filter bandwidths gives

$$\phi_{i.m.} = \frac{i_s^2\beta^2\left(\frac{Y_1+\alpha}{Y_c+\alpha}\right)}{(Y_1+\alpha)^4+\beta^2\left(\frac{Y_1+\alpha}{Y_c+\alpha}\right)\left[3i_s^2+i_c^2\left(\frac{Y_1+\alpha}{Y_c+\alpha}\right)\right]} \qquad \text{.....(34)}$$

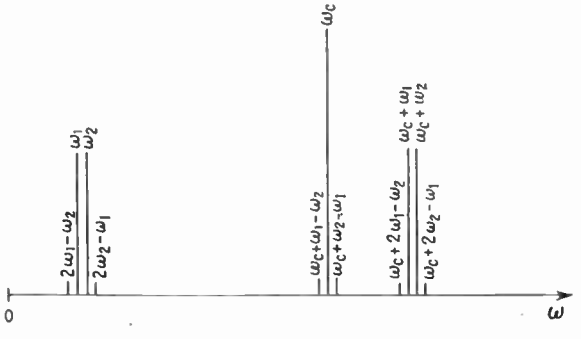


Fig. 4. Spectrum of frequencies for intermodulation calculation.

4.2. *Higher-order Distortion*

Whilst it is usual to specify intermodulation distortion performance in terms of third-order products, higher-order intermodulations can produce unwanted signals in the receiver i.f. band. Which products in fact prove troublesome depends on the input-signal frequency band and the chosen i.f. In the set-up of Fig. 4 the fifth-order products ($\omega_c \pm 3\omega_1 \mp 2\omega_2$) are the next in importance to the third. These arise in λ_5 and may be calculated from eqn. (13).

It can be seen from Appendix 1 that the fifth-order distortion terms in question occur only in the product $2\lambda_0\lambda_4$ and that their amplitude is given by the product $K_{01}K_{4(c+21-22)}$. As before, we may calculate a ratio, ϕ' (i.m.) of fifth-order products to the frequency converted interfering signals. Ignoring contributions to v_{c+1} from higher-order mixes than occur in λ_3 this gives

$$\phi'_{\text{i.m.}} = \frac{\dfrac{i_s^4\beta^4}{(Y_1+\alpha)^4(Y_{c+21-22}+\alpha)(Y_{c+21-2}+\alpha)(Y_{c+31-22}+\alpha)(Y_{c+1-2}+\alpha)}}{\left\{1+\dfrac{\beta^2}{(Y_1+\alpha)(Y_{c+1}+\alpha)(Y_c+\alpha)}\left[\dfrac{i_c^2}{Y_c+\alpha}+i_s^2\left(\dfrac{1}{Y_1+\alpha}+\dfrac{Y_c+\alpha}{(Y_1+\alpha)(Y_{c+1-2}+\alpha)}+\dfrac{Y_{c+1}+\alpha}{(Y_1+\alpha)(Y_{c+2}+\alpha)}\right)\right]\right\}} \quad \ldots\ldots(35)$$

Making the same assumption as before concerning admittances seen by the various products this reduces to

$$\phi'_{\text{i.m.}} = \frac{\dfrac{\beta^4 i_s^4}{(Y_1+\alpha)^2(Y_c+\alpha)^2}}{(Y_1+\alpha)^4+\beta^2\left(\dfrac{Y_1+\alpha}{Y_c+\alpha}\right)\left[3i_s^2+i_c^2\left(\dfrac{Y_1+\alpha}{Y_c+\alpha}\right)\right]} \quad \ldots\ldots(36)$$

Other orders of intermodulation product may be evaluated in the same way.

5. Numerical Evaluation of $\phi_{\text{c.m.}}$ and $\phi_{\text{i.m.}}$

5.1. *Terminations and Bias*

So far we have considered a device solely in terms of the characteristic of eqn. (1) and it is now necessary to determine the bias and matching impedance conditions which must be used in practice.

The practical diode has a characteristic as shown in Fig. 5 and it is known that the optimum efficiency of conversion is obtained when the diode conductance variation in time has as large a component at the local-oscillator frequency as possible relative to the mean value.[3] It is shown in Appendix 2 that this is achieved by allowing the local oscillator to swing the operating point over as large a portion of the diode characteristic as possible while at the same time remaining within the square-law region of the curve. In the case of small input signals this is readily accomplished by arranging the d.c. bias voltage to be half the peak-to-peak local-oscillator swing, but the presence of a large interfering signal, under such conditions, could cause the total voltage across the diode to exceed the permissible limits. Consequently, allowance must be made in assessing the bias voltage for the worst case likely to arise, i.e. the largest interfering signal. Terminations, on the other hand, will be considered fixed at the values for minimum conversion loss to the small wanted signal.

It is shown in Appendix 2 that the optimum values for the signal circuit terminations are

$$Y_{c+1} = Y_1 = \alpha\sqrt{3/2} \quad \ldots\ldots(37)$$

and for the local oscillator,

$$Y_c = \alpha \quad \ldots\ldots(38)$$

and with these values we may now calculate the voltage appearing across the diode due to the large interfering current i_2 (i_m). This is obtained from λ_0 and λ_2 in Appendix 1 and gives

$$v_2 = \frac{i_2}{(Y_1+\alpha)} + \frac{\beta^2 i_2 i_c^2}{(Y_1+\alpha)(Y_c+\alpha)^2(Y_{c+2}+\alpha)} \quad \ldots\ldots(39)$$

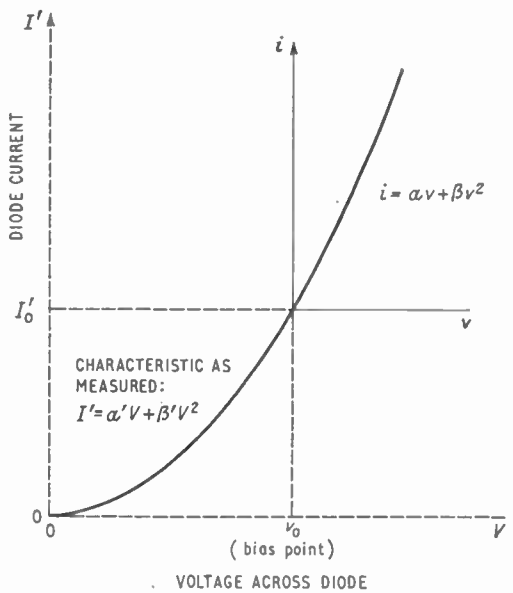


Fig. 5. Relationship between diode characteristic as measured and characteristic used in analysis.

and

$$v_c = \frac{i_c}{(Y_c+\alpha)} + \frac{\beta^2 i_c}{(Y_1+\alpha)^2(Y_c+\alpha)^2} \times \times \left[\frac{i_1^2}{(Y_{c+1}+\alpha)} + \frac{i_2^2}{(Y_{c+2}+\alpha)}\right] \quad \ldots\ldots(40)$$

Now the second term in eqn. (40) may be ignored since $i_1 \ll i_2$ and the i_2^2 term will not be significant in comparison with the first term since $i_2 < 0{\cdot}5 i_c$ in practice. Defining

$$i_2 = \delta i_c \quad \ldots\ldots(41)$$

the largest voltage to appear across the diode (neglecting sidebands) is

$$v_c + v_2 = \frac{i_c}{(Y_c+\alpha)} + \frac{\delta i_c}{(Y_1+\alpha)} \times \times \left[1 + \frac{\beta^2 i_c^2}{(Y_c+\alpha)^2(Y_1+\alpha)^2}\right] \quad \ldots\ldots(42)$$

Now it is also shown in Appendix 2 that

$$\beta^2 i_c^2 \simeq \frac{\alpha^4}{(1+\delta)^2} \quad \ldots\ldots(43)$$

and substituting values for Y_c and Y_1 we find that to a first approximation

$$v_c + v_2 = v_c(1+\delta) \quad \ldots\ldots(44)$$

Thus this value represents the maximum voltage swing across the diode and fixes v_0 for a given value of i_c.

5.2. *Conversion Loss*

We are now in a position to evaluate the modulator conversion loss under the large signal conditions. Again from Appendix 1 the voltages across source and load at input-signal and sideband frequencies can be found and since $Y_{c+1} = Y_1$ the conversion loss may be written

$$L = 20 \log \frac{v_1}{v_{c+1}} \quad \ldots\ldots(45)$$

The complete expression becomes

$$L = 20 \log \left\{ \frac{1 + \dfrac{\beta^2 i_c^2}{(Y_c+\alpha)^2(Y_1+\alpha)^2}}{\dfrac{\beta i_c}{(Y_c+\alpha)(Y_1+\alpha)}\left[1 + \dfrac{\beta^2 i_c^2}{(Y_c+\alpha)^2(Y_1+\alpha)^2} + \dfrac{\delta^2\beta^2 i_2^2}{(Y_c+\alpha)^4(Y_1+\alpha)^3}\right]} \right\} \quad \ldots\ldots(46)$$

Now substituting for Y_c, Y_1 and βi_c, L reduces to

$$L = 20 \log \frac{3{\cdot}732(1+\delta)[(3{\cdot}732)^2(1+\delta)^2+1]}{(3{\cdot}732)^2(1+\delta)^2+1{\cdot}07(1+\delta^2)} \quad \ldots\ldots(47)$$

It is clear that for values of δ of 0·3 or less this expression reduces to

$$L = 20 \log (3{\cdot}732(1+\delta)) \quad \ldots\ldots(48)$$

and for $\delta \ll 1$, i.e. the small-signal modulator, this loss has the value of 11·4 dB which is in agreement with the predictions of linear time-varying theory.[4]

5.3. *Theoretical Results*

It is now possible to examine the performance of a square-law diode modulator as the input level is increased relative to that of the local oscillator.

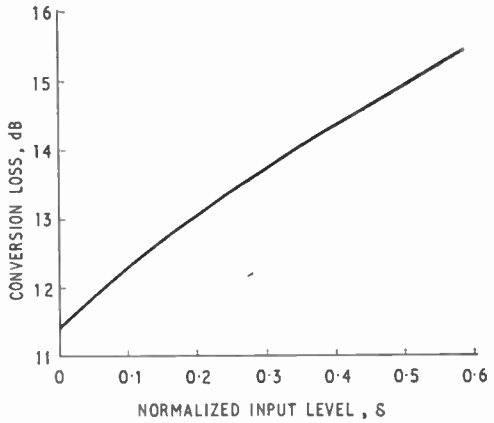


Fig. 6. Variation of conversion loss with normalized input level.

In the graph of Fig. 6, eqn. (47) is used to plot the variation in conversion loss which results from increasing the bias voltage v_0 so that signals are confined to the square-law region of the diode characteristic.

Equation (29) is used to give the plot of percentage cross-modulation against δ of Fig. 7. In this graph the performance of the square-law diode modulator is compared with that of a conventional resistively terminated ring modulator using switching diodes. The results for the latter circuit are derived from theory due to Tucker.[5] It is immediately apparent from this comparison that the square-law diode generates very much less distortion than the ring under large input signal conditions when restricted local-oscillator power is available or when the input signal levels are such that any attempt to achieve low distortion by increasing the local-oscillator level would cause the diodes to fail.

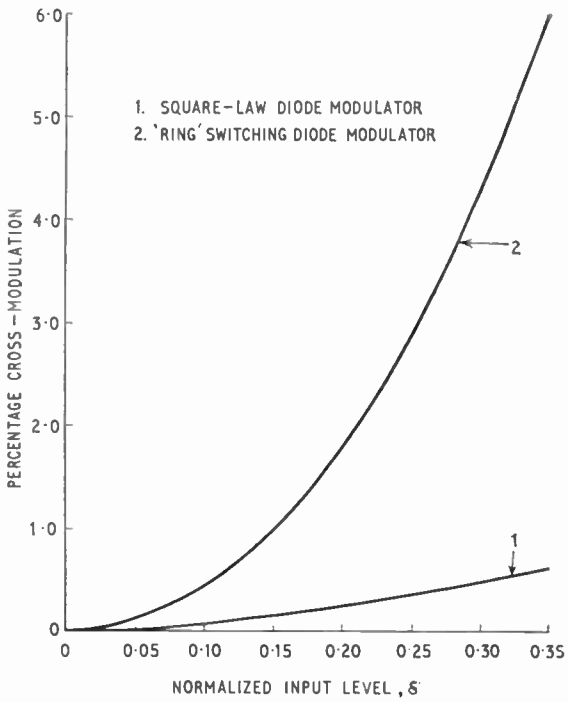


Fig. 7. Cross-modulation performance as a function of normalized input level.

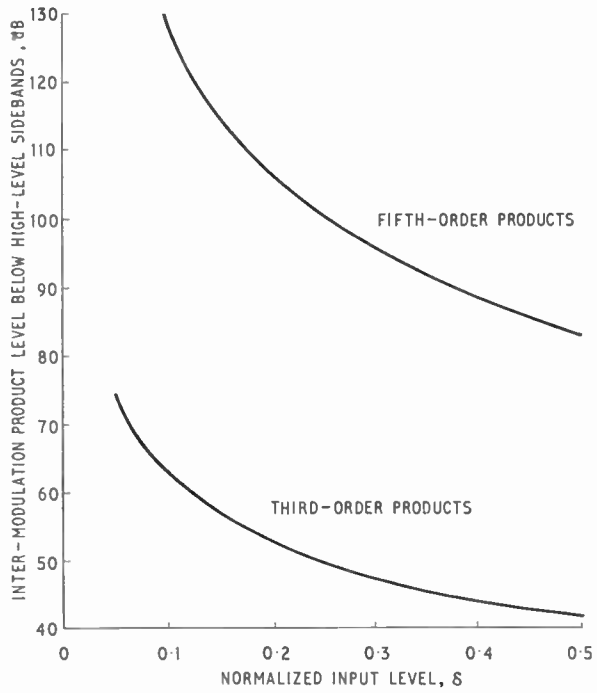


Fig. 8. Intermodulation performance as a function of normalized input level.

Figure 8 shows plots of the third- and fifth-order intermodulation product levels relative to the frequency-converted interfering signal outputs. It is seen that products above fifth-order are unlikely to prove of significance in a practical situation.

In all the above evaluations an ideal diode is assumed (i.e. $\alpha' = 0$) but this is by no means an unrealizable assumption. Conversion loss measurements on practical diodes have shown[4] that loss performances approaching the ideal can be achieved from prototype diodes with very low Q-factors (~ 3) by using relatively modest drive powers—typically 8–10 mW.

6. Conclusions

The cross-modulation and intermodulation performances of a tuned square-law diode converter have been evaluated and it is seen that this device is capable of frequency conversion with significantly less distortion than can be achieved with the ring switching mixer under identical operating conditions. The large conversion loss associated with this circuit is disadvantageous but the demands of linearity now made of converters are such that this order of loss will probably be regarded as acceptable in view of the other advantages of the device.

7. Acknowledgments

This work was supported in the Department of Electronic and Electrical Engineering of the University of Birmingham by the Racal Research Fellowship and the author is indebted to the Board of Directors, Racal Communications Ltd., for encouragement to publish this paper. Thanks are also due to Professor D. P. Howson of the School of Electrical and Electronic Engineering in the University of Bradford for a number of valuable suggestions.

8. References

1. Gikow, E., 'Wide dynamic range preselectors', *Trans. Inst. Elect. Electronics Engrs on Aerospace and Electronic Systems*, **AES-2**, pp. 94–101, January 1966.
2. Wright, G. T., 'Square-law silicon resistor, provisional specification', Department of Electronic and Electrical Engineering, Birmingham University, Memorandum, July 1966.
3. Howson, D. P., 'Single-balanced rectifier modulators', *Proc. Instn Elect. Engrs*, **109**, Part C, pp. 357–64, January 1962. (I.E.E. Monograph No. 500E.)
4. Gardiner, J. G., 'Single-balanced modulators using square-law resistors (s.c.l. diodes)', *The Radio and Electronic Engineer*, **37**, No. 5, pp. 305-14, May 1969.
5. Tucker, D. G., 'Intermodulation distortion in rectifier modulators', *Wireless Engineer*, **31**, pp. 145–52, June 1954.

9. Appendix 1
Evaluation of λ Terms

$$\lambda_0 = K_{01}\cos\omega_1 t + K_{02}\cos\omega_2 t + K_{0c}\cos\omega_c t \qquad \text{.....(49)}$$

where

$$K_{01} = \frac{i_1}{Y_1+\alpha} \qquad \text{.....(50)}$$

$$K_{02} = \frac{i_m}{Y_1+\alpha} \qquad \text{.....(51)}$$

$$K_{0c} = \frac{i_c}{Y_c+\alpha} \qquad \text{.....(52)}$$

$$\lambda_1 = K_{1(c+1)}\cos(\omega_c+\omega_1)t + K_{1(c+2)}\cos(\omega_c+\omega_2)t \qquad \text{.....(53)}$$

where

$$K_{1(c+1)} = \frac{-i_1 i_c}{(Y_1+\alpha)(Y_c+\alpha)(Y_{c+1}+\alpha)} \qquad \text{.....(54)}$$

$$K_{1(c+2)} = \frac{-i_m i_c}{(Y_1+\alpha)(Y_c+\alpha)(Y_{c+2}+\alpha)} \qquad \text{.....(55)}$$

$$\lambda_2 = K_{2c}\cos\omega_c t + K_{21}\cos\omega_1 t + K_{22}\cos\omega_2 t + K_{2(c+1-2)}\cos(\omega_c+\omega_1-\omega_2)t + \\ + K_{2(c+2-1)}\cos(\omega_c+\omega_2-\omega_1)t \qquad \text{.....(56)}$$

where

$$K_{2c} = \frac{i_c}{(Y_1+\alpha)^2(Y_c+\alpha)^2}\left[\frac{i_1^2}{(Y_{c+1}+\alpha)} + \frac{i_m^2}{(Y_{c+2}+\alpha)}\right] \qquad \text{.....(57)}$$

$$K_{21} = \frac{i_c^2 i_1}{(Y_1+\alpha)^2(Y_c+\alpha)^2(Y_{c+1}+\alpha)} \qquad \text{.....(58)}$$

$$K_{22} = \frac{i_m i_c^2}{(Y_1+\alpha)^2(Y_c+\alpha)^2(Y_{c+2}+\alpha)} \qquad \text{.....(59)}$$

$$K_{2(c+1-2)} = \frac{i_1 i_c i_m}{(Y_1+\alpha)^2(Y_c+\alpha)(Y_{c+1}+\alpha)(Y_{c+1-2}+\alpha)} \qquad \text{.....(60)}$$

$$K_{2(c+2-1)} = \frac{i_1 i_c i_m}{(Y_1+\alpha)^2(Y_c+\alpha)(Y_{c+2}+\alpha)(Y_{c+2-1}+\alpha)} \qquad \text{.....(61)}$$

$$\lambda_3 = K_{3(c+1)}\cos(\omega_c+\omega_1)t + K_{3(c+2)}\cos(\omega_c+\omega_2)t + K_{3(c+21-2)}\cos(\omega_c+2\omega_1-\omega_2)t + \\ + K_{3(c+22-1)}\cos(\omega_c+2\omega_2-\omega_1)t \qquad \text{.....(62)}$$

where

$$K_{3(c+1)} = \frac{-i_1 K_{2c}}{(Y_1+\alpha)(Y_{c+1}+\alpha)} - \frac{i_m K_{2(c+1-2)}}{(Y_1+\alpha)(Y_{c+1}+\alpha)} - \frac{i_c K_{21}}{(Y_c+\alpha)(Y_{c+1}+\alpha)} \qquad \text{.....(63)}$$

$$K_{3(c+2)} = \frac{-i_m K_{2c}}{(Y_1+\alpha)(Y_{c+2}+\alpha)} - \frac{i_1 K_{2(c+2-1)}}{(Y_1+\alpha)(Y_{c+2}+\alpha)} - \frac{i_c K_{22}}{(Y_c+\alpha)(Y_{c+2}+\alpha)} \qquad \text{.....(64)}$$

$$K_{3(c+21-2)} = \frac{-i_1 K_{2(c+1-2)}}{(Y_1+\alpha)(Y_{c+21-2}+\alpha)} \qquad \text{.....(65)}$$

$$K_{3(c+22-1)} = \frac{-i_m K_{2(c+2-1)}}{(Y_1+\alpha)(Y_{c+22-1}+\alpha)} \qquad \text{.....(66)}$$

$$\lambda_4 = K_{41}\cos\omega_1 t + K_{42}\cos\omega_2 t + K_{4c}\cos\omega_c t + K_{4(c+2-1)}\cos(\omega_c+\omega_2-\omega_1)t + \\ + K_{4(c+1-2)}\cos(\omega_c+\omega_1-\omega_2)t + K_{4(c+21-22)}\cos(\omega_c+2\omega_1-2\omega_2)t + \\ + K_{4(c+22-21)}\cos(\omega_c+2\omega_2-2\omega_1)t + K_{4(21-2)}\cos(2\omega_1-\omega_2)t + \\ + K_{4(22-1)}\cos(2\omega_2-\omega_1)t. \qquad \text{.....(67)}$$

where

$$K_{41} = \frac{-1}{(Y_1+\alpha)}\left[K_{2c}K_{1(c+1)}+K_{2(c+2-1)}K_{1(c+2)}+K_{0c}K_{3(c+1)}\right] \quad \text{.....(68)}$$

$$K_{42} = \frac{-1}{(Y_1+\alpha)}\left[K_{2c}K_{1(c+2)}+K_{1(c+1)}K_{2(c+1-2)}+K_{0c}K_{3(c+2)}\right] \quad \text{.....(69)}$$

$$K_{4c} = \frac{-1}{(Y_c+\alpha)}\left[K_{21}K_{1(c+1)}+K_{22}K_{1(c+2)}+K_{01}K_{3(c+1)}+K_{02}K_{3(c+2)}\right] \quad \text{.....(70)}$$

$$K_{4(c+2-1)} = \frac{-1}{(Y_{c+2-1}+\alpha)}\left[K_{01}K_{3(c+2)}+K_{02}K_{3(c+22-1)}\right] \quad \text{.....(71)}$$

$$K_{4(c+1-2)} = \frac{-1}{(Y_{c+1-2}+\alpha)}\left[K_{01}K_{3(c+21-2)}+K_{02}K_{3(c+1)}\right] \quad \text{.....(72)}$$

$$K_{4(c+21-22)} = \frac{-1}{(Y_{c+21-22}+\alpha)}\left(K_{02}K_{3(c+21-2)}\right) \quad \text{.....(73)}$$

$$K_{4(c+22-21)} = \frac{-1}{(Y_{c+22-21}+\alpha)}\left(K_{01}K_{3(c+22-1)}\right) \quad \text{.....(74)}$$

$$K_{4(21-2)} = \frac{-1}{(Y_{21-2}+\alpha)}\left[K_{0c}K_{3(c+21-2)}+K_{1(c+1)}K_{2(c+2-1)}\right] \quad \text{.....(75)}$$

$$K_{4(22-1)} = \frac{-1}{(Y_{22-1}+\alpha)}\left[K_{0c}K_{3(c+22-1)}+K_{1(c+2)}K_{2(c+1-2)}\right] \quad \text{.....(76)}$$

10. Appendix 2

Bias and Terminating Conditions for the Practical Diode

Considering the diagram of Fig. 5, two sets of axes are defined, the solid lines labelled i, v representing the characteristic used in the analysis and the dotted axes representing the characteristic as measured and defined as

$$I = \alpha' V + \beta' V^2 \quad \text{.....(77)}$$

Therefore α and β in eqn. (1) are related to α' and β' by the choice of bias condition. Now suppose that

$$V = v_0 + \sum_{p=1}^{\infty} v_p \cos \omega_p t \quad \text{.....(78)}$$

then

$$I = (\alpha'+\beta' v_0)v_0 + (\alpha'+2v_0\beta')\sum_{p=1}^{\infty} v_p \cos \omega_p t + \\ + \beta'\left\{\sum_{p=1}^{\infty} v_p \cos \omega_p t\right\}^2 \quad \text{.....(79)}$$

Thus for the a.c. components at the bias point (v_0, I_0)

$$i = (\alpha'+2v_0\beta')v + \beta' v^2 \quad \text{.....(80)}$$

i.e.

$$\alpha = \alpha'+2v_0\beta' \quad \text{.....(81)}$$

$$\beta = \beta' \quad \text{.....(82)}$$

Also from small-signal theory we know that optimum terminating conditions[3] for minimum conversion loss to the small signal in eqn. (14) are:

$$Y_{(c+1)\text{opt}} = Y_{1\text{ opt}} = \sqrt{g_0^2 - \tfrac{1}{4}g_1^2} \quad \text{.....(83)}$$

where g_0 is the d.c. component of the time-varying conductance $g(t)$ of the diode expressed as a Fourier Series and g_1 is the first harmonic component. Since

$$g(t) = \frac{\mathrm{d}i}{\mathrm{d}v} = \alpha + 2\beta v \quad \text{.....(84)}$$

$$g_0 = \alpha \quad \text{.....(85)}$$

$$g_1 = 2\beta v \quad \text{.....(86)}$$

For the ideal square-law diode ($\alpha' = 0$) and for small-signal operation

$$v = v_c = v_0 \quad \text{.....(87)}$$

$$g_0 = g_1 = \alpha \quad \text{.....(88)}$$

this gives

$$Y_{c+1} = Y_1 = \alpha\sqrt{3}/2 \quad \text{.....(89)}$$

From eqn. (79) for small-signal operation the admittance of the diode to the local-oscillator supply is seen to be

$$\frac{i_c'}{v_c} = (\alpha'+2v_0\beta') = \alpha \quad \text{.....(90)}$$

where i_c' is the current flowing in diode at frequency ω_c due to v_c across its terminals. Thus for optimum power transfer at the local-oscillator port regardless of bias condition

$$Y_c = \alpha \quad \text{.....(91)}$$

(It is seen from eqn. (40) that this condition holds to a good approximation for large values of interfering signal.)

Under large-signal conditions we know from eqn. (44)

$$v_0 = (1+\delta)v_c \qquad \text{.....(92)}$$

Thus from eqn. (80)

$$\beta = \frac{\alpha}{\frac{1}{Q}+2v_0} \qquad \text{.....(93}$$

where Q is the diode quality factor given by the ratio

$$Q = \beta/\alpha \qquad \text{.....(94)}$$

and since

$$i_c = 2\alpha v_c \qquad \text{.....(95)}$$

(to establish $i_c' = \alpha v_c$ in the diode)

$$\beta i_c = \frac{2\alpha^2 v_c}{\frac{1}{Q}+2(1+\delta)v_c} \qquad \text{.....(96)}$$

which for the ideal diode gives

$$\beta i_c = \frac{\alpha^2}{(1+\delta)} \qquad \text{.....(97)}$$

Manuscript first received by the Institution on 20th February 1968 and in final form on 10th January 1969. (Paper No. 1261/CC46.)

STANDARD FREQUENCY TRANSMISSIONS—May 1969

(Communication from the National Physical Laboratory)

May 1969	Deviation from nominal frequency in parts in 10^{10} (24-hour mean centred on 0300 UT)			Relative phase readings in microseconds N.P.L.—Station (Readings at 1500 UT)		May 1969	Deviation from nominal frequency in parts in 10^{10} (24-hour mean centred on 0300 UT)			Relative phase readings in microseconds N.P.L.—Station (Readings at 1500 UT)	
	GBR 16 kHz	MSF 60 kHz	Droitwich 200 kHz	*GBR 16 kHz	†MSF 60 kHz		GBR 16 kHz	MSF 60 kHz	Droitwich 200 kHz	*GBR 16 kHz	†MSF 60 kHz
1	— 300·0	— 0·1	— 0·1	548	470·3	17	— 300·2	0	0	562	480·8
2	— 300·0	— 0·1	0	548	471·1	18	— 300·1	+ 0·1	0	563	480·3
3	— 300·1	— 0·2	— 0·1	549	473·0	19	— 300·1	0	0	564	480·7
4	— 300·2	— 0·1	— 0·1	551	474·0	20	— 299·9	0	0	563	480·6
5	— 300·0	— 0·1	0	551	475·0	21	— 300·0	0	0	563	480·6
6	— 300·1	— 0·2	— 0·1	552	476·6	22	— 300·0	0	+ 0·1	563	480·9
7	— 300·2	— 0·1	0	554	478·1	23	— 300·0	+ 0·1	— 0·1	563	480·3
8	— 300·1	— 0·2	— 0·1	555	478·8	24	— 299·9	— 0·1	— 0·1	562	481·4
9	— 300·2	— 0·1	— 0·1	557	481·2	25	— 300·2	— 0·1	— 0·1	564	481·9
10	— 300·0	— 0·1	— 0·1	557	479·6	26	— 300·1	— 0·1	— 0·1	565	482·6
11	— 300·1	— 0·2	— 0·1	558	480·6	27	— 300·0	— 0·1	— 0·1	565	483·1
12	— 300·0	— 0·1	0	559	481·3	28	— 300·1	— 0·1	— 0·1	566	484·6
13	— 299·9	0	0	558	481·3	29	— 300·0	0	— 0·1	566	483·7
14	— 300·1	0	0	559	481·3	30	— 300·0	— 0·1	— 0·1	566	484·7
15	— 300·0	+ 0·1	0	559	479·8	31	— 300·2	+ 0·2	0	568	483·1
16	— 300·1	— 0·1	0	560	480·6						

All measurements in terms of H.P. Caesium Standard No. 334, which agrees with the N.P.L. Caesium Standard to 1 part in 10^{11}.

* Relative to UTC Scale; $(UTC_{NPL} - Station) = +500$ at 1500 UT 31st December 1968.

† Relative to AT Scale; $(AT_{NPL} - Station) = +468{\cdot}6$ at 1500 UT 31st December 1968.

Contributors to this Issue

L. A. Crapnell read mathematics (with physics) at Trinity College, Cambridge, obtaining his degree in 1960. He then joined the digital systems department of Ferranti Ltd., where he is primarily a logic designer. He has been concerned with the design and implementation of a number of real-time data handling systems for the Royal Navy. Mr. Crapnell is now group leader in charge of all central processor development work for the department.

E. I. Richmond graduated from Queen Mary College, London, in 1955. From 1955 to 1963 he was at the Atomic Weapons Research Establishment, Aldermaston, initially on secondment from the Royal Signals, and subsequently as a graduate apprentice; while from 1959-63 he was with a group designing instrumentation for the detection of underground nuclear explosions by seismic measurements, and for the analysis of the records. Mr. Richmond joined the Ferranti digital systems department at Bracknell, and has been concerned with various peripheral devices to be connected to real-time on-line computing systems. He is now leader of a section responsible for interfacing peripheral devices with the computer.

S. L. Hurst has been a lecturer in the School of Electrical Engineering at Bath University of Technology since 1961. He obtained his initial engineering training and education with British Thomson-Houston, Rugby, and graduated in electrical engineering in 1949. From 1950 to 1961 he was with the Westinghouse Brake & Signal Company at Chippenham, working on varied design and development projects, principally control and supervisory systems for the G.P.O., and home and overseas railways. In 1967 Mr. Hurst was awarded the M.Sc.(Eng.) degree by London University for research into ternary logic systems. He is currently engaged in investigations into the synthesis of single-gate and multi-gate threshold realizations of Boolean functions.

A. V. Lord became head of the signal origination group at the Research Department of the British Broadcasting Corporation in February of this year. He was previously head of electronics group (1966–67) and of physics group (1967–69). He has been with the B.B.C. since 1948, primarily being concerned with television matters: in 1961 he was appointed deputy head of the television group. Mr. Lord is the author of several papers and articles on television engineering and he has been active in professional group activities; he is a member of the Programme Committee for the International Broadcasting Convention.

Peter L. Mothersole (F. 1961, A.M. 1957, G. 1952) is now head of the consumer applications groups at the Mullard Central Application Laboratories, Mitcham. He studied electrical and radio engineering at Croydon Technical College and joined the Mullard Research Laboratories at Salfords in 1953, being concerned with the development of special measuring equipment. He became section leader of the video and signal processing group in the television laboratory in 1955. When the Central Application Laboratory was formed in 1964 he was given responsibility for the consumer activity. Mr. Mothersole is chairman of the Institution's Communications Group Committee, having previously served on the Television Group Committee since its formation in 1960, and as its chairman from 1964. He represents the I.E.R.E. on the management committee for the International Broadcasting Convention.

Peter McAuley Fenwick is a tutor in the Physics Department of the University of Auckland. He obtained his B.Sc. in 1964 and his M.Sc. with honours in 1966, both from Auckland, and he is currently doing work in computer logic design for a Ph.D. degree. He is also assisting with the development of a nuclear physics data acquisition system.

U.D.C. 621.317.32:621.391.822.2

A Radiometer for Measurement of the Noise Temperature of Low-noise Microwave Amplifiers

By

J. W. CARTER, B.Sc.†

H. N. DAGLISH, Ph.D., F.Inst.P., C.Eng., M.I.E.E.‡

AND

P. MOORE, B.Sc.‡

Summary: The noise temperature of a low-noise microwave amplifier can be measured by connecting the amplifier input in turn to thermal noise sources at different physical temperatures. The conventional technique is limited by the necessity for a very stable amplifier and system gain during the measurements. The paper describes a manually-balanced radiometer circuit designed to overcome this limitation, and intended to measure amplifier noise temperatures in the range below 500°K. Examples of experimental measurements on a maser and on uncooled and cooled parametric amplifiers are discussed.

List of Principal Symbols

C	coupling ratio of directional coupler C2. $\left(C = \frac{\text{incident power}}{\text{coupled power}} > 1\right)$
f	frequency
G, $G(f)$	gain of low-noise amplifier
G_R, $G_R(f)$	receiver gain, including where appropriate, that of the t.w.t.
h	Planck's constant
k	Boltzmann's constant
K_1, K_2	attenuation ratios (<1) of two states of diode modulator M1 ($K_1 < K_2$)
l_H, l_C	attenuation ratios (<1) when system is connected to 'hot' and 'cold' loads
R	attenuation ratio corresponding to change from 'hot' to 'cold' load
T_A	amplifier noise temperature
T_C	'cold' load noise temperature
T_C'	effective 'cold' load noise temperature at point of connection to low noise amplifier
T_H	'hot' load noise temperature
T_H'	effective 'hot' load noise temperature at point of connection to low-noise amplifier
T_L	noise temperature of the noise lamp
T_0	ambient temperature
T_T	noise temperature of travelling-wave tube

1. Introduction

Low-noise amplifiers, such as masers and cooled parametric amplifiers, are important components of satellite communication earth stations. In developing such amplifiers, it is essential to have an accurate measure of the amount of noise introduced into the signal by the amplifier. A useful parameter is the effective input noise temperature of the amplifier, by which the noise generated in the amplifier is expressed in terms of the thermal noise generated by a matched termination at the input of the amplifier. The effective input noise temperature (often referred to simply as the 'amplifier noise temperature') is defined as the temperature at which the input termination must be held to produce an output noise power from the amplifier, per unit bandwidth, double that which would occur if the input termination were cooled to absolute zero.[1]

In developing a new type of amplifier, it will clearly be necessary to make a large number of measurements of the amplifier noise temperature; if the amplifier in question is intended to have a large working bandwidth, it will be necessary to measure the noise temperature at a number of frequencies across the amplifier pass-band. The traditional methods of measuring the noise temperature are relatively slow, and because they depend upon the stability of the amplifier gain, the results are often of limited accuracy. As part of the development programme leading to new forms of cooled parametric amplifier, equipment was designed to enable rapid and accurate measurements to be made of the noise temperature of these amplifiers.

The general requirements were as follows:

Range of noise temperatures to be measured:	5–500°K
Range of amplifier gains to be accepted:	10–35 dB
Accuracy of noise temperature:	2%
Maximum time for a single measurement:	30 s
Frequency range (to cover the present civil satellite frequency allocation):	3·7–4·2 GHz

The equipment which was built to meet these requirements will be described, but in order to follow the

† Formerly with P.O. Research Department; now with G.E.C.-A.E.I. Electronics Ltd., Stanmore, Middlesex.

‡ P.O. Research Department, Dollis Hill, London, N.W.2.

description of its method of operation it is first desirable to consider the traditional method of measuring amplifier noise temperatures.

2. The 'Y' Method

The noise temperature T_A of a microwave amplifier is usually measured by connecting the amplifier in turn to matched thermal noise sources at different temperatures T_H and T_C (where T_H refers to the 'hot' and T_C to the 'cold' termination). This is often called the 'Y' method—see Fig. 1. The noise power at the

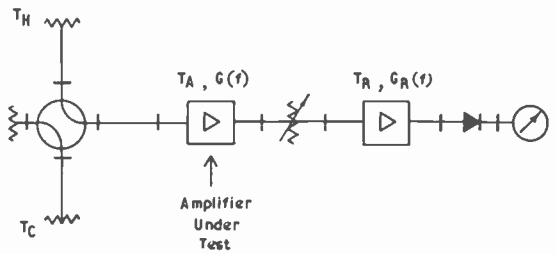


Fig. 1. The 'Y' method of noise temperature measurement.

attenuator output, in an elementary bandwidth df, is equal[2] to

$$dP_H = (T_H + T_A)kl_H G(f)\,df + (1-l_H)kT_0\,df \quad \text{.....(1)}$$

when connected to the 'hot' load, and to

$$dP_C = (T_C + T_A)kl_C G(f)\,df + (1-l_C)kT_0\,df \quad \text{.....(2)}$$

when connected to the 'cold' load. l_C and l_H are the loss ratios (< 1) of the attenuator, which is at ambient temperature T_0. k is Boltzmann's constant. The noise power is detected with a receiver and power meter. Suppose that the receiver gain function is $G_R(f)$ and its noise temperature is T_R. The powers recorded at the power meter input are:

$$P_H = \int_0^\infty [(T_H + T_A)kl_H G(f) + (1-l_H)kT_0 + kT_R]G_R(f)\,df \quad \text{.....(3)}$$

and

$$P_C = \int_0^\infty [(T_C + T_A)kl_C G(f) + (1-l_C)kT_0 + kT_R]G_R(f)\,df \quad \text{.....(4)}$$

Now suppose that the relationship between l_H and l_C is such that $P_C = P_H$.

We now write

$$\frac{l_C}{l_H} = R,$$

where $R > 1$, and we shall suppose that the receiver includes a narrow filter so that significant contributions to the total noise power are only made within the frequency range $f_0 \pm \Delta f$ and the values of T_H, T_C and R can be considered as being independent of frequency.

If T_A is a slowly-varying function of frequency between $f_0 - \Delta f$ and $f_0 + \Delta f$,

$$T_A = \frac{T_H - RT_C}{R-1} + \frac{T_0 \int_0^\infty G_R(f)\,df}{\int_0^\infty G(f)G_R(f)\,df} \quad \text{.....(5)}$$

If the amplifier bandwidth is very much greater than the filter bandwidth, so that

$$\int_0^\infty G(f)G_R(f)\,df \simeq G(f_0)\int G_R(f)\,df \quad \text{.....(6)}$$

then eqn. (5) can be simplified to

$$T_A = \frac{T_H - RT_C}{R-1} + \frac{T_0}{G(f_0)} \quad \text{.....(7)}$$

The amplifier noise temperature is given in terms of the physical temperatures of the two loads and of the attenuator, the attenuation ratio, and the amplifier gain. The noise temperature of the receiver does not have to be known. The last term of eqn. (7) is usually small, so the amplifier gain does not have to be known accurately; it is important, however, that the gain of the amplifier and of the receiver remain constant during the measurement, as a gain change of n dB will lead to an error in R of n dB. The gain in the amplifier and receiver may exceed 100 dB, and this must remain constant to better than $\pm 0{\cdot}02$ dB if an accuracy of ± 1 degK is required. Although the short-term gain stability of a carefully stabilized receiver may approach this value, the gain of a parametric amplifier depends so critically on the pump power that a gain change of this order during a measurement is almost inevitable. It is equally important that the impedances of the two loads at the amplifier connection point are identical as the gain of a parametric amplifier using a circulator with finite isolation varies rapidly with the source impedance. A refined method of noise temperature measurement is obviously required which is less sensitive to amplifier gain variations. The radiometer which is described in this paper represents a practical solution to this problem.

3. A Gain-independent Radiometer

The circuit of the apparatus is shown in Fig. 2 and the associated electronics in Fig. 3. As well as the basic noise temperature measurement, provision has been made for the gain of the amplifier under test to be measured and for the gain/frequency response to be displayed on an oscilloscope.

The essential feature of this radiometer is the use of an internal source of noise which is used, in effect, as

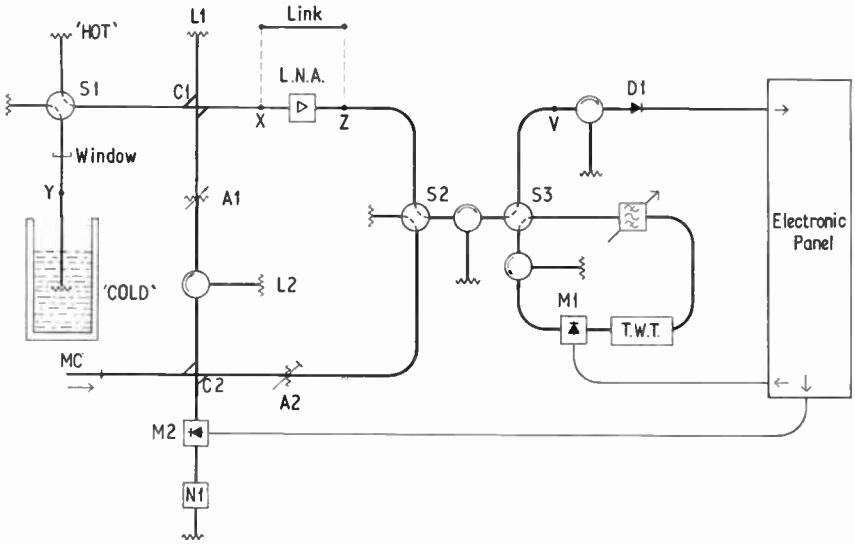


Fig. 2. Schematic circuit of the radiometer.

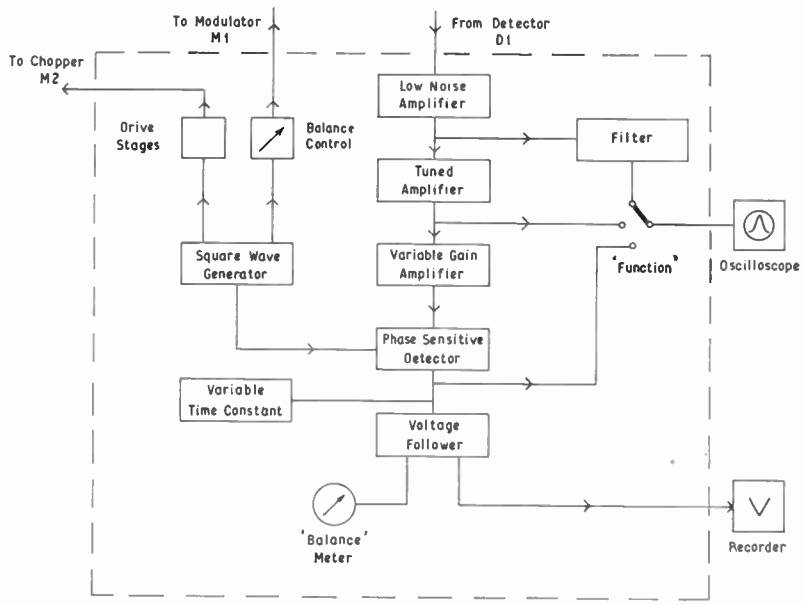


Fig. 3. The electronic panel of the radiometer.

a noise transfer standard, with which the total noise power from the thermal sources and the amplifier itself may be compared. The effective noise temperature of this internal source does not have to be known accurately provided that it remains constant during the measurement. The use of this internal noise source enables all measurements to be made under balanced conditions so that small changes in amplifier gain do not affect the measurements.

Noise from the internal source N1 can be added to the system through a precision attenuator A1 and a 20-dB directional coupler C1. This noise source is a continuously-operating neon-filled discharge lamp, in a waveguide mount. The noise output is chopped at an audio frequency (about 400 Hz) by a p-i-n-diode switch M2.

The amplifier input can be connected, as before, to either a 'hot' or a 'cold' matched load, each consisting

of a matched waveguide termination: the 'cold' load is normally immersed in liquid nitrogen and the 'hot' load maintained at ambient temperature.

The 'cold' load consists of a polyiron pyramid, as shown in Fig. 4, mounted in a length of copper waveguide. The more usual thin-walled stainless steel waveguide[3] was not used, in order to reduce the uncertainty about the actual value (and stability) of the attenuation between the effective position of the load and the reference plane. By using copper waveguide this uncertainty was reduced, at the expense of a large increase in the rate of consumption of liquid nitrogen; the increase is not serious for laboratory

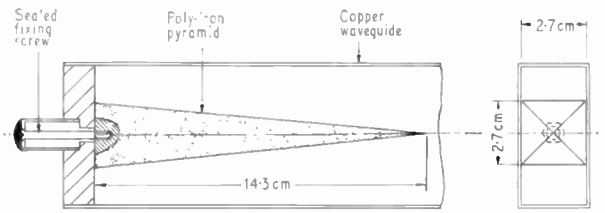


Fig. 4. Construction of the 'cold' load.

experiments. The waveguide aperture is sealed by a low-loss 'melinex' window, which is heated to prevent condensation. The waveguide is immersed in liquid nitrogen contained in a stainless steel Dewar. The selected load is connected to the amplifier input by a 4-port waveguide switch. The v.s.w.r. looking into each load from the amplifier connexion port is plotted as a function of frequency in Fig. 5.

The loss factor S between the 'cold' load connection point Y (see Fig. 2) and the amplifier connection point X was measured as a function of frequency. The calculated effective temperature T_C' of the 'cold' load is shown together with S in Fig. 6. T_C' is equal to 82·5° ±0·5 degK from 3·7 GHz to 4·2 GHz.

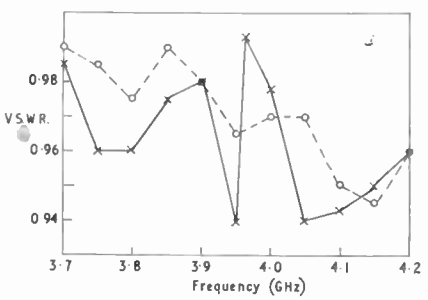


Fig. 5. V.s.w.r.s of the 'hot' and 'cold' loads. (– – ○ – – 'hot' load at ambient temperature. — × — 'cold' load at 77°K.)

Impedance matching in the microwave circuit is ensured by the inclusion of several ferrite circulators used as isolators.

The amplifier under test is followed by a matched modulator M1, identical in construction to the switch M2. The power loss coefficient of this modulator can be varied from K_1 to K_2 (where $K_1 < K_2 < 1$) in phase with the chopping action of switch M2. The attenuation of the high-loss state K_1 can be varied by an uncalibrated control. The output from M1 is amplified by a travelling-wave tube and, after passing through an yttrium iron garnet filter having a fixed bandwidth but variable centre frequency, is passed to a coaxial backward-diode detector D1. The audio output from this detector is amplified in a low-noise transistor amplifier and detected by a synchronous detector driven from the same square-wave oscillator as the p-i-n-diode switch M2 and modulator M1.

The electronic circuit, shown in the block diagram of Fig. 3, uses entirely solid-state components and consists of a low-noise preamplifier, a tuned amplifier, a stepped variable gain amplifier, a synchronous detector with stepped variable time-constant (0·1, 1, 2, 5, 10 s) and a voltage follower which drives the balance indicating meter and an external recorder if required. The variable gain amplifier is included so that the system can work with constant sensitivity when testing amplifiers with gains ranging from 10 dB to 30 dB.

The square-wave oscillator provides a signal whose mark/space ratio is equal to unity within about ±5%, in order to avoid significant loss in sensitivity.

In order to measure the amplifier noise temperature, the switch S3 is arranged to connect the t.w.t. and variable filter to the output of the amplifier under test and the filter is adjusted to select the required section of the frequency spectrum. This adjustment can be

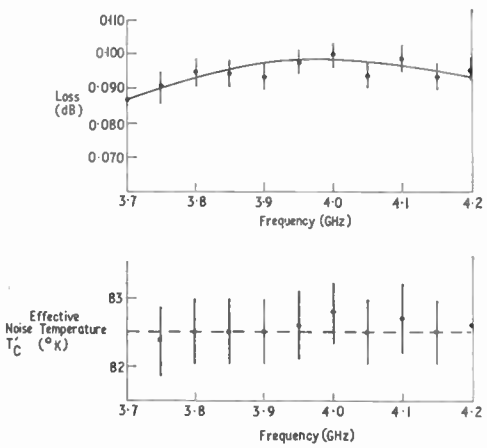


Fig. 6. Variation of interconnexion loss and effective 'cold' load noise temperature with frequency.

made from a directly-calibrated multi-turn control conveniently situated on the front panel of the equipment.

The system is balanced initially with the amplifier input connected to the 'hot' load, and the attenuator A1 set to 0 dB. The balance control is used to vary the depth of modulation of M1 until, when M2 switches the noise source 'on', the attenuation of the high-loss state of M1 is large enough for the total noise powers entering the receiver during the two half-cycles of the switching frequency to be equal. To a first order, this condition is independent of the gain of the low-noise amplifier, provided that any gain changes occur slowly compared with the switching rate.

The waveguide switch is then moved to connect the amplifier input to the 'cold' load; the balance condition is then restored by decreasing the injected noise power from the internal source by increasing the attenuation of A1 from zero to R. The analysis in the Appendix shows that the amplifier noise temperature is given by

$$T_A = \frac{T'_H - R.T'_C}{R-1} + \frac{T_0}{G(f_0)} \qquad \text{.....(8)}$$

Although this equation is apparently the same as eqn. (7), there is now no implicit assumption that the gain remains constant during the measurement;

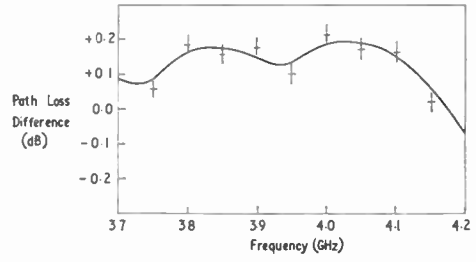


Fig. 7. Differences in attenuation between amplifier and reference paths.

changes in gain while the amplifier input is switched between the 'hot' and 'cold' loads will only affect the second term in eqn. (8), which is usually quite small.

It is obviously desirable that it should be possible to measure and display the gain-frequency response of the amplifier immediately before and after noise measurements. This can be done by switching off the square-wave source, and then injecting a microwave signal from a swept oscillator into the mode changer MC. Most of the power enters the preset attenuator A2 and reaches the detector via a cable link and switch S2; this is called the 'reference' path. A fraction of the power passes to the coupled arm of C2 and

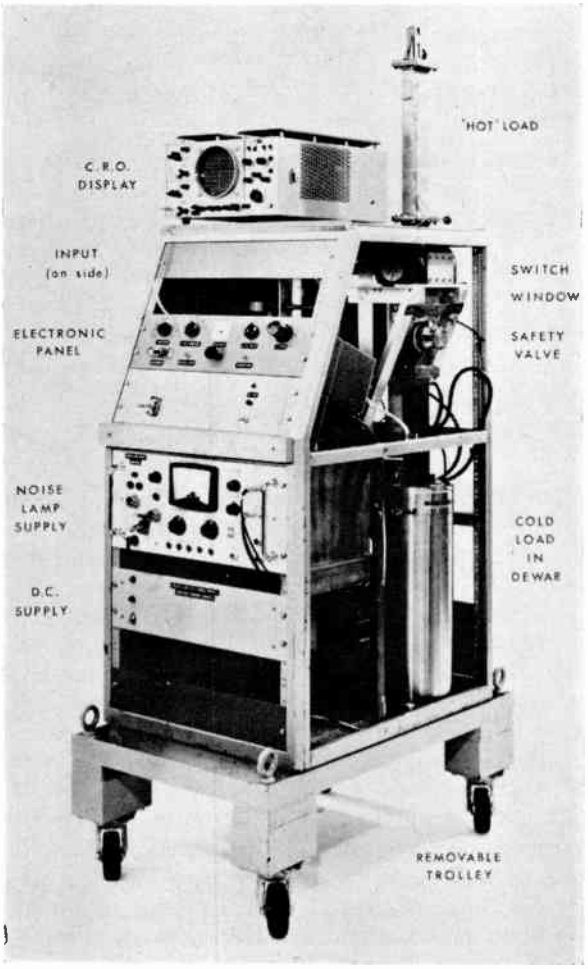


Fig. 8. The radiometer trolley.

reaches the detector via the attenuator A1, the amplifier under test, and S2; this is called the 'amplifier' path. The preset attenuator A2 is adjusted so that, when a cable link is substituted between points X and Z the loss between MC and S2 via the reference path is the same as the loss via the amplifier path when the attenuation of A1 is set to its minimum value. If the link is replaced by the amplifier to be tested and the path losses equalized by adjusting A1, the attenuation introduced is numerically equal to the amplifier gain. The t.w.t. is switched out of circuit by S3 during this measurement. In addition to measuring the amplifier gain, the gain/frequency characteristic can be displayed on an associated oscilloscope.

Components with flat loss/frequency characteristics have been used throughout; the difference between the reference and amplifier channel losses is plotted in Fig. 7 from 3·7 GHz to 4·2 GHz.

The waveguide components and electronic equipment are mounted in a self-contained trolley as shown in Fig. 8. The only other item of equipment required is a swept oscillator which covers the operating band of the amplifier under test.

When connecting the input of the low-noise amplifier under test to the radiometer, the connexion must be as short as possible to avoid adding extra attenuation. It is often useful to lift the whole radiometer trolley on a hydraulic platform until the radiometer and the amplifier are close enough together to be connected either directly or by a short length of waveguide whose loss can subsequently be measured accurately.

4. System Accuracy

The accuracy of the results obtained with the radiometer are limited by both systematic and random errors. The accuracy with which the balance condition can be set is determined by the degree of off-balance caused by changes in the power injected by the noise source, and the amplitude of the meter fluctuations caused by the 'randomness' of the noise. Analysis shows that the system is most sensitive if a noise injection source of infinite power is used; in practice, the coupling factor of the directional coupler C1 must be chosen to afford a compromise between the higher sensitivity resulting from a high level of injected noise and the reduced accuracy resulting from an increase in the effective noise temperature of the 'cold' load (caused by noise from the termination L2) when the noise source is off. A 20-dB directional coupler was used in the system described here (reducing the theoretical sensitivity by a factor of seven, and increasing the effective 'cold' load temperature by 3degK); experience has suggested, however, that the coupling ratio could with advantage be changed to 17 dB.

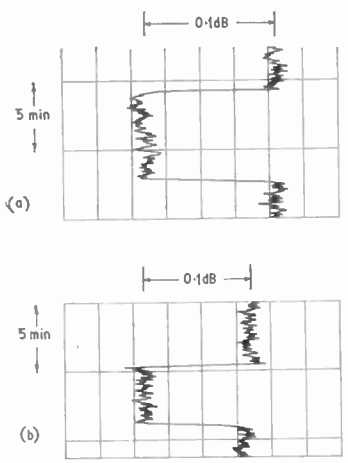


Fig. 9. Balance stability, with (a) 20 dB, and (b) 10 dB gain parametric amplifiers. (Time-constant = 10 s, T_A = 150°K)

The random meter fluctuations are minimized by using a large r.f. noise bandwidth (about 30 MHz) and a long post-detector time-constant (about 5 s.)

The most important systematic errors are caused by uncertainties in the calibration of the precision attenuator A1, and in the noise temperature of the 'cold' load. The attenuator was calibrated from 0 to 6 dB across the 3·7 to 4·2 GHz frequency band, against a 30 MHz piston attenuator. The effective noise temperature of the 'cold' load at the point of connexion with the amplifier is greater than the physical temperature of the refrigerant by an amount dependent on the loss of the interconnecting waveguide. This loss was carefully measured at a number of frequencies within the working range of the equipment.

A somewhat greater accuracy could have been achieved by using a helium-cooled 'cold' load,[4] although this would not have been so convenient for the routine measurements for which the equipment was required.

For most purposes, the full analysis for the error ΔT_A in the noise temperature measurements can be simplified to the approximate equation:

$$\Delta T_A = \pm\left(1 + \frac{1{\cdot}8T_A}{100}\right)\text{degK} \qquad \text{.....(9)}$$

provided that the amplifier noise temperature is less than about 500°K.

5. Results

The radiometer has been used to measure the noise temperatures of masers and parametric amplifiers in the laboratory and in the field. One of the masers used to receive signals from the *Early Bird* satellite was measured *in situ* at Goonhilly Radio Station. A noise temperature of $8° \pm 1$ degK was obtained, the value being independent of the maser gain, which was varied between 20 and 33 dB by changing the level of the pump power.

Cooled and uncooled parametric amplifiers with noise temperatures varying from 15 to 300°K have been measured: the stability obtained with typical single-stage amplifiers is shown in Fig. 9. Some of the results of the measurements on low-noise broad-band helium-cooled amplifiers have been described by Watson.[5]

The radiometer has been used in conjunction with a simple helium-cooled parametric amplifier[6] to measure the effective noise temperatures of several separate experimental cold load assemblies. The amplifier consisted of a conventional balanced-diode

parametric amplifier, using the self-resonance of the diode pair as the idler circuit: the pump frequency was 12·5 GHz. The amplifier could be tuned from 3·7 GHz to 4·2 GHz. The instantaneous bandwidth was 35 MHz at 17 dB gain, and the measured input noise temperature was 23°±1 degK. The experiment was carried out by setting up the equipment in the normal way and obtaining a value for the ratio R, and

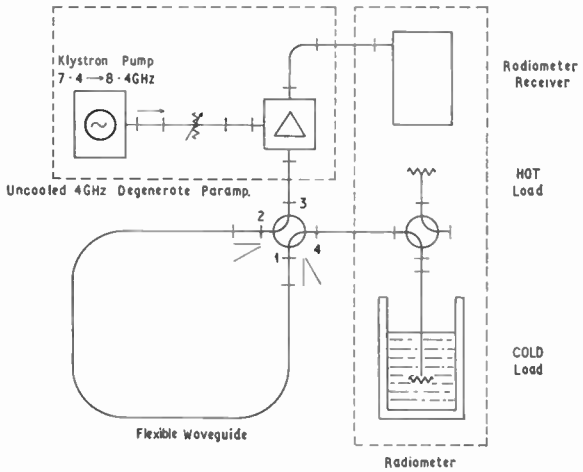


Fig. 10. Measurement of insertion loss of flexible waveguide.

then removing the normal 'cold' load and substituting the 'unknown' load and remeasuring R. For example, the effective noise temperature of one experimental 'cold' load was found to be as follows:

3814 MHz	81·9°K	±1·7 degK
3900	78·9	
4170	81·2	

When a lossy component is placed in front of an amplifier, the input noise temperature of the combination is increased. The insertion loss of the component can be obtained from measurements of the noise temperatures of the amplifier alone and of the combined system. Some experiments were conducted in which the insertion loss of a 10·5 m (35 ft) length of elliptical flexible waveguide was obtained from measurements of the degradation caused to the effective noise temperature of an uncooled degenerate parametric amplifier. The flexible waveguide was connected to ports 1 and 2 of a four-port two-channel waveguide switch as shown in Fig. 10; port 4 was connected to the radiometer and port 3 to the parametric amplifier. The amplifier was operated at 20 dB gain, when it had an instantaneous bandwidth of 40 MHz, a tunable bandwidth of over 500 MHz and a noise temperature of 130°K. The amplifier noise

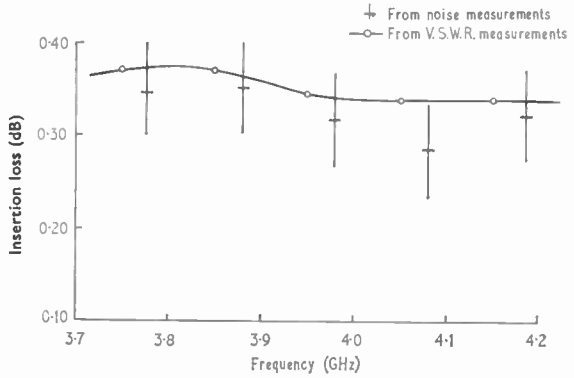


Fig. 11. Insertion loss of flexible waveguide.

temperature T_{A1} was measured with ports 3 and 4 connected directly, and then with the amplifier connected to the noise sources within the radiometer trolley via the flexible waveguide, giving a higher noise temperature T_{A2}. The waveguide loss L was then calculated from:

$$L = 10 \log_{10} \left(\frac{T_{A2} + T_0}{T_{A1} + T_0} \right) \qquad \text{.....(10)}$$

The results are compared in Fig. 11 with values derived from conventional standing-wave measurements.[7] The comparison shows that the two sets of values agree within the limits of the expected experimental error, although there is some evidence of a systematic error of about 0·015 dB. The error in the value of attenuation obtained from the noise measurements is of the order of ±0·05 dB which is about four times greater than that expected for the conventional method.

6. Conclusions

A prototype radiometer has been designed and constructed which meets all the requirements described in the introduction. It has enabled a large number of measurements of noise temperature to be made in a short period of time, and the noise performance of a number of experimental parametric amplifiers has thus been examined in much more detail and with greater accuracy than would otherwise have been practicable.

7. Acknowledgments

Acknowledgment is made to the Senior Director of Development of the Post Office for permission to make use of the information contained in this paper.

The authors also wish to acknowledge the assistance of Mr. J. M. McPherson in the design and construction of the mechanical and cryogenic parts of

the equipment, and of Mr. D. Geden and Mr. N. E. Cartwright who carried out some of the measurements described in this paper.

A shorter version of this paper was presented at the Conference on 'Physical Aspects of Noise in Electronic Devices', held in Nottingham in September 1968 under the sponsorship of the Institute of Physics and The Physical Society.[8]

8. References

1. 'IRE Standard on Methods of Measuring Noise in Linear Two-Ports', *Proc. Inst. Radio Engrs*, **48**, p. 61, 1960.
2. Bedrosian, E., 'On the noise temperature of coupling networks', *I.R.E. Trans. on Microwave Theory and Techniques*, **MTT–8**, p. 463, 1960.
3. Daglish, H. N. and Carter, J. W., 'Cooled terminations for use as 4 Gc/s microwave standard noise sources', *Proc. Instn Elect. Engrs*, **112**, p. 705, 1965.
4. Eisele, K. M., 'Refrigerated microwave noise sources', *I.E.E.E. Trans. on Instrumentation*, IM–13, p. 336, December 1964.
5. Watson, P. A., 'Noise temperature of liquid helium cooled parametric amplifiers', in 'Physical Aspects of Noise in Electronic Devices', (Peter Peregrinus, Stevenage, 1969).
6. Daglish, H. N. *et al.*, 'Low-noise Microwave Amplifiers', p. 76. (Cambridge University Press, 1968).
7. Vogelman, J. H., 'Precision measurement of waveguide attenuation', *Electronics*, **26**, p. 196, December 1953.
8. Carter, J. W., Daglish, H. N. and Moore, P., 'A radiometer for measuring the noise temperature of low noise microwave amplifiers', in 'Physical Aspects of Noise in Electronic Devices', p. 44 (Peter Peregrinus, Stevenage, 1969).

9. Appendix

In order to calculate the relationship between the ratio R and the amplifier noise temperature we shall assume that T_H, T_C', T_L, T_T, C, K_1 and K_2 do not vary significantly with frequency within the pass-band of the receiver. T_H and T_C' are respectively the effective noise temperatures of the 'hot' and 'cold' loads, referred to the input of the low-noise amplifier under test, with the noise lamp 'off'. These effective noise temperatures will include a term T_0/C, due to thermal noise introduced from the side arm of the directional coupler. When the lamp is 'on', this term is replaced by the appropriate term including T_L. To avoid redefining T_H' and T_C' it is necessary to allow for this by inserting the $-T_0/C$ term which will be seen in the equation below.

l_H and l_C are now the attenuation ratios(<1) between the noise lamp and the low-noise amplifier (including the attenuator A1, but not the coupling factor C) when the system is connected to the 'hot' and 'cold' loads respectively. These factors will be assumed to be constant for frequencies within the pass-band of the receiver.

$G_R(f)$ represents the gain/frequency characteristic of the travelling-wave tube plus the narrow band-pass filter.

The noise power at the plane V when the noise source is 'on' and the 'hot' load is connected to the amplifier is given by:

$$P_1 = k\int_0^\infty \left(T_H' + T_A + \frac{T_L l_H + (1-l_H)T_0}{C} - \frac{T_0}{C}\right) \times \\ \times G(f)K_1 G_R(f)\,df + \\ + k\int_0^\infty [(1-K_1)T_0 + T_T]G_R(f)\,df \quad \text{.....(11)}$$

When the noise source is 'off'

$$P_2 = k\int_0^\infty (T_H' + T_A)G(f)K_2 G_R(f)\,df + \\ + k\int_0^\infty [(1-K_2)T_0 + T_T]G_R(f)\,df \quad \text{.....(12)}$$

If K_1 is adjusted so that $P_1 = P_2$, the noise power will not be amplitude modulated at the switching frequency, and there will be a null at the synchronous detector. Thus

$$\int_0^\infty \left(T_H' + T_A + \left(\frac{T_L - T_0}{C}\right)l_H\right)G(f)K_1 G_R(f)\,df - \\ - \int_0^\infty K_1 T_0 G_R(f)\,df \\ = \int_0^\infty (T_H' + T_A)G(\ )K_2 G_R(f)\,df - \\ - \int_0^\infty K_2 T_0 G_R(f)\,df \quad \text{.....(13)}$$

Therefore

$$(T_H' + T_A)(K_1 - K_2) - (K_1 - K_2)T_0 \frac{\int_0^\infty G_R(f)\,df}{\int_0^\infty G(f)G_R(f)\,df} \\ = -l_H \frac{T_L - T_0}{C} K_1 \quad \text{.....(14)}$$

If

$$\int_0^\infty G(f)G_R(f)\,df \simeq G(f_0)\int_0^\infty G_R(f)\,df,$$

then

$$T_H' + T_A - \frac{T_0}{G(f_0)} = -l_H\left(\frac{T_L - T_0}{C}\right)\left(\frac{K_1}{K_1 - K_2}\right) \quad \text{.....(15)}$$

The gain of the amplifier only appears in the small term

$$\frac{T_0}{G(f_0)},$$

so that, if the amplifier gain does not change significantly in one period of the switching frequency (which may be several hundred hertz), the balance condition is largely independent of the amplifier gain. The synchronous detector is used only to detect a null, so that changes in the receiver gain are not significant.

If the amplifier input is connected to the cold load, the system can be re-balanced by increasing the attenuation of A1, l_H then becoming l_C; no other adjustments are made. We then have

$$T'_C + T_A - \frac{T_0}{G(f_0)} = -l_C\left(\frac{T_L - T_0}{C}\right)\left(\frac{K_1}{K_1 - K_2}\right) \quad \text{.....(16)}$$

Dividing (15) by (16), and rearranging,

$$T_A = \frac{T'_H - RT'_C}{R-1} + \frac{T_0}{G(f_0)} \quad \text{.....(17)}$$

where $R = l_H/l_C$. Equation (17) is the same as eqn. (8); the noise source temperature, the modulator transmission coefficients K_1 and K_2, and the coupling ratio C have all disappeared (although C still appears implicitly in T'_H and T'_C).

Manuscript first received by the Institution on 20th November 1968 and in final form on 14th February 1969. (Paper No. 1262/IC.5).

The Authors

Dr. H. N. Daglish read physics at Sheffield University, obtaining a first class honours B.Sc. degree in 1951 and a Ph.D. in 1954. He joined the Post Office Research Station in 1954, working on problems connected with the long-life valves used in the repeaters of transoceanic telephone cables. From 1961 to 1967 Dr. Daglish was leader of the low-noise amplifier group in the satellite communication division, and he is currently working on the use of optical frequencies for telecommunications. He is co-author of a recent book on low-noise microwave amplifiers and of a previous paper in *The Radio and Electronic Engineer*.

J. W. Carter received the B.Sc. degree in physics from the University of Bristol in 1962. After graduating, he joined the Post Office Research Station at Dollis Hill, where he assisted with the development of maser and parametric amplifiers and microwave noise measuring equipment. Since May 1967 he has worked in a satellite transponder development group at G.E.C.-A.E.I. Electronics Ltd., Stanmore.

P. Moore joined the Post Office in 1953. Since graduating from London University in 1963 he has been involved with the design of electronic equipment associated with maser and parametric amplifiers used in connection with satellite communications.

U.D.C.621.373.54

A Pulse Generator using R.T.L. Integrated Circuits

By

P. M. FENWICK,
M.Sc.†

Presented at the Second New Zealand Electronics Conference (Nelcon II), organized by the New Zealand Section of the I.E.R.E. and the New Zealand Electronics Institute and held in Auckland in August 1968.

Summary: A three-channel pulse generator is described which uses r.t.l. integrated circuits as the sole active devices. Each output pulse has a width variable from 100 ns to 1 μs and may be delayed by up to 2 μs with respect to either a fixed synchronization pulse or an external trigger. The circuit as a whole demonstrates the flexibility of some digital integrated circuits in non-digital applications.

1. Introduction

The pulse generator described in this paper was built to fill a specific need which arose during the development of a new type of binary multiplier.‡ It was initially designed to provide control pulses to test the operation of the multiplier logic boards over a wide range of signal timings. Since then the pulse generator has been used to provide some of the timing pulses for the completed multiplier and also to examine the operation of many miscellaneous digital circuits. It has proved to be a very versatile instrument, especially when its simplicity and low cost are considered. The design requirements of the pulse generator were:

(i) The outputs must be compatible with the integrated circuits used in the multiplier. This is readily achieved by using similar integrated circuits in the pulse generator output stages.

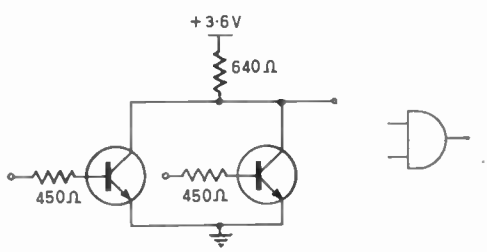
(ii) Several output channels should be provided. Each channel output pulse should be independently variable in width and also in delay with respect to a common synchronization pulse.

(iii) The width and delay controls must operate smoothly and independently, but need not be calibrated. The generator is always used in conjunction with a multi-channel oscilloscope and this can readily and accurately provide the timing information.

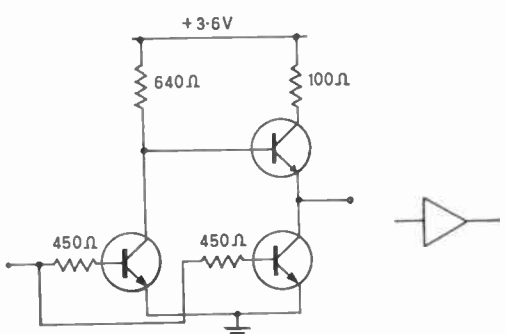
Initially it was planned to use discrete components in most of the pulse generator, but it very quickly became apparent that digital integrated circuits could satisfactorily perform all the necessary active functions if some liberties were taken with the interpretation of their circuits and characteristics. In the final instrument the width of each of the three output pulses may be varied from 100 ns to 1 μs and the pulses may be delayed by up to 2 μs with respect to either a fixed synchronization pulse or an external trigger.

2. General Operation

The main timing oscillator is a linear ramp generator of a rather unusual form. It may be externally triggered or can free-run at about 200 kHz and delivers a linear ramp to each channel delay circuit.



(a) ½ μL914 Gate



(b) μL900 Buffer

Fig. 1. Equivalent circuits of the integrated circuits used in the pulse generator.

† Physics Department, University of Auckland, Auckland, New Zealand.

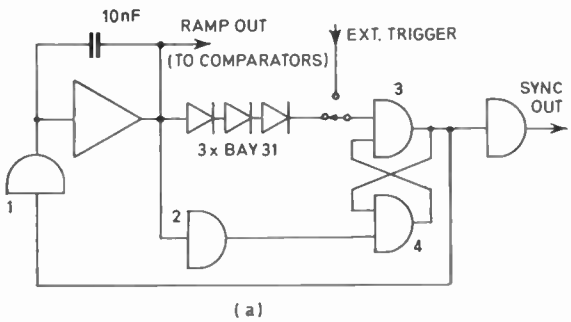
‡ Fenwick, P. M., 'Binary multiplication with overlapped addition cycles', *I.E.E.E. Trans. on Computers*, C–18, No. 1, pp. 71–4, January 1969.

The delay circuit 'fires' at a level on the ramp determined by the appropriate delay control and in turn triggers its respective pulse generator. Each of these is a ramp generator similar to the timing oscillator but without provision for free-running operation. The run-down time, corresponding to the output pulse duration, is variable to give the pulse width control. Only three pulse channels are provided in the original unit but more outputs can easily be added if desired.

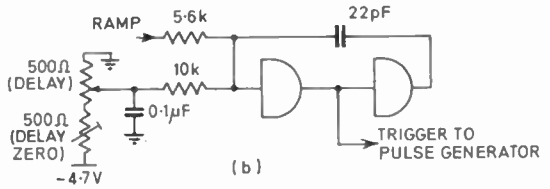
Resistor-transistor logic integrated circuits are used throughout and details of the relevant circuits are given in Fig. 1 as an aid to understanding the detailed operation.

3. Detailed Operation

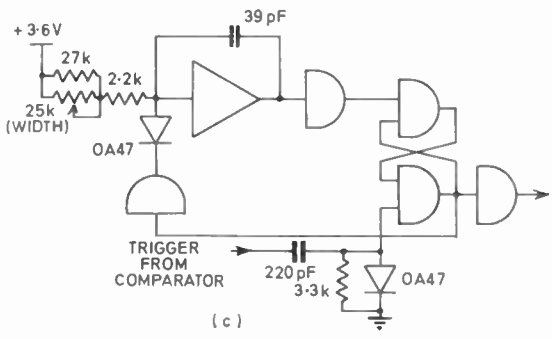
The timing oscillator uses a μL900 buffer with capacitive feedback to give the rundown ramp. Although the gain of the μL900 is very low by operational amplifier standards (about 150), it is nevertheless quite adequate for the purpose. The 640 Ω load resistor in gate 1 connects between the amplifier input and the positive supply line and is used as the timing resistor. Gates 3 and 4 form a reset flip-flop which controls the resetting of the ramp generator.



(a) Main timing oscillator



(b) Comparator/delay circuit



(c) Output pulse generator

Fig. 2.

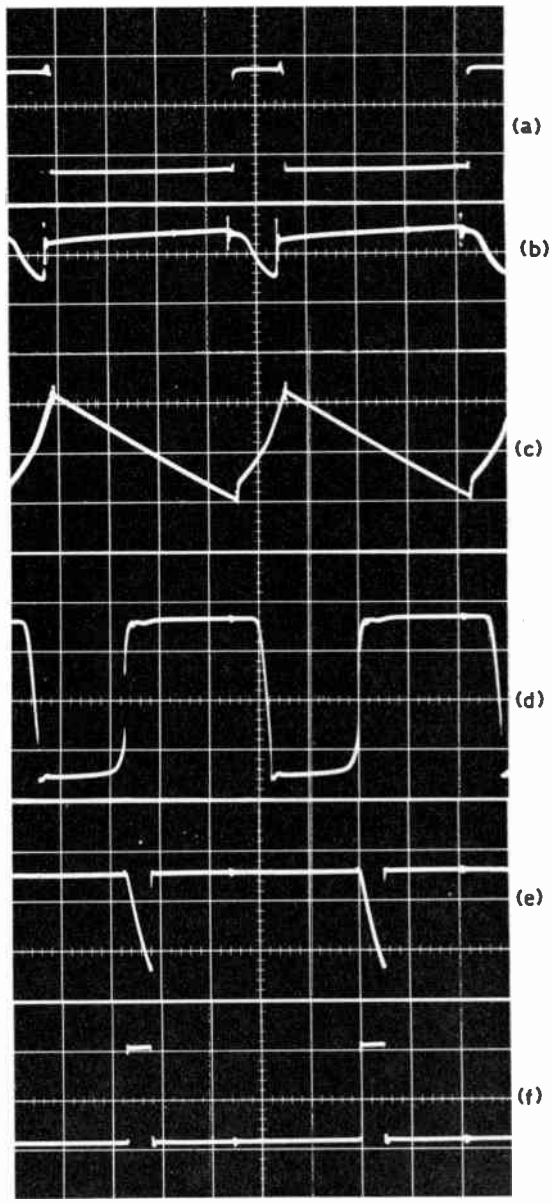


Fig. 3. Waveforms.

(scale 1V/div vert., 1 μs/div hor.)

(a) Synchronization pulse
(b) Timing oscillator—amplifier input (0·5V/div vert.)
(c) Timing oscillator—ramp out
(d) Comparator output
(e) Output pulse generator ramp
(f) Output pulse

Because the timing resistor returns to a positive potential the output ramp is negative-going. Until the ramp reaches the switching threshold of gate 2 (about 0·8 V), the output of gate 2 is low. When the ramp does fall below 0·8 V the output potential of gate 2 rises, setting the reset flip-flop. The input to gate 1 goes positive, turning on its transistor, thereby effectively supplying a large negative current to the amplifier input. As a result the integrator output potential rises until at about 3 V the potential drop across the three series diodes is overcome and the input to gate 3 exceeds the threshold of 0·8 V, re-setting the reset flip-flop. The input clamp (gate 1) is turned off and a new rundown commences.

Following the timing oscillator are three identical comparator/pulse generator circuits, one for each channel. Each comparator consists of a gate fed from a voltage divider between the ramp and a variable negative potential, with the components chosen so that the comparator output change can be set to correspond to any position on the ramp. The output change from one gate is not very abrupt, so the otherwise unused gate in the same package is used to provide some positive feedback to speed the transition.

The variable-width pulse generators for each channel are basically similar to the main timing oscillator, but can operate only in the triggered mode. A differentiating circuit, with a d.c. restoring diode, provides short 'pips' to trigger the pulse generator. The main circuit modification is in the timing circuitry at the integrator input. The timing resistor is variable to control the pulse width and a gold-bonded diode (for low voltage drop) isolates the integrator from the reset circuitry during the rundown interval. The output pulse is taken from the reset flip-flop through an isolating gate.

If desired, the timing oscillator can be triggered externally by a 1–3 V pulse, at least 30 ns long. In most work with the pulse generator it has been found convenient to have the monitoring oscilloscope free-running and to trigger the pulse generator from the oscilloscope 'gate-output' signal. The repetition rate is then controlled directly by the oscilloscope time-base controls and only one pulse sequence occurs for each time-base sweep. This is very convenient if a lengthy operation is initiated by the pulse generator. In addition the 'synchronization' signal from the pulse generator is available as an extra, fixed-length, pulse.

4. Construction

Each channel is individually decoupled by a 22 Ω resistor in its positive supply line, bypassed by a 0·22 μF capacitor. It was also found necessary to place shields between the different comparators to avoid interaction between the channels. The original pulse generator has all the components except the controls on an experimental wiring board about 10 cm by 12·5 cm (4 in by 5 in). The controls and output sockets fit on a panel 25 cm by 10 cm, the size being determined mainly by the need to avoid crowding.

5. Acknowledgments

The author thanks the New Zealand Universities Research Grants Committee for financial support and the Department of Physics, University of Auckland for the provision of research facilities.

Manuscript first received by the Institution on 13th November 1968 and in final form on 7th February 1969. (Short contribution No. 117/CC47).

U.D.C. 681.322

On the Generation of a Standard Computer/Peripheral Interface

By

L. A. CRAPNELL, M.A.†

AND

E. I. RICHMOND,

B.Sc.(Eng.)†

Presented at the Conference on 'Electronic Switching and Logic Circuit Design', organized by the College of Technology, Letchworth, with the support of the I.E.R.E. and held at Letchworth on 24th October 1968.

Summary: The buffer unit forming the interface comprises two distinct sets of logic circuits: computer interrupt equipment and peripheral control unit. Information transfers across the interface are over a bidirectional bussed highway up to 25 bits wide. Eight unidirectional lines control computer to peripheral connexions. Transfer times range from 1 to 5 μs.

1. Introduction

The design and manufacture of real-time systems for military applications involves a large variety of peripherals of widely differing transfer rates: as well as standard data processing devices like magnetic tapes and drums, multiple alpha-numeric and simulated radar displays, radar auto-extractors, analogue-to-digital conversion equipment etc. are used. In many cases, the overall data processing load has required two or three computers to share the task, and in these cases it is required that the system should be configured in such a way that continued service is possible, albeit in a reduced mode, when one computer has failed.

In the past it was the fashion to design a new set of optimized interfaces, and often a new computer, whenever a new requirement appeared. Apart from the development costs involved, this is clearly not a viable approach when one is attempting to offer a range of compatible central processors, any of which may be required to control some standard peripherals and would be expected to use common program packages.

Some three years ago, therefore, the authors' organization embarked upon the design of a standard interface related to real-time on-line multi-computer systems which should still be operable when individual computers or peripherals had failed.

2. Position of the Interface

It was first necessary to determine where this interface should be relative to the whole computing system.

The simplest system is one in which a peripheral is connected directly into the computer via a small amount of special logic. This method is only suitable for a small number of small peripherals over which the program can have complete control, and is not suitable for a real-time system.

It is therefore necessary to have some sort of buffer unit (Fig. 1) between the peripheral and the computer, and the requirements of this unit are fourfold:

(a) It must be able to arrange access to the main core store, and this must be carried out in a way such that the current program is not interrupted.

(b) It must be able to interpret words from the computer as either data or control and pass these as appropriate and at the correct signal levels to the peripheral. Similarly, it must be able to send data and also information about the state of the peripheral, to the computer.

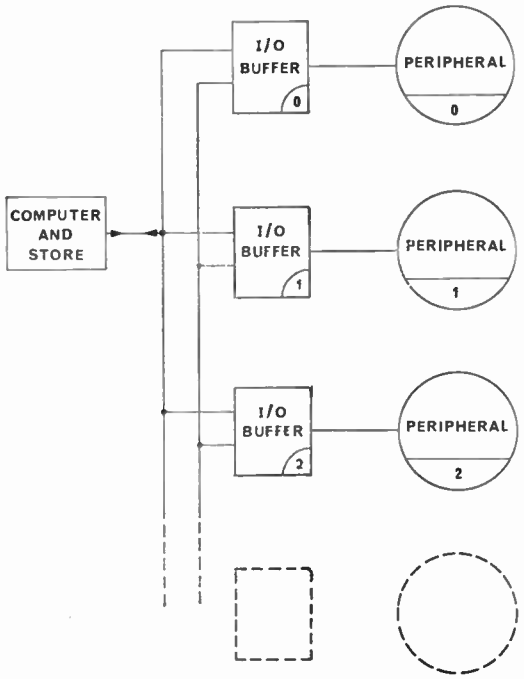


Fig. 1. Basic real-time computer system.

† Ferranti Ltd., Digital Systems Department, Western Road, Bracknell, Berkshire.

(c) In real-time systems, information to or from the core store must not be expected to wait for the peripheral, and so this unit must provide a temporary data store, unless the peripheral device itself contains a store.

(d) It is necessary to provide a facility for giving rapid service to the peripheral whose information is most urgent. Each peripheral in a given system is allocated a priority and simultaneous requests for service must be dealt with according to their relative priorities.

With this arrangement, there could be a common interface between the computer and the various input/output buffer units, but not between the buffer and the peripheral, due to the wide variety of peripheral devices. The buffer units would contain various amounts of special logic, together with the computer interface logic. This latter logic would be similar in all units, and there would need to be interconnection of these units in order to determine priority.

It therefore seemed sensible to split these buffer units into two distinct parts so that the similar blocks of logic could be kept together (Fig. 2). One part, called the computer interrupt equipment (C.I.E.), contains the basic logic for arranging store access, the priority logic, and a number of similar groups of logic depending on the number of peripherals. The other part of the buffer unit is largely unique to each type of peripheral, and is called the peripheral control unit (P.C.U.), the identical parts being represented by the shaded areas. This split also facilitates reduced working requirements in multi-computer systems, since the parts of the buffer unit which effectively cause peripheral or computer failures are isolated from each other and associated with the equipment which would be rendered unserviceable by a failure.

If the interface between the C.I.E. and the P.C.U. is standardized, then the logic at either side of this interface can be standardized, and this leads to a modular expansion concept for the C.I.E. which can result in more flexible system designs.

The C.I.E. is naturally associated very closely with the computer and store, and the P.C.U. is similarly associated with its peripheral. Any long distances which occur in a system would therefore be between the C.I.E. and the P.C.U. It was considered, then, that the standard interface should be between the C.I.E. and the P.C.U.

3. Electrical Characteristics

The method used for transmitting signals across the interface must be capable of fast operation over long distances if the C.I.E. is not to be tied up for long periods with one peripheral. This makes the noise problem and the logical reference level problem more

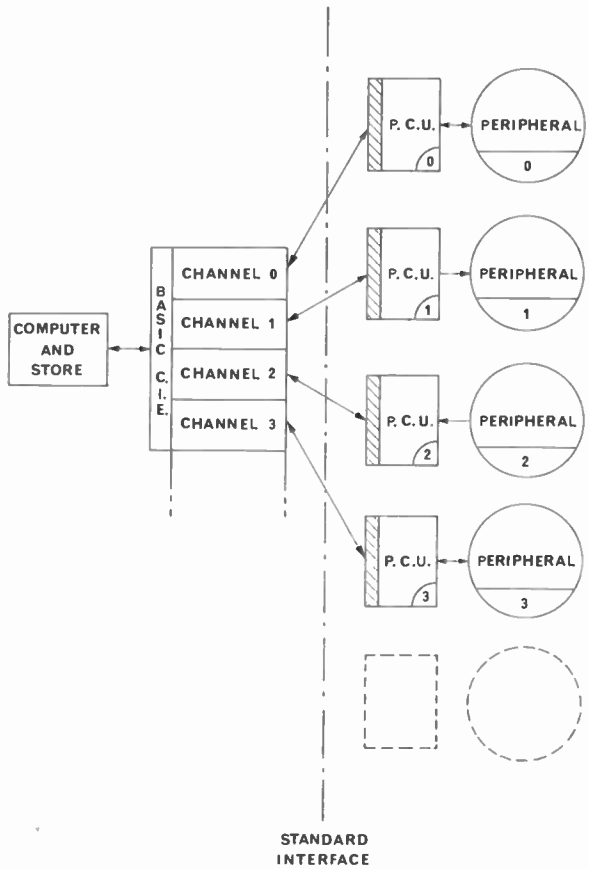


Fig. 2. Modular real-time computer system.

acute. An isolated balanced transmission line system is ideal, giving a large amount of immunity to both induced noise and earth potential differences between the ends of the line. This is not easy to obtain in practice, when some of the control signals across the interface may be steady logical levels. A balanced system was chosen where the line drivers and receivers were referenced to earth, but could tolerate potential differences between line and earth of ± 3 V, and differences of double this amount between the 'earths' at each end of the line. The lines used are twisted pairs in overall screens and therefore the differential noise induced is very small.

The lines are correctly terminated at both ends for both differential and common mode signals. The drivers and receivers are fast balanced devices which present a high shunt impedance to the line, thus causing minimal mis-match.

A further feature of the driver is the ability to gate it off. This enables many drivers to be placed on one line. Since many receivers can also be placed on the line, it can be used to transmit information in both directions. Thus many P.C.U.s can be connected to

one data highway. This is known as a bidirectional bussed data highway system and considerably simplifies wiring and logic at the C.I.E. since only a small number of bussed highways have to be connected instead of individual links to each peripheral.

4. Logical Characteristics

In any standard, it is necessary to strike a balance between cost and facilities offered. Our wide spectrum of requirements has forced us to specify a fairly general interface which would service all the peripherals which we have to control; nevertheless we were, of course, interested in minimizing costs.

4.1. *Data Highway*

It was clear from an inspection of peak data rates on some of the larger systems that it was necessary to cater for the transfer of full length computer words of 24 bits in parallel. To minimize the total number of lines, it was agreed that the data highway would be bi-directional and that it would also be used for the transfer of control and status information and for computer core store addresses when these were required. Peripherals such as paper tape devices, which are naturally less than 24 bits wide, would, of course, only use those lines which they required. A 25th line was also needed to allow the data to be checked for parity when required.

4.2. *Timing*

A further fundamental decision which had to be made was the control timing philosophy. Two basic methods are available: strobe or handshake.

In the strobe system (Fig. 3), the data are accompanied by a valid signal whose leading and trailing edges occur within the data period. Strict timing rules are applied to the width of this strobe pulse and to the minimum amount by which the data must overlap the strobe.

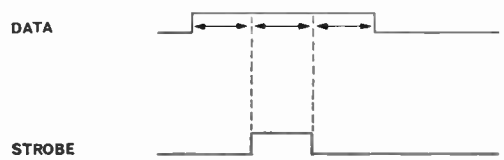


Fig. 3. Strobe system.

In the simplest type of handshake system (Fig. 4), the acceptor requests data with the J signal, the source indicates with the K signal that it has placed the data on the highway, the acceptor removes J when it has accepted the data, and finally the source removes K when it has cleared the highway again. This is the system derived from the N.P.L. interface which is incorporated in the British Standard Interface for Data Collection Systems (B.S. 4421 : 1969).

Compared with the strobe system, handshaking has a number of advantages which derive from the fact that there are no strict timing rules, one J or K edge giving rise to the next: connexion of asynchronous devices is automatically catered for; varying logic speeds at either end of the interface cause no problems and, therefore, improvements in technology can be accommodated without the need to redesign all items of equipment at the same time; the distance between the devices connected does not affect the working of the interface. It does have one disadvantage—more transitions of the interface occur in transferring one word. The available data rate with a given technology is, therefore, slower. Being satisfied with the data rate obtainable from the handshake system, it seemed the obvious choice for our systems.

In the original N.P.L. interface and its derived British Standard, devices are known as acceptors and sources, and there is complete symmetry between the two. A bi-directional device requires two interfaces, one of each type which, for testing purposes, can be interconnected. Computers and peripherals, however, do not have this symmetry. Very often the approximate time at which a transfer occurs is defined by the peripheral, but the precise time is always determined by the computer. This is independent of the direction of transfer and is governed by the relative priority of other peripherals currently making requests and the nature of the instruction being executed at the time.

In the handshake scheme we have adopted, therefore, the J signal always comes from the computer,

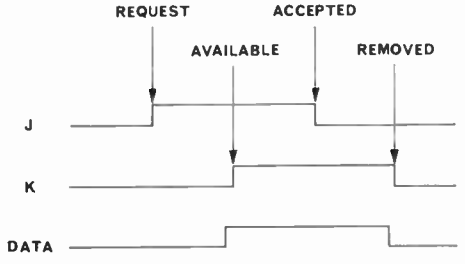


Fig. 4. Handshake system.

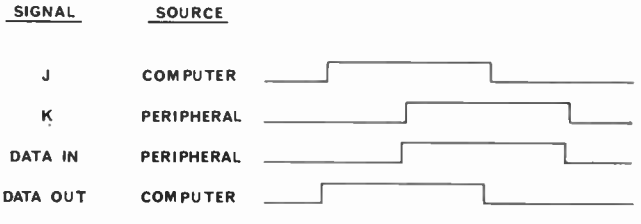


Fig. 5. Ferranti version of handshake system.

and K is the reply from the peripheral (Fig. 5). If data are coming into the computer, the situation is as illustrated earlier: J is the demand for data and K indicates 'data valid'. If data are being output from the computer, however, J indicates 'data valid' and K is the response indicating data accepted.

4.3. *Peripheral Control*

In view of our reduced working requirements, it is clear that in multi-computer systems some peripherals will be connected to more than one computer and a few, for example a drum backing store, could actually be on line to two or three computers at once.

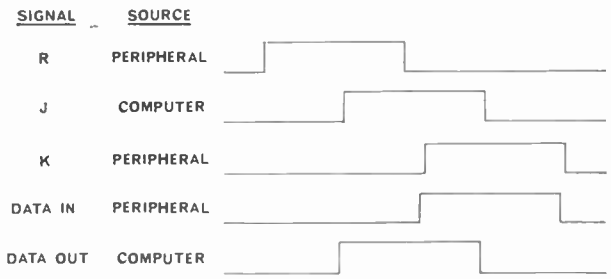


Fig. 6. Interface control.

In addition to the standard problem of peripheral access time, therefore, we had to allow for the possibility of a peripheral actually being busy with another computer when one computer requested a transfer. In a real-time system there must never be a situation in which the computer has to wait on an instruction merely because a peripheral is not ready. The provision of hardware buffer storage in the peripheral was a possibility, but undesirable because of the cost involved. The cleanest alternative appeared to be to arrange that all transfers—control, status, addresses, as well as data—should be requested by the peripheral and result in an interrupt of the computer. The transfer request line, R, completed the trio of control signals (Fig. 6). The peripheral turns R off before raising K if it does not want a further transfer immediately. Since R was the single line which requested one of a number of types of transfer, it had to be qualified by three further lines which defined the transfer required.

Only the input of status information caused a 'program interrupt'. All the remaining requests caused what are referred to as 'data interrupts', which merely required a short pause in the current instruction while the data were transferred to or from the store. These are, of course, equivalent to what elsewhere are referred to by titles like 'autonomous data transfers', 'cycle stealing', 'data break', etc.

Table 1
Control signals

Signal	Name	Source	Use
M	COMMAND	computer	Computer control of peripheral
S	PERIPHERAL SELECT	computer	
R	REQUEST	peripheral	
A	ADDRESS	peripheral	Define R
Q	QUALIFIER	peripheral	
D	DIRECTION	peripheral	
J		computer	Handshake Control
K		peripheral	

4.4. *Computer Command*

The situation so far described was perfectly adequate, provided the computer never required to initiate peripheral transfers. In most cases this was necessary; there was, therefore, a requirement for a computer command signal to enable the computer to instruct the peripheral to take some action. When the single line was not sufficient to define the type of action called for, the peripheral was required to extract further control information from the computer when it was ready.

Recalling the failure mode reduced working requirements, a further signal under program control was required to enable the 'disconnexion' of a faulty peripheral or a peripheral which was on line to another computer. This 'peripheral select' line operated effectively as a gating signal, so that when a peripheral was not selected, commands from the computer were to be ignored and no request signals were to be sent. At the computer end, the select line controlled whether or not request signals from the peripheral were accepted. By these means, it was possible for a computer to ignore a faulty peripheral to which it was connected, and for a serviceable peripheral to work with a serviceable computer even though it was also connected to a faulty computer.

4.5. *Peripheral Addressing*

It is argued above that there were advantages to be gained in having a bussed data highway, but the possibility of putting the control signals on bussed lines has not yet been discussed.

The request signals from the peripherals and the associated lines A, Q, D (Table 1) were all independent

and, therefore, could not be bussed on a common highway. The J signals, also, were sent to individual peripherals and could not be bussed unless accompanied by a separate address highway. 'Computer command' and 'peripheral select' could be long term signals and their state was always important. The only way in which these could have been bussed would have been to use the lines only to indicate changes of state and have repeater flip-flops in the peripherals. Again an associated address highway would have been required. The only control signal which could have been unconditionally bussed was K, since only one K should be up at any time, and its source is known.

To summarize: four of the control signals could not be bussed and four could, but only at the cost of extra address highways and extra peripheral logic. The simple advantages of bussing the data highways were clearly not available for the control signals.

5. The Final Choice

The final choice, then, is an interface on which all information transfers are over a bidirectional bussed highway up to 25 bits wide. In addition, eight individual unidirectional lines (Table 1) control each computer to peripheral connexion. M is the command signal, with which the computer initiates transfers, and S the peripheral select signal, in the absence of which a peripheral is effectively disconnected. All transfers are requested by the peripheral with R, the type of transfer required being defined by A, Q and D. J and K are the handshake control signals.

A, Q, D have a mnemonic interpretation (Table 2): when A is a '1' a core store address is to be transferred; Q is a qualifier bit, when it is a '1' control or status information is to be transferred; D is the direction bit. Of the eight possible states of A, Q, D, seven are defined. 'Data in' and 'data out' are self-explanatory —the core store address involved being held in the computer. 'Status in' is the only one which causes a program interrupt—all the others being data interrupts. 'Control out' is used when M does not provide sufficient information to define the mode of operation required; again the core store address involved is held in the computer. 'Address in' and 'address out' are used in multiplexing type peripherals where it is necessary for the peripheral to define what core store address should be used in a subsequent data transfer. In 'store increment', unity is added to the location whose address is specified on the data highway. This facility is useful in a variety of applications which come broadly under the heading of event counting.

Associated with the standard interface we have a code of practice which covers both hardware and software aspects. In this connexion we have developed a standard usage of the core store (Fig. 7). Four locations are reserved at the end of the store for each interface channel—channel 0 having the four largest addresses, channel 1 the previous four, and so on.

For each channel, the four words have the same uses, which are shown on the right of Fig. 7. The control pointer defines the address from which the next control word will be extracted. During the associated data interrupt, the pointer will be indexed by unity so that a subsequent control word will be extracted from the adjacent location. The data pointer carries out precisely the same job for data transfers. The remaining two locations are both associated with status input; the status information is placed in 3+ and the starting address of the associated interrupt program is held in 2+.

Table 2

Types of transfer

A	Q	D	
0	0	0	DATA IN
0	0	1	DATA OUT
0	1	0	STATUS IN (PROGRAM INTERRUPT)
0	1	1	CONTROL OUT
1	0	0	ADDRESS IN
1	0	1	ADDRESS OUT
1	1	0	STORE INCREMENT
1	1	1	UNDEFINED

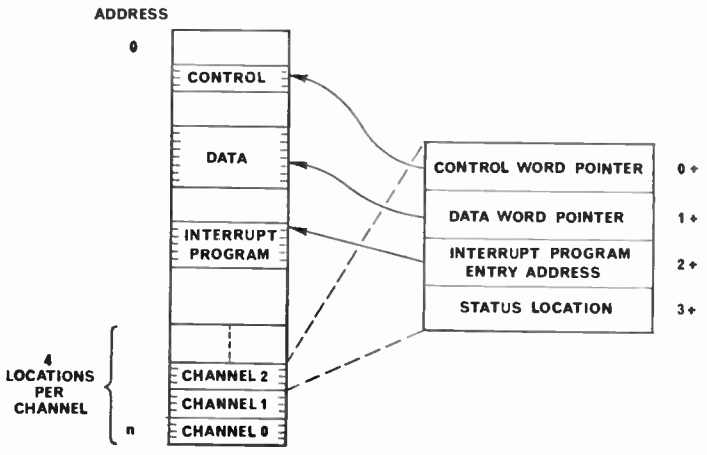


Fig. 7. Core store usage.

6. Practical Implications

6.1. *Speed*

No interface description is complete without an indication of its speed of operation. In a general description of this nature, however, it is difficult to know what figures to quote; in fact we have already stated that one of the reasons for selecting this transfer mechanism was that there are no timing rules and, therefore, no theoretical transfer rates can be predicted.

In any given implementation of the interface there are a number of factors which govern the speed of operation: circuit speed, degree of buffering employed in C.I.E. and P.C.U., distance between them, whether they are synchronous or asynchronous, the priority assigned to a particular P.C.U. at the C.I.E., and the degree of autonomy of the C.I.E. logic associated with each P.C.U. With asynchronous logic and gate delays of the order of 10 ns, it is possible to conceive of an RJ turn-around time of, say, 200 ns when the request is accepted immediately and the priority assessment proves it to be the highest priority request present. Assuming data are always ready for transmission and a register is available for reception, JK turnaround times of, say 50 ns would also be possible —giving a total spread of 350 ns. In a practical system, generous tolerances must be allowed, and one would probably moderate this time to, say, 500 ns. To this must be added cable delay time; since five transitions of the interface occur, an approximate figure for this is 10 ns per foot (33 ns per metre) of cable length.

It must be emphasized that these are theoretical figures which we have not attempted to achieve. In our various implementations we have used synchronous logic, with independent clocks, at either side of the interface, the clock rates varying over the range of 2–6 MHz. We have usually been involved with a 1 or 2 μs core store as a buffer at the computer end and have always had to cater for a variable response time to the R signal: due to variations of interrupt waiting time, in the range 0–3 μs, and to interrupt queues which, in worst case situations, can amount to some tens of microseconds. Excluding the variable response times, variations in transfer time have ranged from 1 to 5 μs, due primarily to differing core store speeds and buffer facilities.

6.2. *Hardware*

Because this interface has eight control lines which are mandatory for all peripherals, and no additions are permitted, the logic at both sides of the interface can be standard for all systems. Having decided on the most practical way of implementing the interface requirements, this logic only had to be designed once, and printed circuit cards have been produced which deal with all the interface lines for both the C.I.E. and the P.C.U.

The printed circuit cards used can contain up to 50 dual-in-line packages. The C.I.E. consists of a few cards of basic logic plus the repetitive logic associated with each highway and each P.C.U. Two identical cards provide gating and drivers/receivers for each 24-bit bussed data highway. The logic associated with each P.C.U. occupies half a card, therefore one further card type allows the channel capability of the C.I.E. to be increased in multiples of 2. The logic on these cards is independent of the type of peripheral with which it may ultimately be used, and therefore the cards are standard.

The P.C.U.s require a standard interface card which can be used as a basis for the control logic. In our systems there are two requirements to be met: one is for simple peripherals to be connected only to one C.I.E.; and the other is for more complex peripherals which may be connected to two C.I.E.s in a multi-computer system.

For the simple peripherals (paper tape or card equipment, plotters, etc.) a standard card contains the standard interface logic, and an 8-bit buffer register, together with all the necessary line drivers and receivers. The data highway width can, of course, be increased by the addition of register and highway driver/receiver cards. For this type of peripheral, part of another card is usually sufficient to contain the rest of the P.C.U. control logic, and interface circuits to the peripheral device, unless that device happens to require high current or voltage input/outputs.

For the more complex periphals (generally any backing store), an alternative card is used which contains the standard interface logic and drivers/receivers for two sets of control lines, together with the logic to determine which of the two C.I.E.s should be given service. For such peripherals several extra cards are usually required to complete the P.C.U.

When medium-scale integration is available for customer design at reasonable cost, then the logic at present contained on the four standard interface cards is an obvious choice for the new technology.

6.3. *Fault Detection Philosophy*

The interface specification does not define any fixed rules for fault detection and diagnosis, but detection and repair of faults in the intimate interface area is relatively straightforward thanks to the use of the standard logic cards at either side of the interface. Failure of the RJ sequence is detected and indicated by the P.C.U., while a JK handshake failure is detected in the C.I.E. The latter detection results in clearance

of the S line and hence isolation of the peripheral on the faulty channel, and an indication to the supervisor program.

Test transfers are always required in an on-line system, and these can be used to detect failures on the data highway. If immediate fault detection is required, or the equipment is in a noisy environment, a parity option is available to give an immediate indication of data failure.

7. Conclusion

The adoption of any standard interface and a related code of practice, which applies to electrical, logical and programming aspects, results in greater efficiency in designing, producing and maintaining a computer system. In general, because it is standard, there is a certain amount of redundancy inherent in facilities unused by individual peripherals. One of the advantages of the interface described is that there are few control lines which may not be needed by a peripheral; since the logic associated with these lines is trivial, there is little hardware redundancy in unused facilities. Thus there is no need to have optional lines across the interface with their attendant design problems.

The standard use of core store locations enables a modular programming approach to be adopted whereby the channel number to which a peripheral is connected is used as a simple parameter. There are thus facilities for straightforward interchangeability of units, which considerably simplifies commissioning and maintenance problems.

Experience in using this interface has shown that the positive virtues of this particular standard more than outweigh the minimal redundancies involved. The only types of peripheral discovered in which problems arise in using this interface are fast multiplexers where data need to be built up in separate core store buffer areas. For such devices there is a time penalty in using this interface; the provision of a separate core store address highway would have been an advantage. Since these devices only occur rarely in the systems with which the interface is used, this omission is of no consequence.

Manuscript first received by the Institution on 14th November 1968 and in final for mon 14th March 1969. (Paper No. 1263/Comp 118).

Radio Engineering Overseas . . .

The following abstracts are taken from Commonwealth, European and Asian journals received by the Institution's Library. Abstracts of papers published in American journals are not included because they are available in many other publications. Members who wish to consult any of the papers quoted should apply to the Librarian giving full bibliographical details, i.e. title, author, journal and date, of the paper required. All papers are in the language of the country of origin of the journal unless otherwise stated. Translations cannot be supplied.

VARIABLE ATTENUATOR

The circuit construction of a wideband transistor variable attenuator, which is suitable as a gain controller for wideband transistor amplifiers operated at high level, is investigated, both theoretically and experimentally, in a paper from the Electrical Communications Laboratory of the Nippon Telephone and Telegraph Company. The characteristics of variable attenuators based on this analysis at the input level of 7 dBm and at the maximum attenuation of 10 dB are as follows:

(1) The variation of the differential gain is less than 0·1%, and the variation of the delay time is less than 1 ns over the transmission bandwidth of 70 ± 10 MHz.

(2) The variation of the amplitude is less than 0·2 dB over the transmission bandwidth of 140 ± 20 MHz.

As a consequence, it is found that this variable attenuator can be used as an a.g.c. (automatic gain control) element for a wideband transistor amplifier operating in the 70 MHz or the 140 MHz band.

'Wideband transistor variolosser', T. Watanabe and T. Nakamura, *Electronics and Communications in Japan* (English language edition of *Denshi Tsushin Gakkai Ronbunshi*), **51**, No. 1, pp. 60–67, January 1968.

DIFFERENTIAL DISTORTION IN COLOUR TELEVISION

The method of chrominance signal transmission varies from one system of colour television transmission to another, so that the effect of non-linear distortion introduced in the transmission path on the quality of the picture reproduction likewise varies.

In a Soviet paper a brief description outlining the susceptibility of various systems of colour television transmission to differential distortion is given. For instance, in the N.T.S.C. system the colour tone is determined by the sub-carrier phase, and the saturation is determined by its amplitude. Distortions in phase and amplitude, which depend on the brightness level, thus cause colour distortion over large areas of the reproduced scene.

Interlaced switching of the phase of one of the chrominance signals in the PAL system leads to differential-phase shifts of different signs, which are compensated in the receiver. The N.T.S.C. and PAL systems are identical as regards the influence of differential gain.

In the SECAM system the colour sub-carrier is frequency-modulated and the colour-difference signals are interlaced. It would seem at first sight that there could be no differential distortion at all in this case, since the colour tone is determined by a frequency ratio, while the amplitude-limiting of the f.m. signals in the receiver eliminates any effect of parasitic variations in the sub-carrier amplitude. Though this is not strictly true, the system is undoubtedly fairly insensitive to non-linear distortion. In this system the differential distortion becomes apparent on transients and this is treated in greater detail.

The amount of correction necessary for a given subjective quality is discussed.

'Differential distortion in colour television', P. M. Kopylov, *Telecommunications and Radio Engineering* (English language translation of *Elektrosvyaz*' and *Radiotekhnika*), No. 3, pp. 47–50, March 1968.

V.H.F. TRACKING ANTENNA FOR FLYING OBJECTS

In order to receive signals transmitted from rockets or artificial satellites, the receiving antenna on earth must possess the ability to track the rocket or satellite by shifting its directional pattern, because the signal direction shifts with time. Parabolic antennas in the microwave band and helical antennas in the v.h.f. band have been used for this purpose, and are installed on a movable mounting of the variable azimuth-elevated type. In this system, however, because the whole antenna system of considerable weight must be rotated, its structure tends to be very complicated. Furthermore, it has the disadvantage that it is difficult in principle to track objects passing near the zenith.

Of the antenna arrays consisting of a large number of elements, the phased array, in which the phase of each element is properly controlled, is suitable for high-speed tracking of a flying object because it is capable of changing the beam direction without mechanically moving the equipment itself. As a variable-phase mechanism in the v.h.f. band, the author of a Japanese paper proposes a cross-dipole element and shows that an electrical phase shift of this element corresponds to a local rotation. With respect to a two-dimensional square array consisting of such elements, the writer analyses its directional characteristic, beam scanning and the rotating means.

The major design points and the results of performance tests are discussed. The gain was 16 dB and the tracking speed was above 1·25%/s. At the present time this antenna array is being used as part of an artificial satellite telemetering receiver system.

'A vehicle-tracking antenna array composed of cross dipoles', K. Takaeo, *Electronics and Communications in Japan* (English language translation of *Denshi Tsushin Gakkai Ronbunshi*), **51** No. 6, pp. 78–85, June 1968.

SUBJECT INDEX

Papers and major articles are denoted by printing the page numbers in bold type

INDEX OF PERSONS

Names of authors of papers published in the volume are indicated by bold numerals for the page reference.
Authors of papers which are given in abstract form are denoted by A.
Biographical references are denoted by B. Contributors to discussions are indicated by D.

INDEX OF ABSTRACTS

This index classifies under subject headings the abstracts throughout the volume in 'Radio Engineering Overseas'.

OTHER JOURNALS FROM WHICH ABSTRACTS HAVE BEEN TAKEN DURING THE FIRST HALF OF 1969

Archiv der Elektrischen Übertragung (Germany)

E.B.U. Review (Switzerland)

Electronics and Communications in Japan (English language edition of *Denshi Tsushin Gakkai Ronbunshi*)

Ericsson Technics (Sweden)

Nachrichtentechnische Zeitschrift (Germany)

Philips Technical Review (Holland)

Proceedings of the Institution of Radio and Electronics Engineers, Australia

Telecommunications and Radio Engineering (Translation of *Elektrosvyaz* and *Radiotekhnika*) (U.S.S.R.)