

Controlling the speed of small induction motors by means of thyristors

A. Walraven

The development of the silicon controlled rectifier (thyristor) has accelerated the development of motor control systems. The application of thyristors to such systems, which formerly used saturable reactors or thyratrons, allows considerable simplification and space saving. In particular, brushless control of the torque and speed of small induction motors can be achieved with a minimum number of components and with the circuit directly fed from the mains supply.

Small electric motors, with powers of up to 100 W, whose speed must be variable stepwise or continuously, are often used in professional equipment and in industrial control. Examples include the capstan drive in tape recorders for professional sound recording or for the recording of measurement data ("instrumentation recorders"), wire or tape winding and the speed control of small laboratory centrifuges. The emphasis in any given application may lie on a few very accurately controlled speeds, as is the case with a tape recorder, or on a continuously variable speed over a given range, e.g. for centrifuges, or again on a speed varying according to a certain programme of operations, e.g. in a wire-winding installation.

Collector motors or eddy current couplings [1], are widely used for these purposes. However, controlled induction motors can be used with advantage for motor powers up to about 100 W. Although the control of induction motors by saturable reactors or thyratrons has long been familiar, various methods based on these principles have become attractive only through the development of the silicon controlled rectifier (the thyristor) [2]. Compared with the gas-filled thyatron, this device has the advantages of a lower internal dissipation and smaller volume, and it does not need the heater voltage source required by the thyatron.

The speed of certain types of induction motor can

be very simply controlled with the help of thyristors, and some examples of the resulting control systems that have been used in practice will be discussed here.

Since the behaviour of a control system for an induction motor depends very greatly indeed on the motor parameters, we shall first of all briefly discuss some important characteristics of the induction motor.

The characteristics of the induction motor

Fig. 1 is a diagram of a three-phase induction motor. The three identical stator windings, or phases, U , V and W are considered as being connected to the symmetrical voltages E_R , E_S and E_T of the three-phase

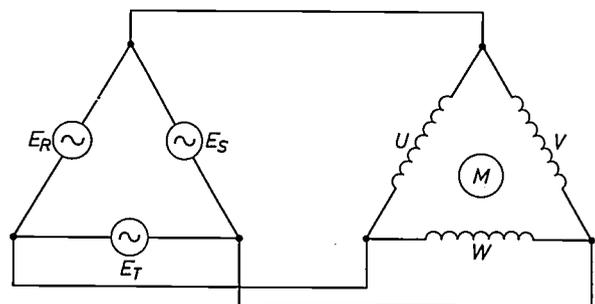


Fig. 1. Diagram of a three-phase induction motor connected to a three-phase supply. Motor phases U , V and W are connected to supply voltages E_R , E_S and E_T .

[1] W. Bähler and W. van der Hoek, An eddy-current coupling employed as a variable-speed drive, Philips tech. Rev. 27, 15-21, 1966 (No. 1).

[2] K. Steimel, Käfigläufermotoren und Thyristor, AEG-Mitt. 54, 87-88, 1964.

supply. The speed n_0 of the rotating field produced at the stator is given by:

$$n_0 = \frac{60\omega}{2\pi p} \text{ r.p.m.,} \dots \dots \dots (1)$$

where ω is the angular frequency of the supply and p is the number of pole-pairs per phase in the stator. As a result of the slip s between the rotating field and the rotor the speed of the rotor is:

$$n = (1 - s)n_0. \dots \dots \dots (2)$$

The motor may now be regarded as a three-phase transformer, with the rotor as its short-circuited secondary, whose equivalent circuit diagram, for a single phase, is given in fig. 2a. Neglecting the stator leakage L_1 , the stator resistance R_1 , the iron loss resistance R_m

of the rotor, we find the following expression for P_m :

$$P_m = T\omega_m = \frac{\omega(1-s)T}{p} \dots \dots \dots (6)$$

Combining (6) with (5) and (3) gives T as a function of s and the motor parameters:

$$T = \frac{pqE^2}{\omega^2 L_2 \left(\frac{\omega L_2 s}{R_2} + \frac{R_2}{\omega L_2 s} \right)} \dots \dots \dots (7)$$

The motor torque reaches a maximum value T_k at a certain value of the slip, the "critical slip" s_k :

$$T_k = \frac{pq}{2L_2} \left(\frac{E}{\omega} \right)^2; \quad s_k = \frac{R_2}{\omega L_2} \dots \dots \dots (8)$$

With the aid of (8), (7) may also be written in a form

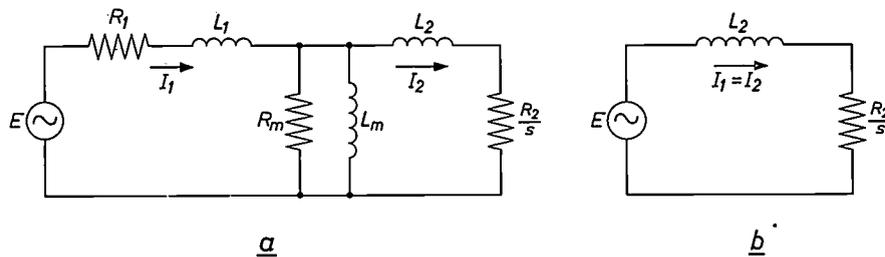


Fig. 2. a) Equivalent circuit diagram of one of the phases of an induction motor regarded as a (symmetrical) three-phase transformer. R_2/s and L_2 are the rotor resistance and rotor leakage transferred to the primary side, while R_m and L_m represent the stator core losses and the magnetization current. R_1 and L_1 are the stator resistance and the stator leakage. b) The simplified equivalent circuit diagram arising from a when R_1 , L_1 , R_m and L_m are neglected. The current I_1 taken from the mains supply is now equal to the reduced rotor current I_2 .

and the magnetizing inductance L_m gives the simplified diagram in fig. 2b. This includes only the rotor leakage L_2 and the rotor resistance R_2/s , connected to the supply voltage E .

The power taken from the supply by a motor with q phases is q times the power taken per phase. From fig. 2b we find the following expression for the total power consumed:

$$P = \frac{qI_2^2 R_2}{s} = \frac{q E^2 R_2}{s\omega^2 L_2^2 + \frac{R_2^2}{s}} \dots \dots \dots (3)$$

The dissipation P_{R_2} occurring in R_2 is a part of P , and is given by:

$$P_{R_2} = qI_2^2 R_2 = sP. \dots \dots \dots (4)$$

The mechanical power delivered to the shaft is found from the difference between P and P_{R_2} :

$$P_m = P - P_{R_2} = (1 - s)P. \dots \dots \dots (5)$$

Since the mechanical power is defined as the product of the shaft torque T and the angular velocity ω_m

identical with an equation for the eddy-current coupling [1]:

$$T = \frac{2T_k}{\frac{s}{s_k} + \frac{s_k}{s}} \dots \dots \dots (9)$$

Like the induction motor the eddy-current coupling has a rotating field set up by d.c.-powered magnets in a wheel (or inductor), driven by an auxiliary motor. The currents induced in the rotor by this rotating field produce the rotor torque in the same way as in the induction motor. Variation of the energizing current produces a proportional variation in the amplitude of the rotating field and therefore a variation in torque proportional to the square of the current variation. With the induction motor, the torque depends in a similar way on E^2 , as is shown by eqs. (7) or (8). With E and ω constant, the behaviour of T as a function of s or n follows from (9) and (2). Fig. 3 shows this dependence of T on s or n at various values of the critical slip s_k .

Finally, it follows from (5) that the efficiency of the

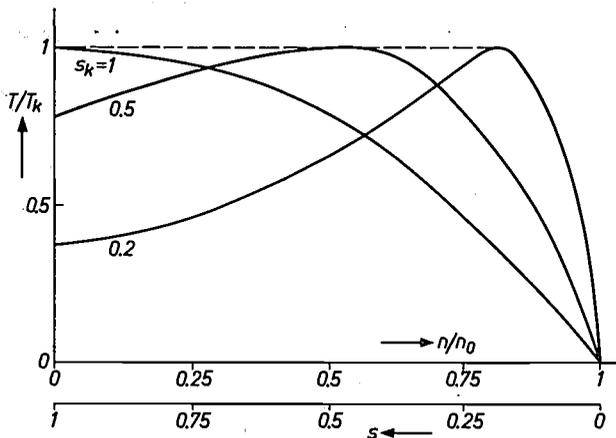


Fig. 3. Normalized torque-speed curves of the induction motor for various values of the critical slip s_k . The curves show the ratio of the torque T to the maximum (critical) torque T_k as a function of the speed n (shown as a fraction of the synchronous speed n_0), or of the slip s .

induction motor is independent of the motor parameters and is determined by the slip:

$$\eta = \frac{P_m}{P} = 1 - s. \quad \dots \quad (10)$$

Methods of controlling induction motors

Induction motors are generally controlled by varying either the stator voltage (E), the rotor resistance (R_2), or the frequency of the supply (ω). All three methods can be realized with the use of thyristors. The simplest method is stator voltage control which has been applied exclusively in the examples discussed in this article. We shall now give a more detailed account of this method.

Varying the stator voltage

By varying the stator voltage E with the frequency of the supply remaining constant, T or n may be controlled. This method is particularly suitable for the control of small induction motors of up to 100 W of

the two-phase capacitor type and of single-phase motors with short-circuit windings (shaded pole motors).

Fig. 4 shows the principle of this method of control applied to a two-phase motor with a phase-shifting capacitor. Only one thyristor (Th) is needed in the circuit shown. This is included in the d.c. branch of a bridge rectifier and controls the fraction both of the positive and of the negative half-cycles of the supply voltage applied to the motor. This is done by firing the thyristor Th periodically with pulses at twice the supply frequency, supplied from a pulse generator PG , the thyristor being automatically extinguished when the load current changes sign. Control of T or n can be obtained if the conduction angle φ is made to depend on the deviation of T or n from a given reference value.

The resultant motor voltage E_m now consists of "segments" from the sinusoidal supply voltage $E_p \sin \omega t$ which correspond to the conduction angle φ of the thyristor. Fourier expansion of E_m gives a series of odd harmonics of the supply frequency:

$$E_m = \sum_{n=1}^{\infty} \{ a_{2n-1} \cos (2n-1)\omega t + b_{2n-1} \sin (2n-1)\omega t \},$$

$$a_{2n-1} = \frac{E_p}{\pi} \left\{ \frac{\cos 2n\varphi - 1}{2n} + \frac{\cos 2(n-1)\varphi - 1}{2(n-1)} \right\},$$

$$b_{2n-1} = \frac{E_p}{\pi} \left\{ \frac{\sin 2(n-1)\varphi}{2(n-1)} - \frac{\sin 2n\varphi}{2n} \right\}.$$

... (11)

The fundamental component of E_m has an amplitude of

$$A_1 = \sqrt{a_1^2 + b_1^2}$$

$$= \frac{E_p}{\pi} \sqrt{\frac{1}{2}(1 - \cos 2\varphi - 2\varphi \sin 2\varphi + 2\varphi^2)} \quad (12)$$

and provides a contribution to T proportional to A_1^2 (expression 7).

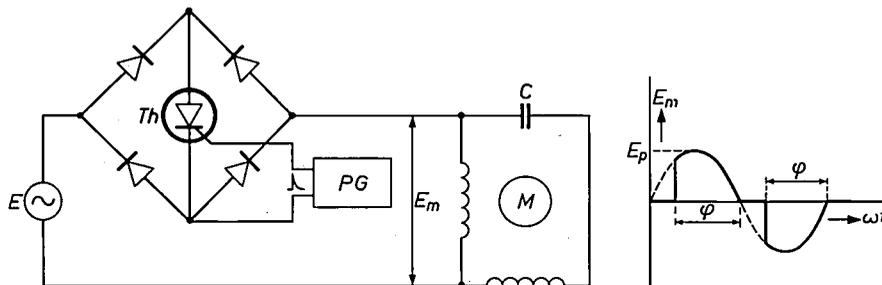


Fig. 4. The supply voltage E_m of the two-phase capacitor motor is varied by means of a thyristor (Th) in a bridge circuit. The voltage E_m consists of segments of the mains voltage E corresponding to the conduction angle φ of the thyristor. This is ignited by ignition pulses from the pulse generator PG . The instant of ignition is controlled by an actuating error signal derived from torque or speed error.

The variation of T as a function of φ and E was examined by the measurement of the starting torque of a 50 W two-phase capacitor motor. This motor had a solid iron rotor to provide a high rotor resistance and starting torque ($s_k \approx 1$, see fig. 3). The contribution to T made by the fundamental component of E_m , as calculated from (12), is shown as a function of φ (with E_p constant) by curve 1 in fig. 5. The measured behaviour of the total starting torque is shown by curve 2. The close agreement between these two curves shows that the effect of the higher harmonics on the starting torque is virtually negligible. This is due not

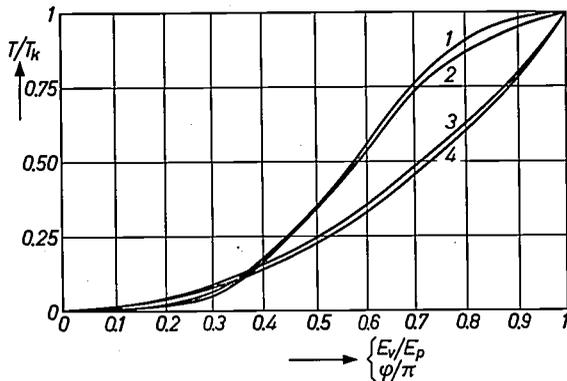


Fig. 5. The starting torque of an induction motor fed via a thyristor circuit as shown in fig. 4 (curves 1 and 2) or via a control transformer (curves 3 and 4). The torque is normalized to the maximum (critical) torque T_k . Curves 1 and 3 were calculated, while 2 and 4 represent the measured variation of T/T_k . E_v is the voltage from the variable transformer, E_p is the rated motor voltage.

only to the lower amplitudes of these harmonics but also to the connection of the second motor phase via the phase-shifting capacitor. The capacitor has a value such that, at the supply frequency, the currents through both stator phases are in the complex ratio desired for the excitation of a symmetrical rotating field (equal amplitudes and a phase shift of 90°). This condition cannot be satisfied at the same time for the higher harmonics, resulting in severe distortion of the harmonic rotating fields and strongly reducing their contributions to the starting torque.

Finally, fig. 5 also shows the measured and calculated curves of T as a function of the amplitude of a purely sinusoidal E_m . The measured curve 4 is a good approximation to the calculated parabolic shape of curve 3. The variable motor voltage $E_v \sin \omega t$ in this case was taken from a variable transformer allowing E_v to be varied from 0 to 100% of the nominal motor voltage E_p . The thyristor method is found to provide a rather better control characteristic than is obtained by means of a variable transformer: over 40 to 70% of the thyristor control range the characteristic is almost linear.

If, in this method, the frequency of the rotating field is kept constant, the controllable speed is automat-

ically restricted to the range $0 < n < n_0$. A high torque over a relatively wide speed range may be obtained by appropriate choice of R_2 and L_2 ($s_k \approx 0.5$, see eq. 8 and fig. 3), or by the use of special rotor designs which enable the torque-speed curve to approach the ideal rectangle [3] [4].

The advantage of this method of control is its simplicity for in certain cases (fig. 4) it allows the torque and the speed of an induction motor to be controlled by only one thyristor. A drawback is the poor efficiency at low speeds ($\eta = 1 - s$), causing the heat evolved to be greatest when the rotor is stationary. Motor cooling can be improved by the use of a separate cooling motor (or "blower motor"). If this is not used, the motor power is restricted to the value of about 100 W mentioned above in connection with the permissible increase in motor temperature.

Varying the rotor resistance

Varying the rotor resistance requires the use of a "slip-ring motor", in which the rotor winding can be connected to an external circuit via slip-rings. External resistors connected to the rotor winding can now be used to vary s_k and with it the motor characteristic (fig. 3). Here, T_k remains constant (eq. 8) but can be moved to any lower speed, even a negative one, by an increase in the rotor resistance. For $s_k = 1$, the maximum torque T_k occurs at a motor speed of zero: this can be used in starting the motor.

The continuous variation of the rotor resistance which this method requires can be obtained with thyristors by means of the scheme shown in fig. 6. The rotor voltages appearing at the three slip-rings are supplied to the external rotor load R_1 via choke L_1 after rectification in a three-phase bridge rectifier. R_1 is periodically short-circuited during a certain time interval T_1 by thyristor Th_1 , which can be extinguished again with the aid of the circuit Th_2-L_2-D-C , coming into operation when thyristor Th_2 is ignited. An average rotor resistance R_{av} is produced on periodic repetition of period T_2 :

$$R_{av} = \frac{T_2 R_1 (R_1 + R_1)}{T_2 R_1 + T_1 R_1}, \quad 0 \leq T_1 \leq T_2, \quad \dots \quad (13)$$

where R_1 is the internal resistance of the rotor winding. Thyristors Th_1 and Th_2 are ignited with a time difference T_1 by the pulse generator PG . Varying T_1 gives a continuous variation of R_{av} between the limits R_1 and $R_1 + R_1$. The switching frequency $f_s = 1/T_2$ is not critical [5] and, in practice, is between 50 and 1500 Hz.

This method is not suitable for small motors, since the slip-

[3] W. J. Gibbs, Induction and synchronous motors with unlaminated rotors, J. IEE 95-II, 411-420, 1948.

[4] P. L. Alger, G. Angst and W. M. Schweder, Saturistors and low starting current induction motors, IEEE Trans. Power App. Syst., June 1963, pp. 291-298.

[5] An interesting variant of this method ensues by using a variable switching frequency and a fixed proportion of T_1 and T_2 . In this case the speed n can be synchronized with the switching frequency. For a more elaborate description of this system see J. Lemmrich, Der synchronisierte Induktionsmotor, ETZ A 85, 724-726, 1964.

[6] L. Abraham and U. Patzschke, Pulstechnik für die Drehzahlsteuerung von Asynchronmotoren, AEG-Mitt. 54, 133-140, 1964.

[7] D. A. Bradley, C. D. Clarke, R. M. Davis and D. A. Jones, Adjustable frequency inverters and their application to variable-speed drives, Proc. IEE 111, 1833-1846, 1964.

ring design is used only for higher powers (more than 2 h.p.). The R_1 inherent in the motor and the permissible degree of heating of the motor determine the limits to torque and speed. As with control by varying the stator voltage, the efficiency is poor at low speeds because of the dissipation in R_1 . At high powers, this energy can be removed by feeding it back into the supply mains [6].

Varying the supply frequency

It follows from eqs. (1) and (2) that n can be changed by varying the supply frequency ω . In order to maintain the same T_k as ω increases, the mains voltage must be varied as well (eq. 8). The principle of this method of control is given in fig. 7. The mains voltage is first rectified with a controlled rectifier (R), the output signal of which is converted by an inverter (I) into a new three-phase supply, now with variable frequency. The control signal e controls both the rectifier and the inverter in order to obtain a motor voltage which increases with the output frequency.

Because of the large number of thyristors required [7], this system is an attractive proposition only where the motor power exceeds 1 h.p.; low slip and hence high efficiency can then be obtained by the use of a motor with a low s_k (squirrel-cage motor).

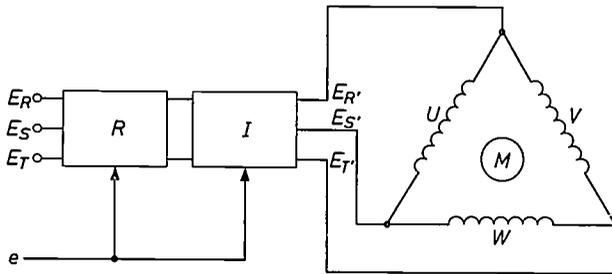


Fig. 7. Diagram of the control of an induction motor by variation of the stator frequency. The 50 Hz three-phase mains supply (E_R, E_S, E_T) is converted by rectifier R and inverter I into a new three-phase supply ($E_{R'}, E_{S'}, E_{T'}$) with a variable frequency and proportionately varying voltages. The desired increase of the motor voltage with frequency is obtained by controlling both R and I with the control signal e .

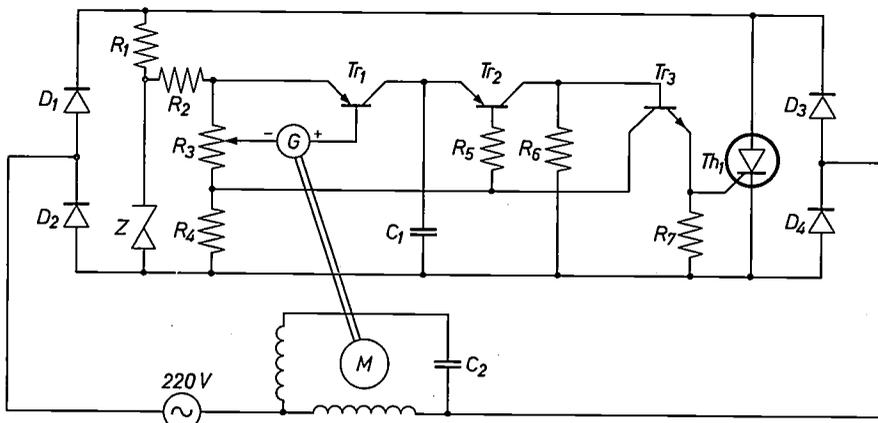


Fig. 8. Simple speed control of an induction motor by means of a thyristor Th . The motor speed is measured by the tachometer G , while the desired speed is set by the potentiometer R_3 . The signal containing the speed fluctuations is introduced at the base of Tr_1 and controls the conduction angle of Th_1 and hence the motor torque via trigger circuit Tr_2, Tr_3 .

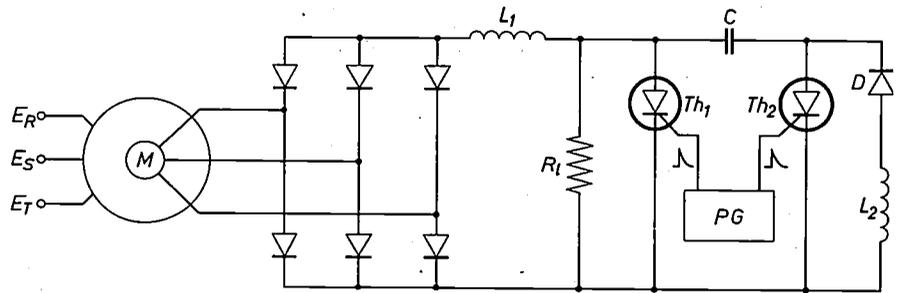


Fig. 6. Diagrammatic representation of a slip-ring motor controlled by variation of the rotor resistance. The rotor voltage rectified by a three-phase bridge is connected to the (external) rotor load R_1 , which is periodically short-circuited with the aid of Th_1 and Th_2 . Th_1 can be extinguished by C, Th_2, D and L_2 . The motor torque is controlled by varying the average rotor current flowing through R_1 (pulse width modulation).

Tape-speed control for tape recorders

In professional tape recording installations, the tape speed may deviate from the rated speed by no more than $\pm 0.1\%$. This means that the mechanism for moving the tape has to meet very exacting requirements. The tape is transported by pressing it by means of a rubber roller against an accurately machined spindle, the "capstan". This is generally driven by a relatively expensive synchronous motor, in which two different tape speeds, e.g. 15 and $7\frac{1}{2}$ inches per second, can be obtained by switching the number of stator poles.

An alternative solution, employing a thyristor-controlled induction motor, can be obtained as shown in the diagram of fig. 8.

When thyristor Th_1 is non-conducting, the rectified supply voltage appears across the d.c. side of the bridge rectifier formed by D_1 to D_4 . This voltage, limited by the Zener diode Z , serves as a combined supply and synchronization voltage for the ignition circuit. The thyristor is automatically extinguished at the reversal of sign of the load current, and its anode current drops below the holding current, whereupon the Zener voltage attains the value E_z within a short time (fig. 9). A saw-tooth voltage now appears across capacitor C_1 through the charging action

of transistor Tr_1 acting as a current source. As soon as this voltage rises beyond the base voltage of transistor Tr_2 , the PNP-NPN switch made up of Tr_2 and Tr_3 becomes conducting, thus discharging C_1 through R_7 . The thyristor now ignites and remains conducting until the process is repeated once more when the load current next passes through zero.

Fig. 9 shows that the conduction angle φ is deter-

mined by the charging current of C_1 , and therefore depends on R_2 and the base voltage of Tr_1 . This voltage is the difference between the reference voltage (the slider of R_3) and the d.c. voltage E_g provided by G , which is proportional to the motor speed. Control of the speed of motor M is thus obtained. The desired speed is set by means of potentiometer R_3 , while the accuracy with which this speed is maintained against a loading torque depends on the loop gain of the control system. This gain can be simply adjusted by R_2 . The tachometer G is a brushless a.c. voltage generator directly coupled to the motor shaft and has a ferroxdure rotor and a five-phase stator winding. The a.c. voltage obtained is rectified by the circuit shown in *fig. 10*, yielding a d.c. output signal with very low ripple content.

Using this method, a control system for a two-phase induction motor with a four-pole stator has been designed which enabled the speed to be varied by a factor of 50.

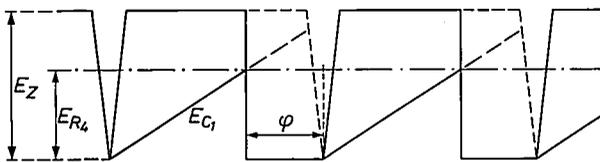


Fig. 9. The control of the conduction angle φ with the aid of the control signal. The saw-tooth voltage E_{C1} is started after the thyristor extinguishes (see *fig. 8*), and has a slope proportional to the actuating error signal (base of Tr_1). The thyristor is ignited when E_{C1} reaches the fixed level E_{R4} .

The performance as a tape recorder drive was examined on the tape drive of an EL 3501/02 studio recorder, the original synchronous motor of which was replaced by the controlled induction motor. The tape speed fluctuation was measured in three different ways: a) in a series of octave bands from 5 to 320 Hz, b) with a flat frequency characteristic from 0.4 to 300 Hz, c) with a frequency characteristic designed to compensate for the sensitivity of the human ear to fluctuations at various frequencies. The peak-to-peak speed variations measured are given in *Table I* in % of the rated speed. The fluctuations measured are well below the permissible peak-to-peak limit of 0.2% (measured with the weighting characteristic) for the standard professional tape speeds of 15 and $7\frac{1}{2}$ inch/s.

As the tape runs, the decrease in the diameter of the reel of tape on the wind-off side causes a corresponding increase in the tape tension as a result of the practically constant torque exerted by the motor driving the wind-off reel. This torque is necessary to keep the tape taut, the tape tension therefore loads the motor under control. The increase in the tape tension causes

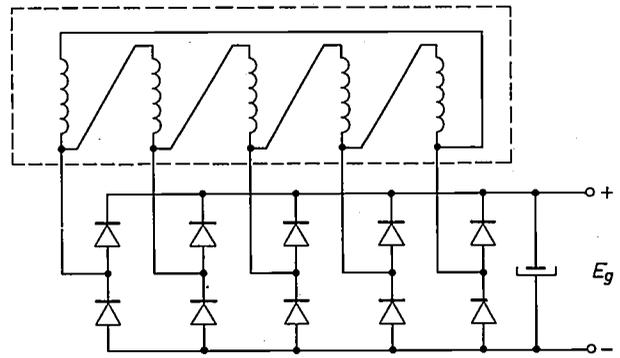


Fig. 10. Detail of the tachometer circuit. The brushless tachometer has a permanently magnetized ferroxdure rotor and a five-phase stator winding. After the stator voltages have been rectified by the rectifier bridge there is a d.c. voltage proportional to the speed, with a low ripple content.

an increase in the loading torque of the controlled induction motor, and, because of the finite loop gain of the speed control system, a slow decrease in the tape speed. At a rated speed of $7\frac{1}{2}$ inch/s, this fluctuation was found to be 0.2% over the entire reel. This is a well known effect in speed control systems in which the control signal is proportional to the speed error. It can be avoided by applying phase control instead of speed control.

Phase control can be obtained by deriving from the capstan a frequency f_0 proportional to the speed, e.g. by means of a perforated disc and a photodetector. Comparison of this frequency with a fixed reference frequency f_1 by means of a phase discriminator then provides an actuating error signal corresponding to the phase difference between f_1 and f_0 . If this phase error signal is fed to the base of Tr_1 in *fig. 8* via a phase-advance network, a stable control system is once more obtained. The tachometer can now be dispensed with, while a slow variation in the motor load has a negligible effect on the speed (at constant load the speed error becomes zero). When used in a tape recorder, this system has the further advantage that, by division of

Table I. Fluctuations in the tape speed (peak-to-peak values in % of the rated speed) for a tape recorder drive with the system of *fig. 8*.

frequency range	tape speed			
	15	$7\frac{1}{2}$	$3\frac{3}{4}$	$1\frac{7}{8}$ inch/s
motor speed	1000	500	250	125 r.p.m.
5 - 10 Hz	0.05	0.072	0.116	0.08
10 - 20	0.04	0.08	0.09	0.06
20 - 40	0.04	0.09	0.072	0.11
40 - 80	0.056	0.09	0.14	0.18
80 - 160	0.1	0.14	0.24	0.24
160 - 320	0.13	0.19	0.3	0.34
0.4-300	0.2	0.32	0.48	0.84
with weighting characteristic	0.06	0.1	0.2	0.4

f_1 or f_0 the tape speed can be accurately varied by a factor of 2^m , in accordance with the range of standard speeds used for tape recorders.

Fig. 11 is the block circuit diagram of such a drive made in this laboratory. The reference generator supplies the crystal-stabilized frequency f_1 to the phase discriminator. A perforated disc with 300 holes is mounted on the capstan and together with a photo-detector provides a voltage at the frequency

$$f_a = \frac{n}{60} 300 \text{ Hz.} \quad \dots \quad (14)$$

Dividing f_a by means of a frequency divider with a division factor of 2^m results in a frequency f_0 :

$$f_0 = \frac{1}{2^m} \frac{n}{60} 300 \text{ Hz.} \quad \dots \quad (15)$$

The output of the phase discriminator serves, via the stabilization network, as the actuating signal for a thyristor control circuit like that in fig. 8. The operation of the control system makes f_0 equal to f_1 . The tape

Table II. As Table I, but for the system of fig. 11 (phase control instead of speed control).

tape speed motor speed	30	15	7½	3¾	1¾	15/16	inch/s
frequency range	2000	1000	500	250	125	62.5	r.p.m.
0.4-300 Hz	0.14	0.14	0.18	0.22	0.42	0.6	
with weighting characteristic	0.05	0.07	0.12	0.14	0.3	0.3	

cuts were made up of standard circuit units (Icoma Circuit Blocks).

Where different tape recording installations with drives like that in fig. 11 are used, e.g. in a studio, the reference frequency f_1 may be taken from one common reference generator, in order to keep relative fluctuations in speed to a minimum. As the average tape speed is determined by the mean frequency of a quartz crystal, a system like that in fig. 11 gives better stability than those with synchronous motors, where the speed is governed by the frequency of the supply mains.

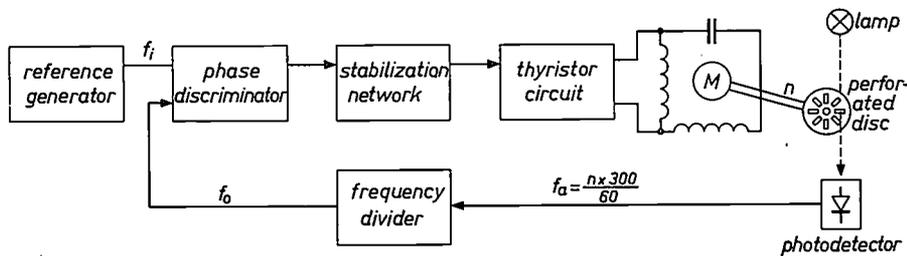


Fig. 11. Block diagram of a speed control system using a phase discriminator. The reference generator provides a quartz-stabilized reference frequency f_1 , which is compared by a phase discriminator with the divided shaft frequency $f_0 = 2^{-m} f_a$. The phase error signal controls the induction motor via a stabilization network and thyristor control circuit.

speed may then be found from the relationship:

$$v = \frac{2^m \pi D f_1}{300} \text{ inch/s,} \quad \dots \quad (16)$$

where D is the diameter, in inches, of the capstan transporting the tape.

This system, used with a two-pole motor, enabled the six standard tape speeds of 15/16 to 30 inch/s to be obtained on an EL 3501/02 tape drive system.

The fluctuations measured at various tape speeds are shown in Table II. For reasons of brevity, only those results corresponding to a flat frequency characteristic from 0.4 to 300 Hz and to the weighting characteristic are given. The figures shown are the measured peak-to-peak speed fluctuations in % of the rated tape speed. Comparison with the values in Table I shows that, particularly at low tape speeds, the speed variations are considerably reduced.

As in the simple system in fig. 8, the drive shown in fig. 11 is fully transistorized; the majority of the cir-

Driving a centrifugal spreader

Another application of the thyristor-controlled induction motor developed in this laboratory is in the centrifuge shown in fig. 12 for spreading photosensitive emulsions on to flat glass plates. The plates are used for making photomasks for the manufacture of integrated circuits [9].

This machine is based on the centrifugal principle employed in the printing industry for applying uniform coatings of ink. The fluid is poured on to the centre of a flat rotating plate, and centrifugal force causes the fluid to flow evenly over the entire surface.

Irregularities in the spread layer may be caused by inhomogeneities in the fluid, pouring off-centre or irregular rotation of the plate during pouring and centrifuging. It is difficult, especially at relatively slow speed (up to 20 r.p.m.), to make the plate rotate evenly.

[8] M. van Tol, Loop gain and stability of simple control systems, Philips tech. Rev. 23, 151-155, 1961/62 (No. 5).

[9] See A. Schmitz, Solid circuits, Philips tech. Rev. 27, 192-199, 1966 (No. 7).

With the a.c. collector motor originally employed, jerky movement was obtained at low speeds because of brush friction and the rotor slots. These effects are prevented by the use of an induction motor with a solid iron rotor. The centrifugal spreader shown in fig. 12 has a four-pole motor of this type directly driving the 30 cm diameter turntable. The speed of the motor is controlled

elements is required in relation to the number of control facilities provided. Spreading takes place at a fairly low speed N_1 (selected by S_6 : 20-100 r.p.m.; S_{2a} and S_{2b} in position B). Once the spreading process is completed, the turntable is uniformly accelerated to the final speed N_2 (selected by S_5 : 100-1000 r.p.m.) by switching S_4 from N_1 to N_2 . This acceleration is achieved by



Fig. 12. Centrifugal spreader which operates either at constant speed, or at constant acceleration over a selected range. The turntable has lugs to hold the square glass plates (in three sizes). All controls are to be found on the sloping front panel. The control knob in the bottom right-hand corner (*Temp*) controls the heating element in the cover, used for drying the spread-out layer of emulsion.

by a speed control system based on the principle shown in fig. 8, with the addition of a circuit by which the speed may be increased with constant acceleration through a selected centrifuging speed range.

The complete circuit diagram is given in fig. 13, and shows that only a relatively small number of compo-

feeding a saw-tooth control voltage to the base of the control transistor Tr_1 from a Miller integrator (Tr_4 , Tr_5 , Tr_6) via tachogenerator G . About one minute is needed to accelerate from 20 to 1000 r.p.m. Apart from programmed speed control, continuous manual speed control is also possible with the aid of R_8 . The set

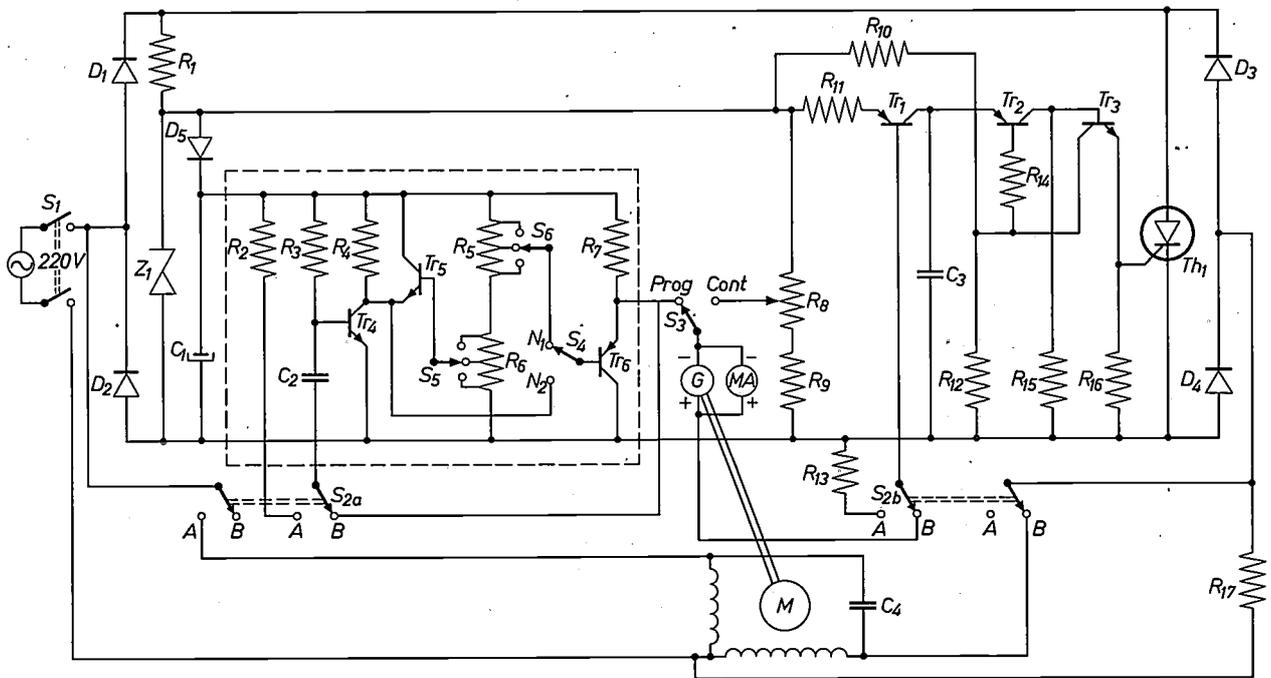


Fig. 13. Complete circuit diagram of the controlled centrifugal spreader. The circuit with transistors Tr_1, Tr_2, Tr_3 is identical to that in fig. 8. The speed can be programme-controlled with the aid of the Miller integrator circuit Tr_4, Tr_5, Tr_6 (shown inside the dotted lines). This enables the speed to be increased at a constant rate of acceleration from N_1 (20-100 r.p.m.) to N_2 (100-1000 r.p.m.).

speed can at all times be read off from the meter MA , connected in parallel with G and calibrated in r.p.m.

The turntable is electrically braked by the reversal of the motor rotating field (S_{2a} and S_{2b} in position A).

It was found possible, with the above-described apparatus, to spread photosensitive emulsions to a thickness down to a few microns with a tolerance of $\pm 4\%$.

Reversible control with two thyristors

Reversible control of small induction motors is often required in servosystems for instrumentation, such as temperature recorders or positional control systems for reversibly controlling the position of a valve or the motion of a shaft.

Fig. 14 shows the principle of a method by which it is possible to obtain reversing control of a two-phase capacitor motor. A second thyristor is used in this case. The latter, Th_2 , is also included in the d.c. branch of a bridge rectifier connecting the a.c. supply to the other side of the capacitor. This second bridge enables the direction of rotation of the stator rotating field, and hence the torque, to be reversed. Because of the voltage difference between the thyristor cathodes, the ignition pulses from the generator PG are applied to Th_1 and Th_2 through small isolating transformers (Tf_1 and Tf_2). For a system using gas-filled thyatrons instead of thyristors, a separate heater supply would be required as well.

The following classes of operation may be distin-

guished for Th_1 and Th_2 (fig. 15), using the conventional terms for push-pull amplifiers.

1) Class AB. The ignition pulses from the pulse generator PG are supplied to both thyristors if the actuating error signal e , again derived from the deviation of the controlled quantity from its reference value, changes

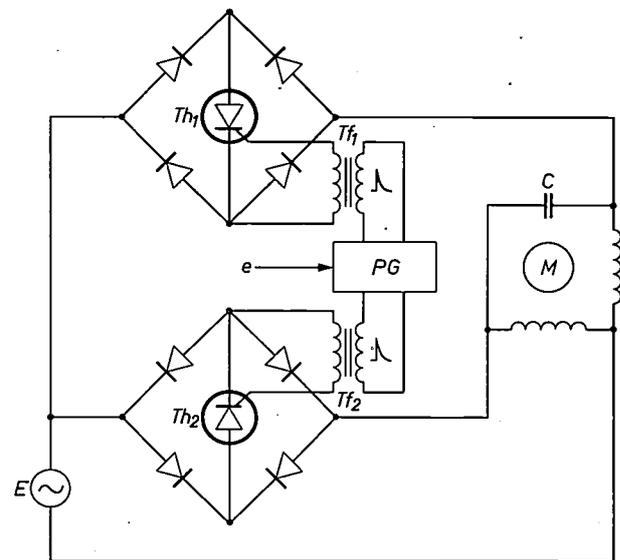


Fig. 14. Principle of a reversible control for a two-phase capacitor motor. The second thyristor Th_2 enables the direction of rotation of the stator rotating field — and hence the torque — to be reversed. Both thyristors are ignited from the same pulse generator (PG), connected by isolating transformers Tf_1 and Tf_2 . The instants at which ignition occurs are controlled by the actuating error signal e .

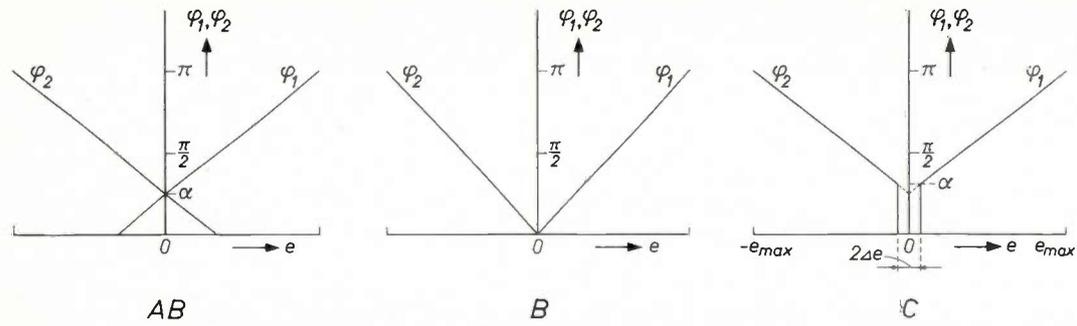


Fig. 15. The various settings for reversible control using two thyristors. The variation of the conduction angles φ_1, φ_2 of the two thyristors as a function of the control signal e is shown for operation in class AB, B and C (using the same classification as in push-pull amplifiers). In class AB, both transistors are operational at small values of $|e|$.

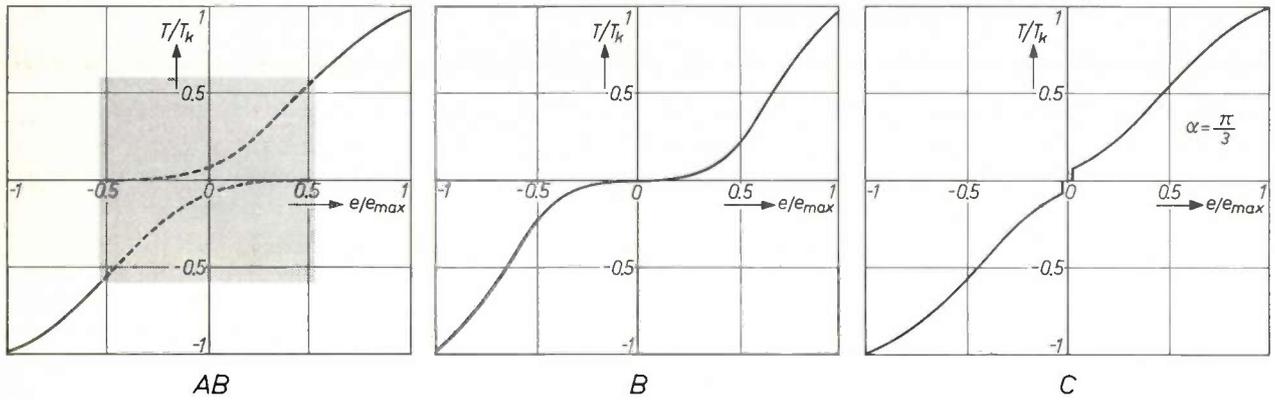


Fig. 16. The resulting control characteristic for the reversible control system of fig. 15, for various types of operation. Class AB operation gives a greatly distorted control characteristic inside the shaded area because of the simultaneous operation of both thyristors for small control signals. In class B, this no longer occurs, but there is a flat region around the origin (see also fig. 5). Class C gives better linearity, with a slight discontinuity in the origin. The characteristic shown is for a starting angle of $\alpha = \pi/3$.

about zero within a restricted range. Outside this range either Th_1 or Th_2 is operational, according to the polarity of e . The conduction angles φ_1 and φ_2 are linear functions of e and may vary between 0 and π .

2) Class B. According to the polarity of e , the ignition pulses are supplied to either Th_1 or Th_2 . The angles φ_1 and φ_2 are linear functions of e and vary between 0 and π . When $e = 0$, we have $\varphi_1, \varphi_2 = 0$.

3) Class C. Th_1 or Th_2 comes into operation only when $|e|$ exceeds a certain threshold value Δe (about 1% of the maximum value e_{max}), the conduction of Th_1 or Th_2 being initiated with an initial angle α . φ_1 and φ_2 are again linear functions of e and now vary between α and π .

The effect of these different classes of operation was examined with the aid of a four-pole motor with a solid iron rotor ($s_k \approx 1$) in a circuit like that of fig. 14. The resultant control characteristics can be seen in fig. 16. With class AB operation a highly distorted and discontinuous control characteristic occurs within the shaded area. This is caused by the simultaneous operation of Th_1 and Th_2 for small control signals, causing capacitor C to be short-circuited. Class B operation does give a continuous monotonic control character-

istic, but with a very flat section about the origin. This is undesirable for reversing control, as it causes the loop gain, and hence the accuracy of the control system, to be least precisely in the operating range, i.e. about $e = 0$. The characteristic shown in fig. 16c for class C operation exhibits better linearity, and is therefore best suited to a reversing control, in spite of a slight discontinuity at the origin.

The dynamic characteristics of this reversible control system (in class C operation) were examined with the arrangement shown diagrammatically in fig. 17. The step-function response of the torque (T) was measured with dynamic torque meter

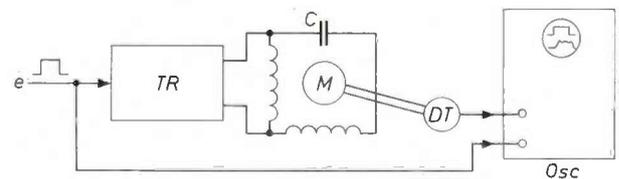


Fig. 17. Diagram of an arrangement for measuring the dynamic starting torque. The control unit TR , which uses thyristors, is controlled by a stepped or sinusoidal control signal. The corresponding response of the torque T is displayed on the oscilloscope by means of the dynamic torque meter DT , together with the control signal used.

DT and displayed on the oscilloscope. The frequency response was examined in the same way with the aid of a sinusoidal control signal.

The response to a rectangular pulse signal is shown in fig. 18a. The ripple in the torque is of relatively high frequency and disappears when a 0-80 Hz low-pass filter is used with the display. This makes the rise and decay of the mean torque easier to see (fig. 18b). However, the rise time of the filter (fig. 18c) must be deducted

The transfer function of the system when used in a reversible speed control was examined using the circuit in fig. 20. Negative feedback of the tachometer voltage E_o , which is proportional to the speed, gives a speed control with the complex transfer function^[8] (closed-loop control system):

$$E_o/E_i = A. \dots \dots \dots (17)$$

Here E_i is the reference voltage corresponding to the

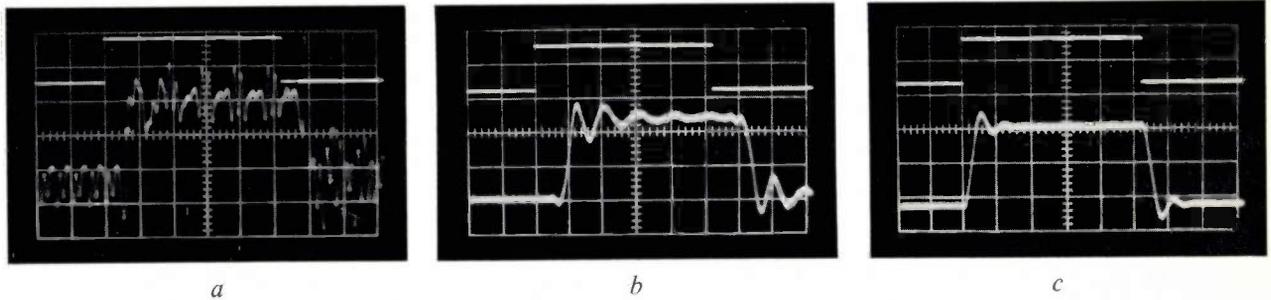


Fig. 18. Oscillograms of the torque (with locked shaft) with a square-wave control signal, the torque varying in steps from $-0.75 T_k$ to $+0.75 T_k$. The time scale is 20 ms per division. a) Variation of the torque signal, the superimposed ripple originating from the switching on of the thyristors. b) The same signal but passed through a 0-80 Hz low-pass filter to show the average torque. c) The step-function response of the filter alone, shown so that the effect of this filter can be assessed.

in finding the rise time of the torque from fig. 18b. The time required to attain 90% of the final value of the torque is found to be less than 20 ms.

The response to a sinusoidal control signal of 4, 10 or 23 Hz when the filter is used with the display is shown in fig. 19. As in fig. 18, the control signal itself is also shown for purposes of comparison. After the filter phase-shift has been subtracted, the phase-shift between control signal and torque is found to be:

f (Hz)	φ (degrees)
4	22
10	37
23	86

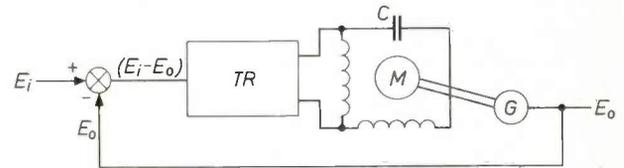


Fig. 20. Diagram of the reversible thyristor control (TR) applied in a speed control system. Negative feedback of the tachometer signal E_o , which is proportional to the speed, gives a speed control (cf. fig. 8) with the transfer function $A = E_o/E_i$.

desired speed and occurs as an input signal, while E_o is the output signal corresponding to the shaft speed. The corresponding open-loop transfer function G follows from the transfer function defined in (17) for the closed-loop control system:

$$G = \frac{E_o}{E_i - E_o} = \frac{A}{1 - A} \dots \dots \dots (18)$$

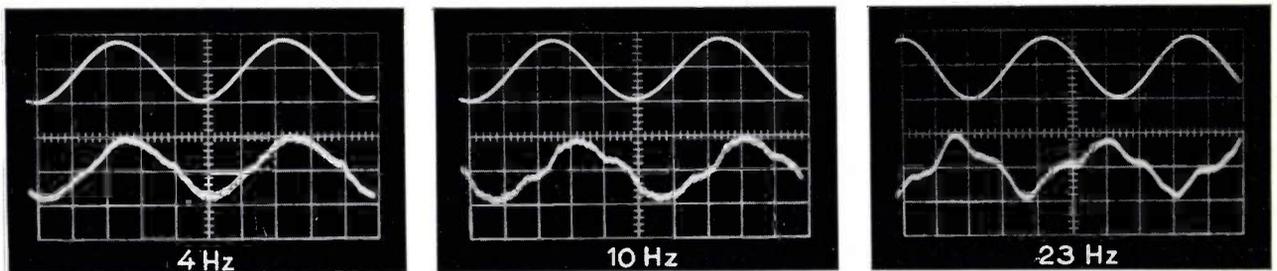


Fig. 19. The behaviour of the starting torque when a sinusoidal control signal with a frequency of 4 Hz, 10 Hz or 23 Hz is employed. The signal proportional to the torque is displayed after passing through an 80 Hz low-pass filter.

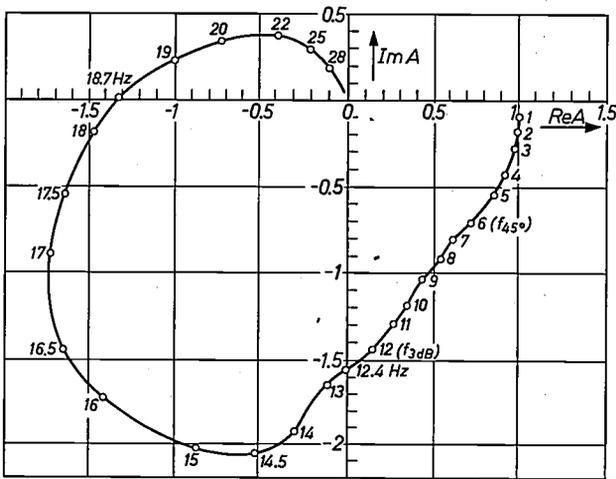


Fig. 21. Polar diagram of the transfer function A (with a closed loop) at 1000 r.p.m. for the speed control shown in fig. 20. The 3 dB frequency f_{3dB} of the transfer function is 12 Hz.

These functions were measured for various frequencies at a shaft speed of 1000 r.p.m. Fig. 21 is the polar diagram of A . The frequency f_{3dB} , where A differs by 3 dB from its value at $f = 0$, is found to be 12 Hz, while the phase angle of A is 45° at $f_{45} = 6$ Hz. These figures give an idea of the maximum allowable rate of variation of E_i , for E_o to be still a good approximation of E_i . Both frequencies can be increased by the use of a suitable stabilization network (as in fig. 11).

Such a network can be calculated from a knowledge of the open-loop transfer function (G) by well-known

methods [8]. Fig. 22 shows the amplitude and phase of G as a function of the frequency (Bode diagram). The frequency at which $|G|$ cuts the 0 dB line (the "cross-over frequency") is a measure of the dynamic characteristics of the system. This frequency is 8.5 Hz according to fig. 22.

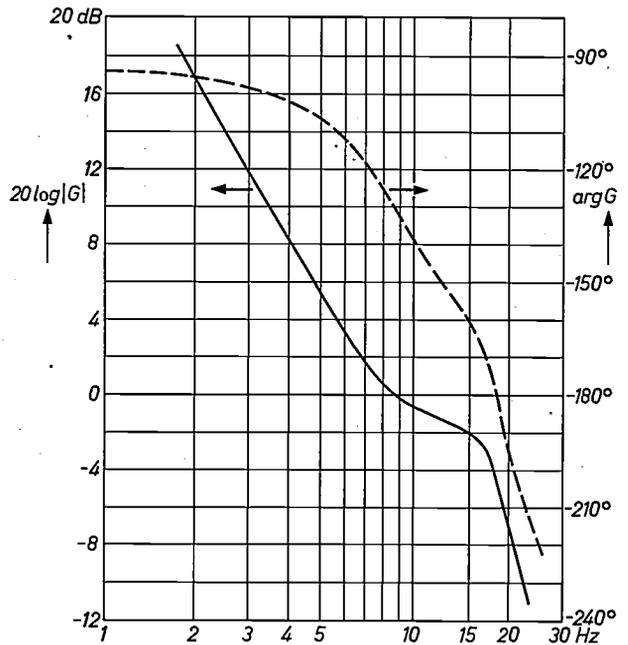


Fig. 22. Amplitude and phase response curve of the open-loop transfer function (G).

Summary. The torque or speed of an induction motor can be controlled in a simple manner with the aid of thyristors. By using the control method based on variation of the stator voltage, particularly simple control systems can be obtained requiring only one thyristor. At low speed (high slip) the efficiency becomes low, and the resulting increase of rotor dissipation limits this control method to motors rated below about 100 W. The method can be used for higher-power motors if forced cooling is provided. In the

control systems dealt with, use is made of two-phase capacitor motors with a fairly high rotor resistance (critical slip nearly 1). These motors have been used in a tape recorder drive with a speed range of 1 : 32 and in a laboratory centrifuge with a speed range of 1 : 50. In both cases, the use of the controlled induction motor gave better results than the synchronous or universal motors formerly used. Reversible control was obtained by using a second thyristor. Class C operation gave the best results

Crystal chemistry and magnetism of oxide materials

- I. Principles and application of crystal field theory
- II. Magnetic compounds with spinel structure

The relation between crystal structure, chemical composition and physical properties occupies the attention of many chemists and physicists in our laboratories. The discovery of materials of such technological importance as the ferrites, for example, was a direct result of investigations in this field. The article below illustrates the method of investigation, making use of new theoretical insight. Part I deals with crystal field theory, which is today the basis for the theoretical treatments of compounds of the transition elements. Part II deals with various properties of ferrites with spinel structure, and shows how materials can be given specific physical properties by an appropriate choice of chemical composition.

I. Principles and application of crystal field theory

P. F. Bongers

Certain compounds of transition metal ions, that is to say ions whose d shell is only partly filled with electrons, possess interesting magnetic properties. Well-known cases in point are the ferrites. These magnetic oxides, which include the various kinds of ferroxcube and ferroxidure, have found a wide field of application as magnetic materials [1] [2]. The usefulness of these substances is determined by factors such as the magnitude of the magnetic moment per unit volume and by the stiffness with which this moment is bound to a particular crystal direction. These properties depend not only on the external shape, grain size, etc., but also to a considerable extent on the type and valence of the magnetic ions of which the compound is composed and on the stacking of the ions: in other words, on the crystal structure.

A good general understanding of the crystal structure of many oxides can be obtained if it is assumed that the bonding in these substances is ionic, i.e. that the arrangement of the charged ions in the crystal is the one which gives the minimum energy for the system of attraction and repulsion forces between the ions. In a number of transition element oxides, however, the crystal structure deviates from that of corresponding compounds that do not contain transition elements. For example, ZnMn_2O_4 has a tetragonal structure, with an axial ratio of 1.14. This crystal structure can

be thought of as a spontaneous deformation of the crystal structure of ZnAl_2O_4 .

The effect of the crystal structure on the magnetic properties is neatly demonstrated by the two modifications α and γ of iron trioxide: $\alpha\text{-Fe}_2\text{O}_3$ has a hexagonal structure and is not magnetic, whereas $\gamma\text{-Fe}_2\text{O}_3$ has cubic symmetry and is magnetic below 234 °C.

In investigations into the relationship between chemical composition, crystal structure and magnetic properties, *crystal field theory* has proved to be a particularly useful tool. This theory starts from the electronic state of free ions (in vacuo) and deals with the effect of the surrounding ions (ligands) on the electronic state of a central ion. The surrounding ions are considered here as point charges giving an electric field at the position of the central ion. Purely ionic bonding is assumed, i.e. the electrons in the complex are each localized around one nucleus. This is not entirely correct. Methods of describing the state of the electrons using a combination of states of both the positive and the negative ion are generally extremely complicated, however. In these theories, which are usually grouped under the name "ligand-field" theory, it is difficult to introduce simplifications which are applicable to large numbers of compounds. What in fact makes crystal field theory so useful is the relatively simple method of

Dr. P. F. Bongers is with Philips Research Laboratories, Eindhoven.

[1] J. J. Went and E. W. Gorter, Philips tech. Rev. 13, 181, 1951/52.

[2] J. J. Went, G. W. Rathenau, E. W. Gorter and G. W. van Oosterhout, Philips tech. Rev. 13, 194, 1951/52.

description employed; it can be used to give a semi-quantitative explanation for the magnetic, optical and crystallographic properties of numerous compounds [3].

This article examines the manner in which the crystal field of an octahedral and tetrahedral environment formed of negative ions — an arrangement frequently found in oxides — affects the electronic states and the magnetic moment of the various transition metal ions. Some examples are then discussed of spontaneous lattice deformations which arise because the ions are not truly spherical. Finally, consideration is given to the magnetic interaction between neighbouring ions.

One *d* electron

The ions of the first series of transition elements — the iron group — have an incompletely filled 3*d* shell ($n = 3, l = 2$). For the trivalent ions, the number of 3*d* electrons increases with ascending atomic number from zero for Sc^{3+} to 10 for Ga^{3+} . The Ti^{3+} ion has one 3*d* electron ($3d^1$). There are five 3*d* wave functions, and these all correspond to the same energy. The free ion has therefore a fivefold degenerate level as its ground state. If we now introduce a perturbation due to six negative point charges arranged in a regular octahedron (see *fig. 1*), we shall find that all the orbitals no longer correspond to the same energy. It is found that the two wave functions of the form

$$(x^2 - y^2) f(r) \text{ and } (2z^2 - x^2 - y^2) f(r),$$

indicated by $d_{x^2-y^2}$ and d_{z^2} respectively, correspond to

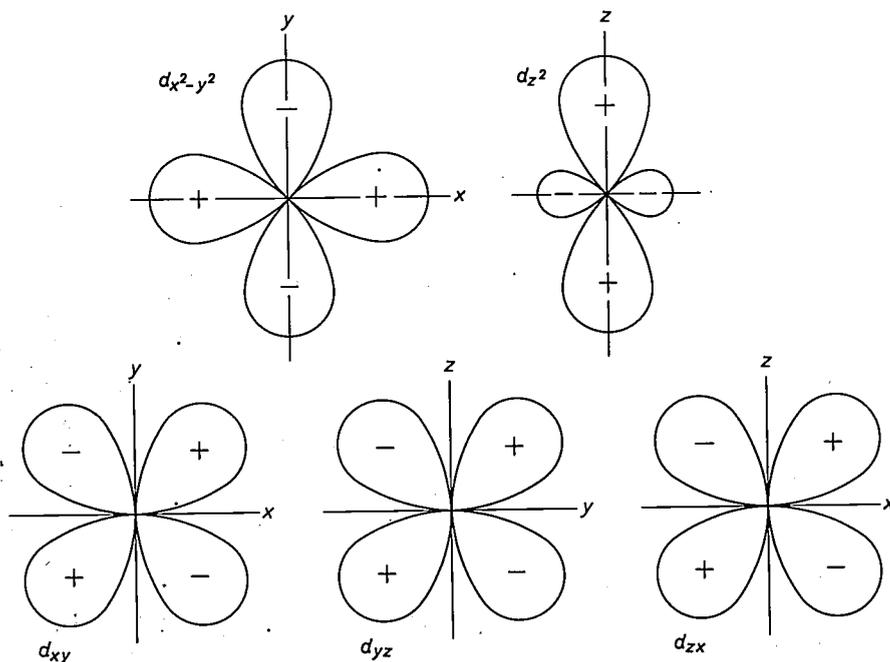


Fig. 2. The *d* orbitals of an ion in a cubic crystal field (diagrammatic). The sign of the wave function is indicated in the various regions. The charge distribution of the electron is proportional to the square of the wave function. The contours are lines of constant wave-function amplitude.

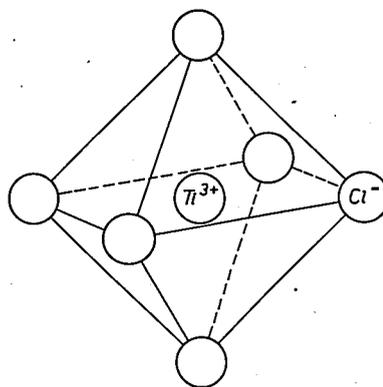


Fig. 1. A Ti^{3+} ion ($3d^1$) at an octahedral site formed by six negative ions.

an energy different from that of the three wave functions

$$xy f(r), yz f(r) \text{ and } zx f(r),$$

indicated by d_{xy} , d_{yz} and d_{zx} respectively. Here $f(r)$ depends on the distance to the nucleus; x , y and z are the direction cosines ($x^2 + y^2 + z^2 = 1$). The charge distribution of the electron, when it is situated in one of these orbitals, is shown diagrammatically in *fig. 2*. It can readily be seen that the ion will have the same energy if the electron is in d_{xy} , d_{yz} or d_{zx} . This energy will be lower than when the electron is in $d_{x^2-y^2}$ or d_{z^2} . The reason for this is that the charge cloud of these orbitals is directed towards the negative ions, whereas the charge cloud of $d_{x^2-y^2}$, d_{yz} or d_{zx} points in the directions between the negative ions. The fivefold degeneracy of the ground state therefore disappears

and it splits up into a doublet e_g and a triplet t_{2g} (*fig. 3a*). The energy separation between the doublet and the triplet, referred to as crystal field splitting, is called $10 Dq$. Its value is roughly 10^4 cm^{-1} for bivalent and trivalent ions in oxides and hydrates.

If the ion is not surrounded by six ions arranged in an octahedron but by four ions forming a tetrahedron, the splitting is reversed (*fig. 3b*). The splitting is smaller because there are now only four negative charges instead of six. *Fig. 4* shows clearly that the crystal field splitting of a cube differs in sign from that of an octahedron. The centre diagram shows an octahedron

Fig. 3. Splitting of the d level as a result of the symmetry of the crystal field when there is one electron in the $3d$ shell. It can be seen on the left of both figures that the free electron in each of the five d orbitals has the same energy. On the right the level is shown splitting into a doublet and a triplet, as a result of a crystal field with *a*) octahedral symmetry and *b*) tetrahedral symmetry. The crystal-field splitting, the energy separation between doublet and triplet, is $10 Dq$.

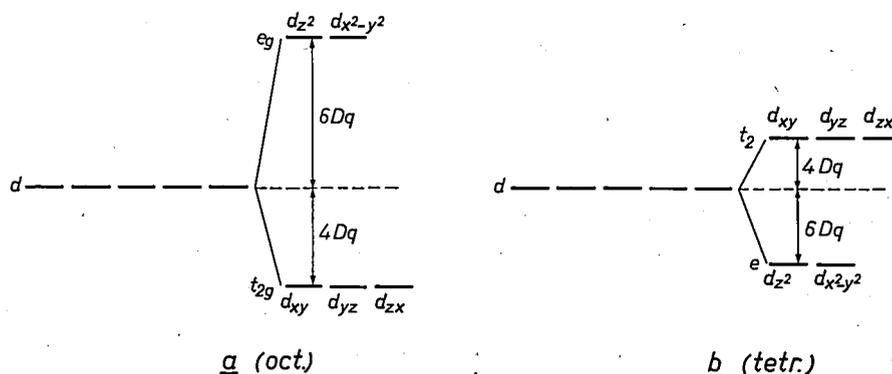
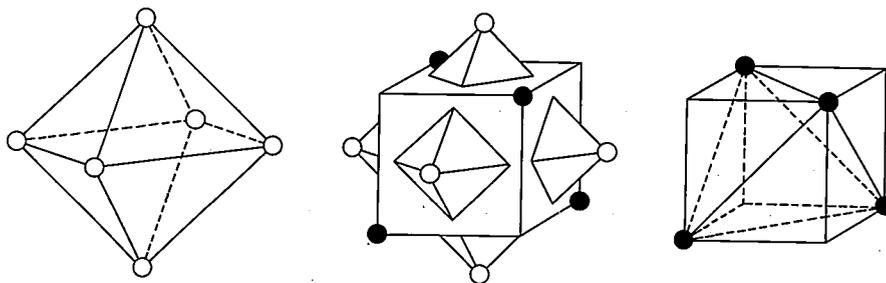


Fig. 4. The difference between the three-dimensional charge distributions in fourfold and in sixfold co-ordination. On the left is drawn an octahedron and on the right a cube with a tetrahedron inside. The drawing in the middle shows a cube and octahedron combined.



and a cube combined. Seen from the origin the corners of the cube are exactly behind the centre points of the sides of the octahedron, and vice versa. From the point of view of charge distribution, the completely occupied cube is therefore the "negative" of the octahedron. A tetrahedron is obtained from this cube by occupying only half the corners with ions.

More than one d electron

If there are more electrons in the $3d$ shell, these will repel one another. This repulsion is at a minimum when the spins of the electrons are parallel. In this case the electrons cannot be at the same place at the same time, as they would then be characterized by the same state (the same orbital and spin quantum number), and this is not permitted because of the Pauli exclusion principle. The electrostatic repulsion between the electrons therefore leads to Hund's first rule, which states that for free ions the state in which as far as possible the spins are oriented parallel is the state with the lowest energy.

If the crystal field (V_{CF}) is small with respect to the electron repulsion, which is also called the Hund energy E_H , then little change in this will be brought about. We now obtain the electron configuration of the various ions, as a first approximation, by filling up the levels, beginning with the low-energy levels, with one electron each until all of them are half filled. This is the case, for example, with the ferric ion Fe^{3+} . The five $3d$ orbits here are half filled with electrons; all with their magnetic spin moments parallel: $S = 5/2$.

The Fe^{2+} ion has one electron more. This "sixth" $3d$ electron is situated in a t_{2g} orbital, now with its spin moment antiparallel to the other spin moments, so that the configuration is $t_{2g}^4 e_g^2$ with $S = 2$ (fig. 5).

If, however, the crystal field is very strong, stronger

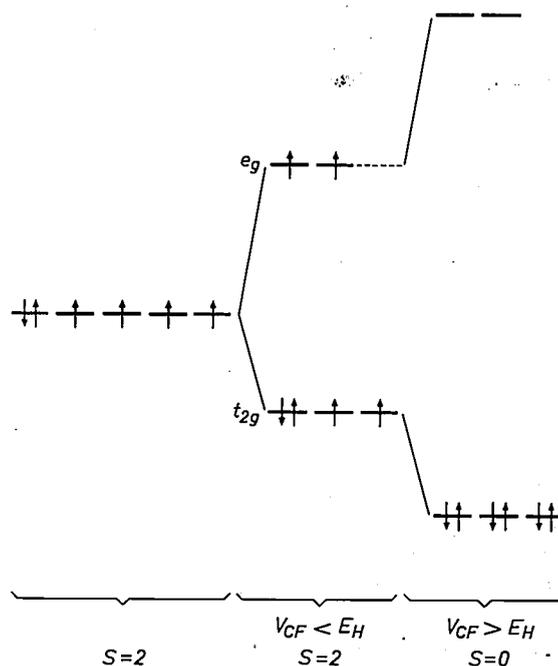


Fig. 5. Diagrammatic representation of the electron configuration of an ion with six d electrons, e.g. Fe^{2+} , on the left for the free ion and on the right for the ion in an octahedral environment. If the crystal field V_{CF} is small with respect to the electrostatic repulsion of the electrons (Hund energy E_H), then the occupation of the d orbitals is to a first approximation equal to that of the free ion. The total spin moment is in both cases $S = 2$. For the case where V_{CF} is greater than E_H only the t_{2g} orbitals are occupied. The moment of the ion is then $S = 0$.

[9] L. E. Orgel, An introduction to transition-metal chemistry; Ligand-field theory; Methuen, London 1960.

than the electrostatic repulsion between the electrons that leads to the parallel orientation of the moments ($V_{CF} \gg E_H$) then the t_{2g} orbitals will first be completely filled with electrons and the remaining electrons will fill the e_g orbitals. This is the case, for example, with the Fe^{2+} ions in $K_4Fe(CN)_6$; the electron configuration here is $t_{2g}^6e_g^0$ (fig. 5). The ferrous ions then have zero magnetic moment. The ions are said to be in a low spin state. Another example is given by the Co^{3+} ion, also with d^6 configuration. In the compound $LiCo^{3+}O_2$ the cobalt ion is in the low spin state, with $S = 0$. In the compound K_3CoF_6 however, in which the crystal field is weaker, because the fluorine ions, unlike the oxygen ions, have only a monovalent negative charge, the Co^{3+} ions have a moment $S = 2$.

Crystal field stabilization

As a result of the d level splitting due to the crystal field the various ions undergo an energy stabilization, called crystal field stabilization, which depends upon the number of d electrons and upon the symmetry of the crystal field. This can be made clear with the aid of some examples. The electron distribution of Cr^{3+} in an octahedron is $t_{2g}^3e_g^0$. The energy gain resulting from the fact that the three electrons are in the low-energy t_{2g} orbitals is $3 \times 4Dq_{oct} = 12Dq_{oct}$ (cf. fig. 3). For Mn^{3+} in an octahedron ($3d^4$; $t_{2g}^3e_g^1$) the stabilization is $(3 \times 4 - 1 \times 6)Dq_{oct} = 6Dq_{oct}$. For this same ion in a tetrahedron the stabilization is $4Dq_{tet}$ (see Table I). If we now know the value of Dq both for octahedral and tetrahedral surroundings, it is possible to determine the stabilization energy for both co-ordinations. Table II gives the crystal field stabilization values for various transition metal ions with fourfold and sixfold co-ordination in oxides. The Dq values were determined from the absorption spectra of these ions in these two types of co-ordination. The third column in Table II gives the difference between the stabilization energy for these ions in octahedral and tetrahedral co-ordination. This difference may be regarded as a measure of the preference for octahedral co-ordination.

The transition metal ions do not all have the same ionic radius. Since the size of the ions also affects the stability of a compound, it is not easy to verify the correctness of the octahedral preference. The Cr^{3+} ion shows the strongest preference, the Fe^{3+} ion has no preference. No oxides of trivalent chromium are known in which the Cr^{3+} ion is located on the tetrahedral sites of the oxygen lattice, but this is indeed the case for Fe^{3+} ion in many oxides, for example in nearly all $MeFe_2O_4$ ferrites with spinel structure. A detailed analysis of the octahedral preference in the spinel structure is given in part II of this article.

Table I. Electron configuration and magnitude of the total spin moment S of ions with n electrons in the d shell when these ions are situated in an octahedral or in a tetrahedral crystal field. The ion takes a different configuration depending on whether the crystal field V_{CF} is stronger or weaker than the electrostatic repulsion between the electrons (the Hund energy E_H).

n	high spin ($V_{CF} < E_H$)					low spin ($V_{CF} > E_H$)				
	oct.		tetr.		S	oct.		tetr.		S
	t_{2g}	e_g	e	t_2		t_{2g}	e_g	e	t_2	
1	1		1		$\frac{1}{2}$	1		$\frac{1}{2}$		$\frac{1}{2}$
2	2		2		1	2		1	2	1
3	3		2	1	$\frac{3}{2}$	3		$\frac{3}{2}$	3	$\frac{1}{2}$
4	3	1	2	2	2	4		1	4	0
5	3	2	2	3	$\frac{5}{2}$	5		$\frac{1}{2}$	4	$\frac{1}{2}$
6	4	2	3	3	2	6		0	4	2
7	5	2	4	3	$\frac{3}{2}$	6	1	$\frac{1}{2}$	4	3
8	6	2	4	4	2	6	2	1	4	4
9	6	3	4	5	$\frac{1}{2}$	6	3	$\frac{1}{2}$	4	5
10	6	4	4	6	0	6	4	0	4	6

Table II. Crystal-field stabilization (in kcal/mole) for transition metal ions with octahedral and with tetrahedral environment by oxygen ions.

	oct.	tetr.	"preference" for oct.
Mn^{2+}	0	0	0
Fe^{2+}	12	8	4
Co^{2+}	22	15	7
Ni^{2+}	29	9	20
Cu^{2+}	21	6	15
Zn^{2+}	0	0	0
Ti^{3+}	21	14	7
V^{3+}	38	25	13
Cr^{3+}	54	16	38
Mn^{3+}	32	10	22
Fe^{3+}	0	0	0

Crystal structure deformations

Jahn-Teller effect

For certain electron configurations additional crystal-field stabilization can be obtained by deforming the environment of the ion. Fig. 6a shows the charge cloud of the orbitals $d_{x^2-y^2}$ and d_{z^2} of the doublet e_g . Elongation of the octahedron in the z direction causes the electrostatic repulsion energy between an electron in d_{z^2} and the two negative ions on the z axis to decrease, and compressing it in the x and y directions (so as to keep the volume the same) causes an increase in the repulsion energy between an electron in $d_{x^2-y^2}$ and the four ligands in the xy plane. The e_g level thus splits as indicated in fig. 6b. When there is an odd number of electrons at the e_g level, this deformation

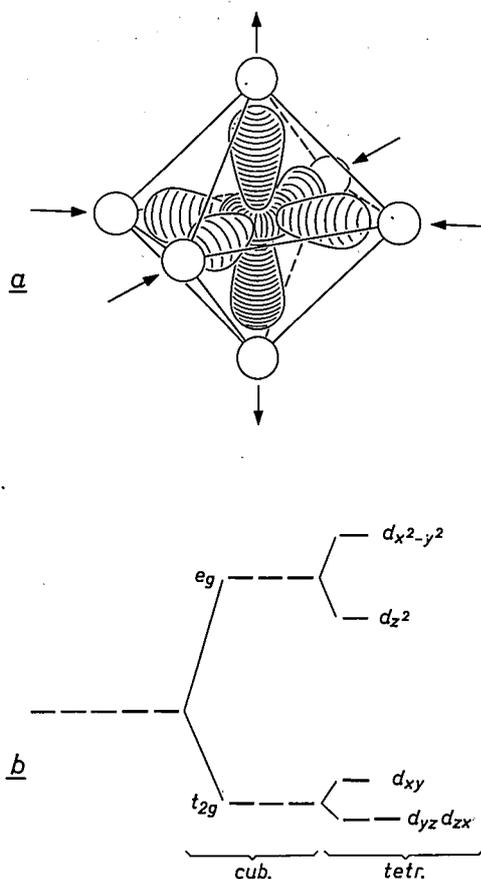


Fig. 6. The effect of deformation of the octahedron on the energy of electrons in the e_g orbitals. *a*) In the octahedron the charge cloud of an electron in the d_{z^2} orbital and in the $d_{x^2-y^2}$ orbital is shown. Owing to deformation of the octahedron in the direction of the arrows, an electron in d_{z^2} acquires a lower energy and an electron in $d_{x^2-y^2}$ a higher energy. *b*) The resultant splitting of the e_g doublet is indicated in the figure, as well as the accompanying, but smaller, splitting of the t_{2g} triplet.

gives rise to a net energy gain. A deformation of this nature is known as a Jahn-Teller effect. Generally it can be said that a "molecule" or complex with a ground state which is orbitally degenerate always has one direction along which it can be deformed and by which the energy is lowered. The Jahn-Teller effect removes the orbital degeneracy of the ground state. The splitting of the t_{2g} triplet is smaller because the t orbitals have a greater separation from the negative ions (fig. 6*b*). When all octahedra in a crystal lattice are randomly deformed in the x , y or z directions, strains arise in the lattice. Minimum energy will be achieved if all octahedra are deformed in the same direction. This can cause a change in the crystal structure; if for example this was initially cubic, it can become tetragonal as a result of a Jahn-Teller effect.

Ions with d^4 and d^9 configuration at an octahedral site have an odd number of e_g electrons. The structure of many compounds of Cr^{2+} and Mn^{3+} (d^4) and Cu^{2+} (d^9) is deformed by a Jahn-Teller effect. Table III gives some examples, for oxides with spinel structure. The positive ions in the spinel structure are located

Table III. Compounds with spinel structure. The crystal lattice of substances which contain Mn^{3+} or Cu^{2+} at octahedral sites is deformed tetragonally as a result of the Jahn-Teller effect of these ions. Ni^{2+} ions at tetrahedral sites give a Jahn-Teller effect with $c/a > 1$, while Cu^{2+} ions in this co-ordination give rise to a deformation with $c/a < 1$.

	symmetry	c/a	Jahn-Teller ion and its electronic state	
$\text{Zn}[\text{Al}_2]\text{O}_4$	cubic	1	—	
$\text{Zn}[\text{Mn}_2]\text{O}_4$	tetragonal	1.14	Mn^{3+}	$t_{2g}^3 e_g^1$
$\text{Mg}[\text{Mn}_2]\text{O}_4$	tetragonal	1.15	Mn^{3+}	$t_{2g}^3 e_g^1$
$\text{Fe}[\text{CuFe}]\text{O}_4$	tetragonal	1.06	Cu^{2+}	$t_{2g}^6 e_g^3$
$\text{Mg}[\text{Cr}_2]\text{O}_4$	cubic	1	—	
$\text{Ni}[\text{Cr}_2]\text{O}_4$	tetragonal	1.025	Ni^{2+}	$e^4 t_2^4$
$\text{Cu}[\text{Cr}_2]\text{O}_4$	tetragonal	0.92	Cu^{2+}	$e^4 t_2^5$

both at the tetrahedral and the octahedral sites. The ions at octahedral sites are placed between square brackets. With the exception of $\text{Cu}[\text{Cr}_2]\text{O}_4$ and $\text{Ni}[\text{Cr}_2]\text{O}_4$, all the tetragonally deformed spinels contain Mn^{3+} or Cu^{2+} ions in sixfold co-ordination. The deformation of copper and nickel chromite is caused by a Jahn-Teller effect of the ions at tetrahedral sites. This may be understood as follows. The electron configurations of these ions are d^8 ; $e^4 t_2^4$ for Ni^{2+} , and d^9 ; $e^4 t_2^5$ for Cu^{2+} . In nickel one of the three t_2 orbitals is doubly occupied, in copper one is singly occupied. If we assume that this is the xy orbital, fig. 7*a* shows the

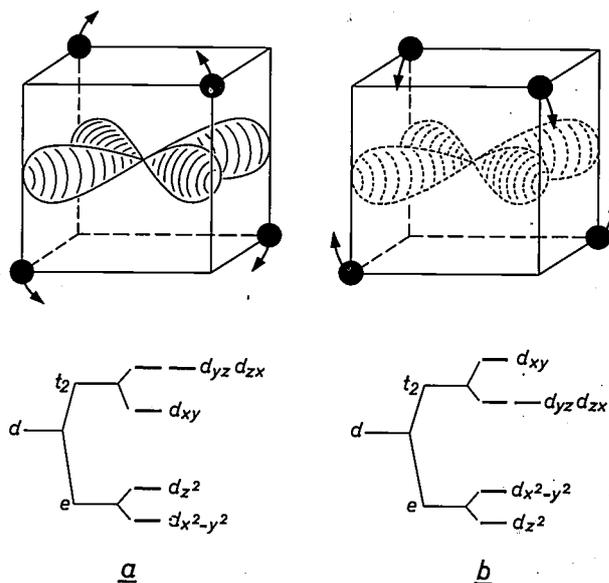


Fig. 7. Splitting of the t_2 level as a result of deformation of the tetrahedron.

- Jahn-Teller stabilization for one and four electrons respectively in the t_2 orbitals. If the negative ions move in the direction of the arrows, an electron in the d_{xy} orbital drawn will acquire a lower energy. This deformation will therefore occur spontaneously when d_{xy} contains more electrons than d_{yz} and than d_{zx} .
- Jahn-Teller stabilization for one "electron hole", that is to say five electrons in the t_2 orbitals. By deformation of the tetrahedron in the direction of the arrows an electron in d_{xy} acquires a higher energy than an electron in d_{yz} or in d_{zx} .

deformation that gives this orbital a lower energy, which is advantageous for the Ni^{2+} ion, and fig. 7b shows the deformation that gives d_{xy} a higher energy and hence d_{zx} and d_{yz} a lower one, as a result of which the Cu^{2+} ion is stabilized. The first deformation gives $c/a > 1$, as found for $\text{Ni}[\text{Cr}_2]\text{O}_4$, and the second $c/a < 1$, as is observed in the case of $\text{Cu}[\text{Cr}_2]\text{O}_4$ (Table III).

We see that the Jahn-Teller effect provides a qualitative explanation for many deformations. The splitting of the levels as a result of the Jahn-Teller deformation may be of the order of 10^3 cm^{-1} .

Spin-orbit coupling

Apart from the deformations due to the Jahn-Teller effect, there are a number of deformations due to spin-orbit coupling energy. This interaction between the spin moment and orbital moment of an ion can be understood as follows. Let us consider a nucleus around which an electron moves. Seen from the electron, the nucleus moves around the electron. The circular current that corresponds to the rotating nuclear charge gives rise to a magnetic field at the position of the electron. The magnetic interaction between the electron spin and this magnetic field is referred to as spin-orbit coupling. For an ion the coupling energy is λLS , where L is the total orbital moment, S the total spin moment and λ is a characteristic constant of the ion. The magnitude of this energy term is 10^2 - 10^3 cm^{-1} for $3d$ and 10^3 - 10^4 cm^{-1} for $4d$ and $5d$ transition metal ions.

It will be clear that in order to gain spin-orbit coupling energy an ion has to be in a state with $L \neq 0$. It has long been known, however, that the ions of the iron group in a crystal lattice have a small orbital moment. This is due precisely to the crystal field, as may be demonstrated with a two-dimensional example [4]. Fig. 8

shows a nucleus with one electron. The movement of the electron around the nucleus may be considered as a travelling wave. The energy is identical for left-hand and right-hand rotation. If a perturbation by four negative charges is now introduced, the electron will try as far as possible to evade the negative charge. The two states with a travelling wave change into two standing waves, one with low energy and one with high. There is now no longer any circular current and hence no orbital moment. The wave functions d_{xy} , $d_{x^2-y^2}$, etc., may be regarded as standing waves of this type. There is, however, an orbital moment if the ground state of the ion is degenerate, that is to say if the electron is in an orbit which is a combination of two standing waves (for example d_{zx} and d_{yz}).

The Jahn-Teller deformation has precisely the effect of removing the degeneracy of the ground state. There will therefore be competition between the Jahn-Teller effect and the spin-orbit coupling. In dealing with the Jahn-Teller effect we have assumed that the spin-orbit coupling energy is small with respect to the energy gained through the Jahn-Teller deformation. This is in fact the case for Mn^{3+} and Cu^{2+} in an octahedral environment. As we shall see later, this assumption is not true for Ni^{2+} and Cu^{2+} at tetrahedral sites.

In ferromagnetic and antiferromagnetic compounds the spin moments are ordered. Due to spin-orbit coupling the orbital moments in this case are also ordered, that is to say the electron orbits of all ions assume the same angular position with respect to the lattice. Instead of an ion possessing cubic symmetry averaged over time, the environment of the ion now "sees" an ion with axial symmetry. The lattice will adapt itself to this situation and become deformed. Kanamori has been able to attribute the deformation of CoO and FeO that occurs below the antiferromagnetic ordering temperature T_N (see below) to the spin-orbit coupling. This author [5] has also shown that the deformation due to spin-orbit coupling of an ion has the opposite sign to the Jahn-Teller deformation.

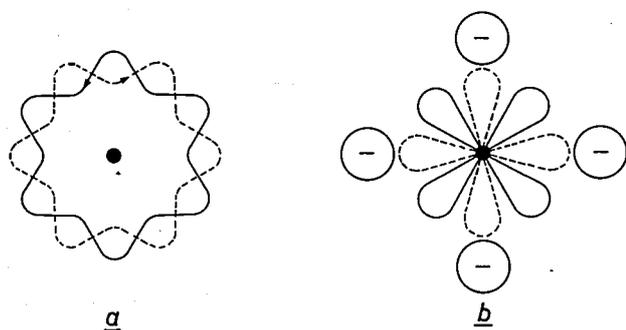


Fig. 8. Illustrating the effect of the crystal field on the orbital moment of a single electron [4].

- a) Free ion; the electron movement can be represented by two opposite travelling waves.
 b) Ion subject to the effect of four negative charges. The two travelling waves are changed into two standing waves, one with high, and one with low energy.

- [4] J. Smit and H. P. J. Wijn, *Ferrites*, Philips Technical Library, 1959.
 [5] J. Kanamori, *J. appl. Phys.* **31**, suppl., page 14 S, 1960.
 [6] R. J. Arnott, A. Wold and D. B. Rogers, *J. Phys. Chem. Solids* **25**, 161, 1964.
 [7] Arnott, Wold and Rogers [6] find an orthorhombic deformation in this region; see fig. 9 and 10. This was not observed by us at 120 °K. We did, however, observe X-ray reflections of a cubic phase in this region, whose intensity decreased upon further reduction of temperature. This is probably due to non-uniformity in the ion distribution or stresses in the powder grains.
 [8] If the orbital moment is zero, g is 2. Deviations of g from 2 are a measure of the orbital contribution to the magnetic moment $\mu = \beta g / \sqrt{S(S+1)}$ of the ion.

In the series of mixed crystals between $\text{Ni}[\text{Cr}_2]\text{O}_4$ and $\text{Fe}[\text{NiFe}]\text{O}_4$ the competition between spin-orbit coupling and the Jahn-Teller effect is clearly demonstrated [6]. These mixed crystals contain the ions Cr^{3+} (d^3) at octahedral sites, Ni^{2+} (d^8) and Fe^{3+} (d^5) at octahedral and tetrahedral sites. In a cubic structure only the Ni^{2+} ion with tetrahedral environment has a degenerate ground state. In the part of the series $\text{Ni}[\text{Cr}_{2-x}\text{Fe}_x]\text{O}_4$ in which $0 < x < 1$, the ions are distributed over both types of site: $\text{Ni}_{1-x}\text{Fe}_x[\text{Ni}_x\text{Cr}_{2-x}]\text{O}_4$. As already mentioned, nickel chromite with all the Ni at tetrahedral sites ($x = 0$) is deformed by a Jahn-Teller effect with $c/a > 1$. With increasing x the percentage of nickel ions at tetrahedral sites decreases. The mechanical coupling between the deformed tetrahedra is lost and the compound therefore becomes cubic. This happens at $x \approx 0.3$ (fig. 9). These compounds, however, are magnetic, and therefore the spin moments are ordered. Consequently a deformation due to spin-orbit coupling can be transmitted through the entire lattice. This deformation, now with $c/a < 1$, therefore occurs as soon as the Jahn-Teller

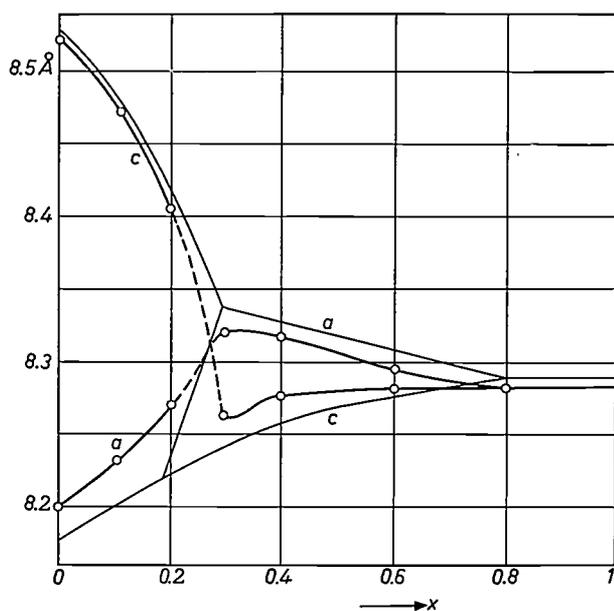


Fig. 9. Lattice constants a and c of the system $\text{NiFe}_x\text{Cr}_{2-x}\text{O}_4$. Thick curves: our own measurements at 120 °K. Thin curves: measurements at 90 °K [6]. For $0 < x < 0.3$, $c > a$. This tetragonal deformation is due to the Jahn-Teller effect of the Ni^{2+} ions at the tetrahedral sites. For $0.3 < x < 0.8$ we have $c < a$, owing to deformation resulting from spin-orbit coupling. For a discussion of the difference between the thin and thick curves see reference [7].

deformation has disappeared [7]. Fig. 10 gives a T - x diagram of this series; the diagram also shows the magnetic ordering temperature T_C . As required by the theory, the region with tetragonal structure ($c/a < 1$) is found only below this temperature.

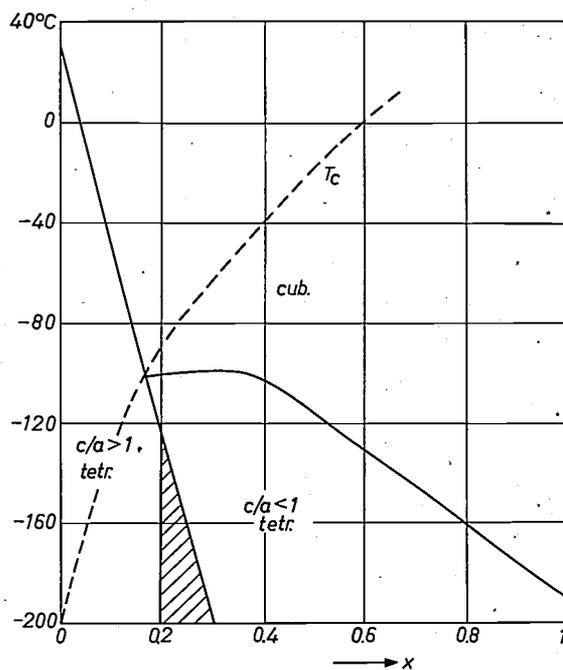


Fig. 10. x - T cross-section of the phase diagram of the system $\text{NiFe}_x\text{Cr}_{2-x}\text{O}_4$, after Arnott, Wold and Rogers [6]. The diagram shows a high-temperature cubic phase, two tetragonal phases and an orthorhombic phase (shown dashed) [7]. The tetragonal region with $c/a < 1$ is attributed to deformation due to spin-orbit coupling, and can therefore occur only below T_C . This is in fact the case, unlike the situation in the Jahn-Teller region (where $c/a > 1$).

Magnetic properties

When a substance containing magnetic ions which do not interact with each other is placed in a magnetic field, the moments of these ions will be oriented by the field. In a field H a magnetic moment $M = \chi H$ is produced. The magnetic susceptibility χ is determined by the magnitude of the moment of the ions. We have already seen that the orbital moment of the metal ions of the first transition series is absent or very small, so that the total spin moment S determines the magnitude of χ . It can be shown that

$$\chi_{\text{mole}} = \frac{Ng^2S(S+1)\beta^2}{3kT} = \frac{C}{T},$$

where N is Avogadro's number, g the Landé factor [8], β the Bohr magneton (unit of magnetic moment), k Boltzmann's constant, T the temperature and C the Curie constant. The slope of the χ^{-1} - T curve of such a paramagnetic substance (fig. 11) gives the value of S , and from this the electron configuration can be derived.

If there is magnetic interaction between the ion moments, then at low temperatures there will also be ordering of the moments without an external magnetic field. The interaction is said to be positive or ferromag-

netic if the moments are oriented parallel. If the interaction is antiferromagnetic or negative, then each moment cannot be oriented antiparallel to each of its neighbours; one or more sub-lattices are then formed in which the moments are parallel. The moments of the sub-lattices, which we can represent by vectors, are oriented antiparallel: $\uparrow \downarrow$.

In most magnetic oxides the interaction is negative. The fact that a spontaneous magnetic moment never-

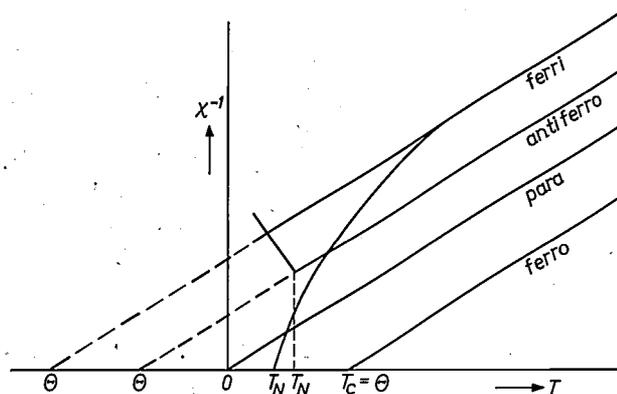


Fig. 11. Variation of the reciprocal of susceptibility as a function of temperature, shown diagrammatically. The susceptibility is infinite at $T = 0$ for a paramagnetic material and at $T = T_C$ for a ferromagnetic material. The asymptotic Curie temperature Θ is negative for antiferromagnetic and ferrimagnetic materials. Below the ordering temperature T_N the susceptibility is small for an antiferromagnetic material because the moments are antiparallel. For a ferrimagnetic material, χ is infinite at $T < T_N$ because the sub-lattice magnetizations do not completely compensate each other.

theless occurs below the ordering temperature is a consequence of the dissimilar magnitude of the sub-lattice magnetizations: this case is known as ferrimagnetism or non-compensated antiferromagnetism, indicated by $\uparrow \downarrow$.

If a magnetic field is applied to a substance in which there are magnetic interactions between the ions, the magnetic moments are subject to the effect of the neighbouring moments as well as the external magnetic field H . In the Weiss approximation this effect is accounted for by a hypothetical magnetic field which is proportional to the magnetization; it is then assumed that:

$$H_{\text{int}} = H_{\text{ext}} + \sum_j w_{ij} M_j; \quad w_{ij} = z_{ij} J_{ij}.$$

M_j is the magnetization of the sub-lattice j , J_{ij} is the interaction between an ion i and its neighbours j ; z_{ij} is the number of neighbours of type j . In the Weiss field approximation the susceptibility is given by

$$\chi = \frac{C}{T - \Theta}, \quad \text{where } \Theta = C \sum_j z_{ij} J_{ij}.$$

This relationship is valid only at high temperature ($T \gg |\Theta|$). The asymptotic Curie temperature is thus a measure of the algebraic sum of the interactions between the ions. If Θ is positive (ferromagnetic interaction), the moments assume parallel orientation below the Curie temperature $T_C \approx \Theta$. If Θ is negative (antiferromagnetic interaction), then ordering in sub-lattices occurs below a certain temperature T_N known as the Néel temperature (see fig. 10).

Magnetic susceptibility measurements can thus supply not only data on the electron configuration of the ions but also data on the interaction between the ions. The magnetic ions are often so far apart that any direct interaction between the electrons must be negligible. The interaction is often strong in those cases where a negative ion is situated on the connecting line between two positive ions. It is thus clear that the anion plays an essential role in the magnetic interaction. This is attributed to a deviation from the rigorous ionic model hitherto used. The pure d orbitals are replaced by a combination of a d orbital with a p orbital of the negative ion: $\varphi = d + ap$. The electrons which are localized around the metal nucleus in the purely ionic model are now also noticeable near the nucleus of the anion [9]. This deviation from the ionic model is known as covalence.

A detailed discussion of the theories of magnetic interaction [10] would be out of place here. It is possible however, using a simplified example, to see why the interaction is positive in certain configurations and negative in others.

As an example we take the magnetic interaction between Ni^{2+} ions, in the case where these are distributed among the octahedral sites of a close-packed structure of chlorine ions in such a way that all angles $\text{Ni}^{2+}\text{-Cl-Ni}^{2+}$ are 90° . The t_{2g} orbitals of the metal ions are completely filled. Each of the e_g orbitals contains one electron. Fig. 12 shows the $d_{x^2-y^2}$ orbitals of two nickel ions and the p_y and p_z orbitals of a chlorine ion. The p orbital parallel to the connecting line between the positive and the negative ion is called p_σ ; the p orbital perpendicular to this line is called p_π . Looking at the signs of the wave functions it can be seen in fig. 12 that the overlap $\int d_{x^2-y^2} p_\sigma d\tau$ between $d_{x^2-y^2}$ and the p_σ orbital is positive, whereas the overlap $\int d_{x^2-y^2} p_\pi d\tau$ is zero. The unpaired electrons from both nickel ions are therefore present together at the chlorine ion in different p orbitals. The same electro-

[9] Paramagnetic resonance experiments have shown that in MnF_2 , for example, a certain concentration of d electrons is present in the neighbourhood of the fluorine nuclei.

[10] P. W. Anderson in *Magnetism*, Vol. I (ed. G. T. Rado and H. Suhl), Academic Press, New York 1963.

[11] J. B. Goodenough, *Magnetism and the chemical bond*, Interscience Publ., New York 1963.

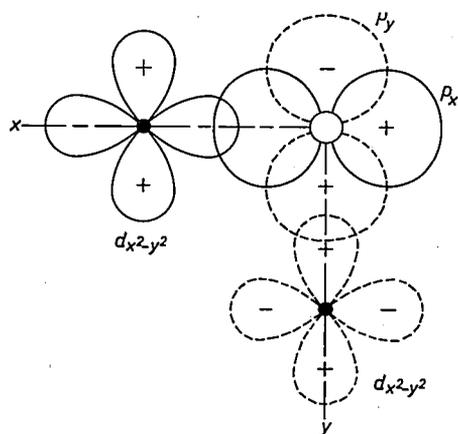


Fig. 12. Overlapping of $d_{x^2-y^2}$ orbitals of two nickel ions with the p_x and the p_y orbitals of a chlorine ion for the case where the angle Ni-Cl-Ni is equal to 90° . For orbitals drawn in the same way (unbroken or dotted curves) the overlapping $\int d_{x^2-y^2} p d\tau$ differs from zero.

static interaction that gives parallel spin orientation for the d electrons (Hund's rule) here causes a positive interaction between the spins of the nickel ions. The overlap between the d_{z^2} and the p orbitals of the chlorine ion is analogous with that for $d_{x^2-y^2}$.

An example of a compound in which only Ni-Cl-Ni angles of about 90° are present is NiCl_2 . The asymptotic Curie temperature θ is 80°K , which, in agreement with the above points to positive Ni-Ni interaction.

If the angle Ni-Cl-Ni is equal to 180° , the $d_{x^2-y^2}$ orbitals of the nickel ions overlap the same p_σ orbital (fig. 13). In this case the electron of the one nickel ion can be transferred via this p_σ orbital into the d orbital

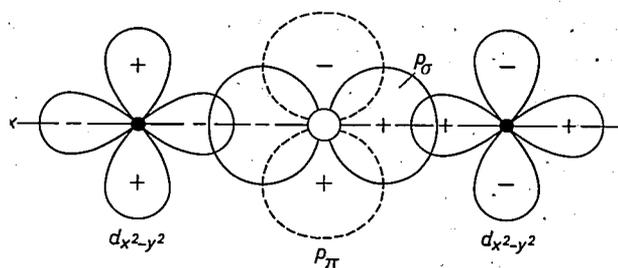


Fig. 13. Overlapping of $d_{x^2-y^2}$ orbitals of two nickel ions with the p_σ and the p_π orbital of a chlorine ion when the angle Ni-Cl-Ni is equal to 180° . There is no overlapping between $d_{x^2-y^2}$ orbitals and the p_π orbital.

of the other nickel ion. In view of the Pauli exclusion principle, however, this is possible only if the two spins are antiparallel. This transfer effect, which therefore leads to negative interaction, is assumed to be stronger than the positive interaction discussed earlier. In KNiF_3 all the angles Ni-F-Ni are 180° . This compound is in fact antiferromagnetic. The ordering temperature is $T_N = 275^\circ\text{K}$.

In the manner described above, the interaction between the ion spins has been estimated for various configurations of the $3d$ ions [11]. In many cases the sign predicted for the interaction corresponds to the sign of the experimentally determined magnetic interaction.

Finally, we shall consider the interaction between chromium ions for the case where the angle Cr-anion-Cr is 90° . The Cr^{3+} ion has three d electrons. The t_{2g} orbitals each contain one electron. Fig. 14 shows a few combinations of t_{2g} orbitals of two chromium ions with p orbitals of the anion. In the case of the t_{2g}

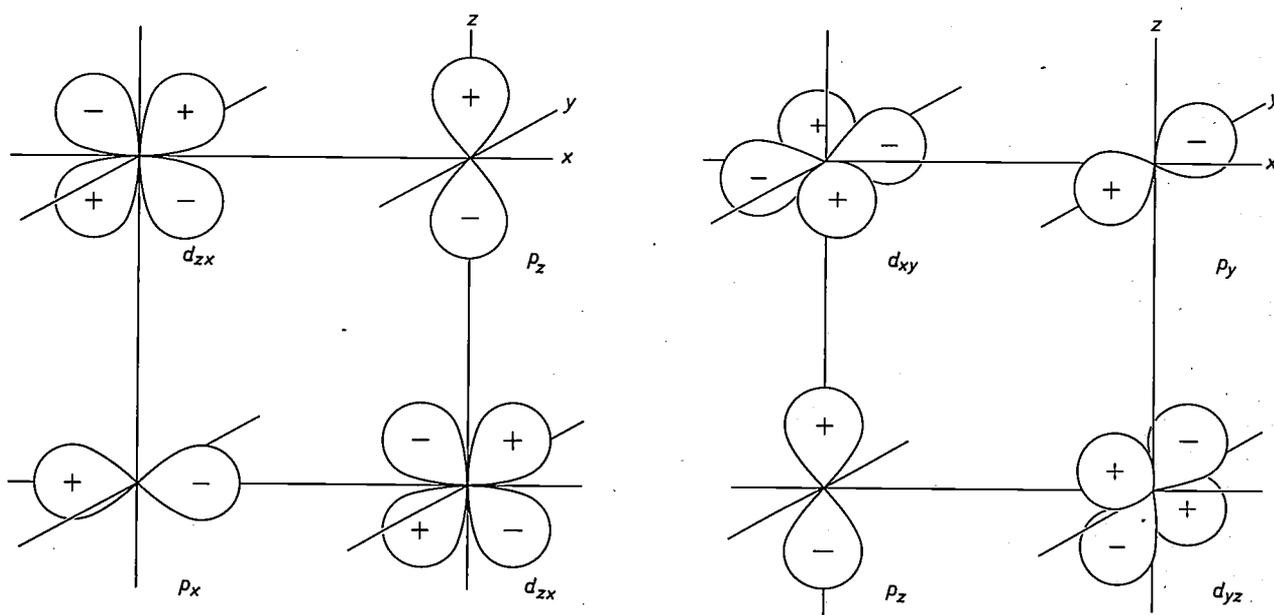


Fig. 14. Overlapping between the d_{xy} orbitals of two chromium ions with the p_x , p_y and p_z orbitals of an anion.

orbitals there is overlapping with $p\pi$ orbitals. The transfer of a d electron is possible only with the combination $d_{xy}-p_y-d_{yz}$. This combination gives a negative contribution to the interaction, whereas three other combinations lead to positive interaction. In this situation the resulting interaction is expected to be weakly positive or negative. In oxides, however, the Cr-O-Cr interaction over 90° is strongly negative. This effect is attributed to the direct overlapping of t_{2g} orbitals (fig. 15).

The compounds NaCrX_2 ($X = \text{O, S, Se}$) have the ordered NaCl structure, the compounds ACr_2X_4 ($A = \text{Zn or Cd, X = O, S, Se}$) the spinel structure. In both crystal structures only Cr-anion-Cr angles of about 90° occur. In the oxides the interaction is negative, in the sulphides and selenides it is positive (see Table IV). Since the Cr-Cr distance between the

Table IV. Cr-Cr distance and asymptotic Curie temperature Θ of a compound in which only Cr-anion-Cr angles of about 90° occur.

	Cr-Cr distance (Å)	Θ (°K)
NaCrO_2	2.98	-354
NaCrS_2	3.53	+ 30
NaCrSe_2	3.71	+130
ZnCr_2O_4	2.89	-380
CdCr_2S_4	3.62	+156
CdCr_2Se_4	3.80	+204

chromium ions in the series O, S, Se increases, there will be a decrease in the negative contribution to the interaction, attributable to *direct* overlapping of the t_{2g} orbitals. Owing to increasing covalence, i.e. increasing α , both the positive and the negative contribution to the interaction *via the anion* will on the contrary increase in the series O, S, Se. The picture given above therefore corresponds to the experimental findings if the Cr-anion-Cr interaction is positive.

In the foregoing we have attempted to show how crystal field theory provides a qualitative explanation for the properties of numerous compounds of transition metal ions, such as lattice deformation at low temperature, the preference for a particular co-ordination and the occurrence of high and low spin-states. Although the electronic states cannot (yet) be calculated in quantitative terms, qualitatively correct rules

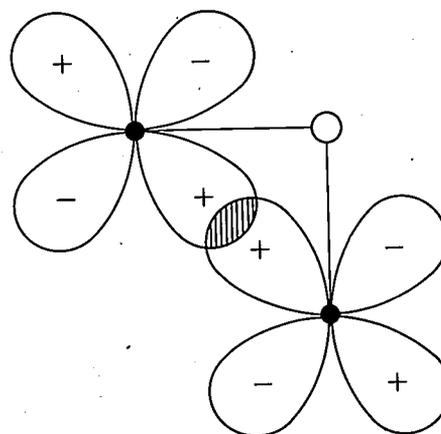


Fig. 15. Direct overlapping between the d_{xy} orbitals of two chromium ions, for a smaller Cr-Cr distance than in fig. 14.

governing the magnetic interaction are arrived at by taking simple combinations of anion and cation wave functions. In part II of this article it will be shown how these rules, combined with knowledge of crystal structures, are being used in realizing compounds with certain desired magnetic properties.

Summary. Crystal field theory is a useful tool for studying the crystal chemistry of oxides of transition metal ions. The theory is applied here first to an ion with one d electron; the splitting of the d level is discussed for octahedral and tetrahedral environment with 6 or 4 negative ions respectively. The electron state and spin moment are then derived for ions which have more than one d electron in the same two types of co-ordination. If the crystal field is very much stronger than the electrostatic repulsion between the d electrons, Hund's rule is no longer valid, and as a result the ions occur in a "low spin-state". This effect is discussed with a few examples.

The metal ions undergo a stabilizing effect in the crystal field, and this effect depends on the number of d electrons. This crystal-field stabilization is different for octahedral and tetrahedral anionic environment. The result is that certain ions, e.g. Cr^{3+} , show a marked preference for octahedral sites. Lowering of the symmetry of the crystal field gives additional crystal-field stabilization for certain ions. This effect, the Jahn-Teller effect, explains a number of spontaneous deformations of the crystal lattice that occur in oxides in which Mn^{3+} or Cu^{2+} ions are found at octahedral sites, or Cu^{2+} or Ni^{2+} ions at tetrahedral sites. The lattice deformations that frequently occur (e.g. in CoO) if the magnetic moments are ordered can be explained by the coupling between spin moment and orbital moment of the metal ions. The competition between the Jahn-Teller effect and the deformation due to spin-orbit coupling is discussed in connection with the system NiCr_2O_4 - NiFe_2O_4 .

Finally, it is shown briefly how the magnitude of the magnetic moment and the interaction between the moments of the metal ions can be derived from susceptibility measurements. Qualitative considerations show that the magnetic interaction between Ni^{2+} ions is positive if the angle formed by Ni^{2+} -anion- Ni^{2+} is 90° , and that the interaction is negative if this angle is 180° . The interaction between Cr^{3+} ions in oxides, sulphides and selenides, in which the angle Cr-anion-Cr is roughly 90° , is also discussed.

II. Magnetic compounds with spinel structure

G. Blasse

An important part of the research on the oxides of the transition elements is concerned with the relation between their crystal structure and chemical composition and their magnetic properties. By giving the oxides the appropriate chemical composition it is possible, within certain limits, to predetermine their magnetic properties. In this article we shall give a number of examples to illustrate this procedure, often referred to as "ionic engineering". We shall confine ourselves to materials with spinel structure, because many magnetic materials in use have this structure. To obtain scientific insight it is often necessary to study compounds with a simpler structure, such as the perovskite or rock salt structure. These compounds, however, show no magnetic properties that can be put to use. There are other materials, on the other hand, that have highly complex crystal structures (e.g. garnets and hexagonal ferrites), but with these the relation between properties and chemical composition is much more difficult to understand [1].

A typical example of the influence of composition on magnetic properties is the following. Zinc ferrite (ZnFe_2O_4) and magnesium ferrite (MgFe_2O_4) are similar compounds of the form $\text{Me}^{2+}\text{Fe}_2^{3+}\text{O}_4$ (Me = metal). Both have the spinel structure, and the constituent ions of the one are magnetically equivalent to those of the other: Fe^{3+} is paramagnetic, Mg^{2+} and Zn^{2+} are both diamagnetic. The magnetic properties, however, are widely different. Magnesium ferrite exhibits a spontaneous magnetic moment and has a very high Curie temperature (715 °K), whereas zinc ferrite has no magnetic moment, even at low temperature. These differences are apparently due to the choice of the bivalent diamagnetic ion in the composition $\text{Me}^{2+}\text{Fe}_2\text{O}_4$.

The spinel structure

It is not possible to build up a magnetic material properly if the crystal structure is unknown. First, therefore, we shall give a description of the spinel structure.

Compounds with spinel structure have the general chemical formula AB_2X_4 . In this formula X represents an anion (in most cases the O^{2-} ion). These anions form a cubic close-packed structure. In this structure

the smaller cations fit into the interstices, which are surrounded by four or by six anions arranged in a regular tetrahedron or octahedron (see *fig. 1*). A represents a cation at a tetrahedral site (or A site) and B a cation at an octahedral site (or B site). We see that there are therefore two possible sites in the structure for the cations. It is because of this structural peculiarity that spinel structure ferrites can have a magnetic moment. This means however that the properties of

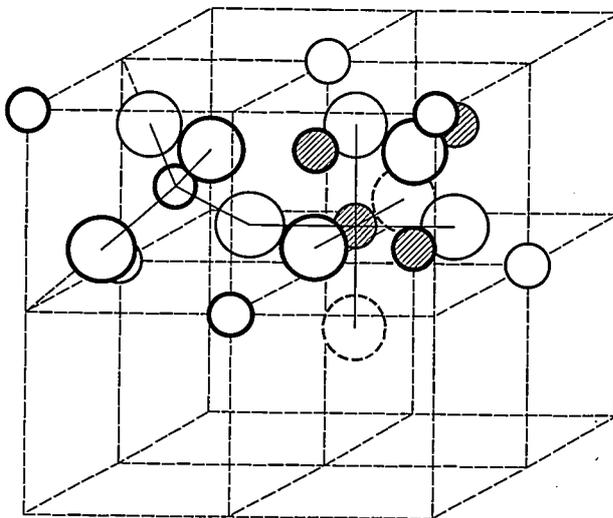


Fig. 1. The spinel structure. The positions of the ions in two octants of the unit cell are shown. The large circles represent anions, the small hatched circles octahedrally surrounded cations (B sites) and the small open circles tetrahedrally surrounded cations (A sites).

compounds of this kind are more difficult to explain, as we shall see shortly.

Let us first consider, very generally, ferrites with the above-mentioned composition $\text{Me}^{2+}\text{Fe}_2^{3+}\text{O}_4$. The cation distribution over the A and B sites may be as follows:

- $\text{Me}^{2+}[\text{Fe}_2^{3+}]\text{O}_4$, i.e. A = Me^{2+} and B = Fe^{3+} .
- $\text{Fe}^{3+}[\text{Me}^{2+}\text{Fe}^{3+}]\text{O}_4$, i.e. A = Fe^{3+} and B = Me^{2+} and Fe^{3+} .

Following the usual notation, we have put brackets round the cations at B sites. Distribution (a) is called normal, and distribution (b) inverse. Various kinds of intermediate distributions are possible.

It may be asked which distribution has the lowest

Dr. G. Blasse is with Philips Research Laboratories, Eindhoven.

[1] Compounds with other crystal structures are dealt with in a recent review article by G. Blasse, in: Progress in Ceramic Science (ed. J. E. Burke), Pergamon Press, Vol. 4, p. 133, 1966.

energy. The cation distribution is determined by a large number of energy terms, which are to some extent conflicting. While some terms favour the normal distribution, other terms will oppose it because they favour the inverse distribution. It is therefore difficult, and usually in fact impossible, to predict which distribution will materialize. The following are some of the energy terms involved [2]:

a) The Madelung energy, which originates from the electrostatic attraction and repulsion between the positive and negative ions. It follows from electrostatic considerations, for example, that lower energies result with large ions of low charge at A sites or small ions of high charge at A sites.

b) The Born energy, which is related to the repulsion between the electron clouds of the ions.

c) The ordering energy, where cations of different charge are located on one type of lattice site: the effect of ordering is to minimize electrostatic repulsion between the cations (the highly charged cations surround themselves as far as possible with ions of low charge). Calculations of these three energy terms were carried out in the forties by members of our laboratory [3].

d) The energy connected with the preference of particular ions for occupying tetrahedral or octahedral sites (see part I of this article). Familiar examples are the preference of the Zn^{2+} ion ($3d^{10}$ configuration) for tetrahedral sites and of the Cr^{3+} ion ($3d^3$ configuration) for octahedral sites. The crystal field theory discussed in part I shows that in general ions of configuration d^1 , d^2 , d^4 , d^6 , d^7 and d^9 have some preference for octahedral sites, and ions with configuration d^3 and d^8 a strong preference for octahedral sites, while ions with configuration d^0 , d^5 and d^{10} show no preference. The crystal field theory does not, however, take account of deviations from the ionic bonding model, which certainly occur (cf. part I). We have therefore attempted to approach this preferential energy with the aid of the more general but more complicated ligand field theory [2]. Quite a number of simplifications have to be introduced, however, in order to obtain a result, and this will therefore not be much more reliable than that obtained with crystal field theory. We found that ions with configuration d^5 , d^6 , d^7 , d^9 and d^{10} have a preference for tetrahedral sites, ions with configuration d^0 , d^1 , d^2 , d^4 and d^8 have no preference, and ions with configuration d^3 have a preference for octahedral sites.

e) The polarization energy of the anion. The anion is surrounded more or less tetrahedrally by one cation at an A site and three cations at B sites. By making an extreme distribution of the positive charges, for example low-charged ions at A sites and highly charged ions at B sites, a strong electric field is produced around

the anion, causing it to become polarized.

A detailed study of cation distribution in the spinel structure has shown that the last two energy terms are often critical, as the first three roughly compensate one another in various distributions. This is no more than a rule of thumb, however.

We shall now consider a few examples.

The spinel $MgAl_2O_4$, which contains exclusively ions with the inert gas configuration, is normal: $Mg^{2+}[Al_2^{3+}]O_4$. If the Mg^{2+} ion is replaced by other bivalent ions, for example Mn^{2+} , Fe^{2+} , Co^{2+} , Ni^{2+} or Cu^{2+} (the electron configurations being $3d^5$, $3d^6$, $3d^7$, $3d^8$, $3d^9$ respectively), then the normal distribution remains, except with $NiAl_2O_4$, which is almost entirely inverse. This indicates that the Ni^{2+} ion has a greater preference for B sites — or at any rate less preference for A sites — than the other bivalent ions mentioned, in agreement with the two rules given under (d), so that the inverse distribution of $NiAl_2O_4$ must be attributed to the preference energy.

If the Al^{3+} ion in $MgAl_2O_4$ is replaced by other trivalent ions, for example Ti^{3+} , V^{3+} , Cr^{3+} , Mn^{3+} , Fe^{3+} or Ga^{3+} (electron configurations $3d^1$, $3d^2$, $3d^3$, $3d^4$, $3d^5$, $3d^{10}$ respectively), a normal distribution is again found, except with Fe^{3+} and Ga^{3+} . The latter ions drive out the Mg^{2+} ion from the A sites. This is in agreement with the second rule of preference under (d).

The effect of the polarization energy is well illustrated by the cation distribution of spinels containing lithium, for example $LiMe^{2+}Me^{5+}O_4$ and $LiMe^{3+}Me^{4+}O_4$. In these spinels the Li^+ ion is always found at A sites, so that the ions of low and high charge are grouped asymmetrically around the anion and the polarization energy is high. There is a deviation from this distribution only if cations are present which have a strong preference for A sites, as in the case of $LiFeTiO_4$ and $LiGaTiO_4$.

Insight into the factors determining the distribution of cations can best be obtained by studying a series of compounds in which only one cation is varied. This method has been followed in our work.

Magnetic properties

The magnetic properties of compounds with spinel structure, in particular of ferrites, have been explained by Néel [4]. In a spinel $A[B_2]O_4$ with paramagnetic ions A and B the following magnetic interactions may be distinguished:

- a) AB interaction, a magnetic interaction between cations at tetrahedral and cations at octahedral sites;
- b) BB interaction, a magnetic interaction between the cations at octahedral sites. (The analogous interaction between cations at tetrahedral sites — AA interaction — is very weak owing to the considerable distance be-

tween the cations on A sites. This interaction is therefore disregarded.)

Néel assumed that the AB interaction is negative and strong compared with the BB interaction. The magnetic structure of a spinel, that is the manner in which the magnetic moments in the crystal lattice are oriented with respect to one another, can then be represented symbolically as $\vec{A}[\vec{B}\vec{B}]O_4$. The sign of the BB interaction is not relevant here. The AB interaction orients the moments at all the A sites antiparallel to those at all the B sites. Since, generally speaking, the magnetic moment of two B ions will not be identical with that of one A ion, the result is a magnetic moment which is equal to the difference between the magnetic moments at A and at B sites. This phenomenon is known as ferrimagnetism.

An actual example is nickel ferrite with the cation distribution $Fe[NiFe]O_4$. The Fe^{3+} ion ($3d^5$) has a magnetic moment of $5\mu_B$, the Ni^{2+} ion ($3d^8$) a magnetic moment of about $2.3\mu_B$ [5]. The resultant magnetic moment is thus:

$$m = m_B - m_A = (5\mu_B + 2.3\mu_B) - 5\mu_B = 2.3\mu_B.$$

Values have been found experimentally between $2.2\mu_B$ and $2.3\mu_B$ (see, for example, reference [2]) so that the agreement between theory and experiment is good.

The difference between the magnetic properties of magnesium ferrite and zinc ferrite, referred to in the introduction are thus explained. Whereas magnesium ferrite is almost inverse ($Mg_{0.1}\vec{Fe}_{0.9}[Mg_{0.9}\vec{Fe}_{1.1}]O_4$), zinc ferrite is normal ($Zn[Fe_2]O_4$) because Zn^{2+} has a preference for A sites. Magnesium ferrite therefore has a magnetic moment $(1.1 - 0.9) \times 5\mu_B = 1\mu_B$, owing to the negative interaction between unequal numbers of Fe^{3+} ions at A and at B sites. In zinc ferrite, however, there is no AB interaction (Zn^{2+} is diamagnetic). In fact, zinc ferrite shows paramagnetic behaviour down to a very low temperature. This also proves that the BB interaction between Fe^{3+} ions must be very weak.

It is interesting to note that the magnetic moment of magnesium ferrite differs from zero owing to the slight deviation from the inverse cation distribution. For the hypothetical completely inverse $Fe[MgFe]O_4$ one would expect a magnetic moment equal to zero.

Néel's hypothesis implies the possibility of anomalous magnetization-temperature curves ($M-T$ curves).

The saturation magnetization of a ferromagnetic material decreases with increasing temperature and vanishes at the Curie temperature; this is the Brillouin curve, shown in fig. 2. Under certain conditions, completely different $M-T$ curves may be expected for a ferrimagnetic material. This is because the magnetization of a ferrimagnetic compound represents the difference between two or more sub-lattice magnetizations. Now if the temperature-dependence is not the same for these sub-lattice magnetizations, anomalous $M-T$ curves may be found. This is illustrated diagrammatically in fig. 3, where two dissimilar sub-lattice magnetizations, M_A and M_B , are shown as a function of temperature. The difference $M_B - M_A$ is the magnetization M of the compound. If $|M_B| > |M_A|$ at $0^\circ K$

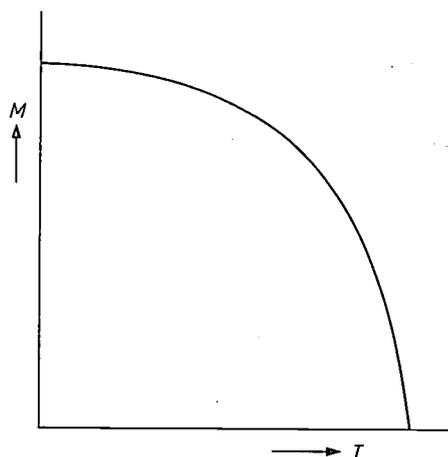


Fig. 2. Magnetization M of a ferromagnetic material as a function of temperature T .

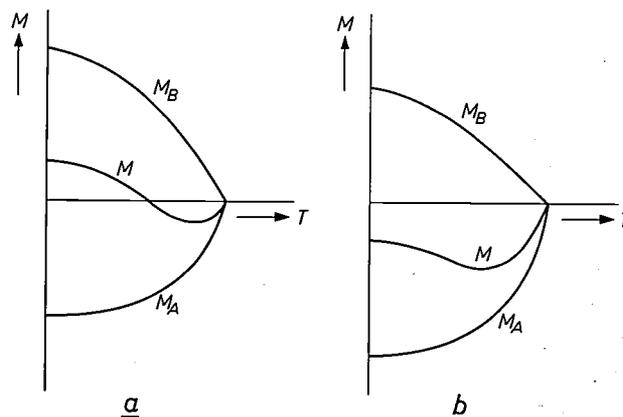


Fig. 3. How anomalous magnetization-temperature curves ($M-T$ curves) for a ferrimagnetic material can arise. The sub-lattice magnetizations M_A and M_B (negative and positive respectively) are shown as functions of temperature. These disappear at the same temperature (the Curie temperature) because they owe their existence to their exchange interaction (AB interaction). The difference $M_B - M_A$ is the magnetization M of the compound. This is measured experimentally (M_A and M_B are not directly measurable).

It is assumed in the figure that M_B decreases faster with increasing T than M_A . If now $|M_B| > |M_A|$ at $0^\circ K$, an $M-T$ curve with a compensation point may be found (a); if $|M_B| < |M_A|$, the result may be a curve showing a maximum (b). Néel called the first a type N curve and the second a type P curve.

[2] For further details, see G. Blasse, Philips Res. Repts., Suppl. 1964, No. 3.

[3] E. J. W. Verwey, F. de Boer and J. H. van Santen, J. chem. Phys. 16, 1091, 1948.

F. de Boer, J. H. van Santen and E. J. W. Verwey, J. chem. Phys. 18, 1032, 1950.

[4] L. Néel, Ann. Physique 3, 137, 1948.

[5] The spin moment of $Ni^{2+}(3d^8)$ is $2\mu_B$, but there is also a contribution from the orbital momentum.

and M_B decreases faster with temperature than M_A , an M - T curve is found with a compensation point at which north and south poles change sign (type N magnetization curve, fig. 3a). If, however, $|M_B| < |M_A|$ at 0 °K, the resultant curve has a maximum (type P magnetization curve, fig. 3b). Both anomalous magnetization curves were found experimentally after Néel had predicted them: the type N curve was found in our laboratory by Gorter and Schulkes [6], the type P curve by Maxwell and Pickart [7]. This phenomenon clearly indicates that there must be more than one magnetic lattice present in a ferrimagnetic compound. The effect revealed by the type P magnetization curve opens up the possibility of making magnetic materials whose magnetization is independent of temperature in a particular temperature range.

The magnetic sub-lattices can be given a widely different temperature dependence by introducing in one way or another a strongly negative BB interaction. This interaction as it were disturbs the parallel orientation of the moments of the B cations, so that the magnetic moment of the B lattice decreases rapidly. This situation is assumed in fig. 3a and b.

We shall now show how certain magnetic properties can be produced by chemical substitutions.

The nickel ferrite-vanadite system ($\text{NiFe}_{2-t}\text{V}_t\text{O}_4$)

If the Fe^{3+} ion in nickel ferrite, $\text{Fe}[\text{NiFe}]\text{O}_4$, is replaced by the V^{3+} ion (configuration $3d^2$), the magnetic properties are drastically changed. With increasing vanadium content the magnetic moment and the Curie temperature decrease, and the shape of the M - T curve undergoes marked variation. In order to be able to interpret these effects, it is necessary first of all to know the cation distribution of this system. For the time being, we shall consider only low vanadium concentrations, i.e. $0 \leq t \leq 1$. The cation distribution then proves to be $\text{Fe}[\text{NiFe}_{1-t}\text{V}_t]\text{O}_4$ [8]. This means that the vanadium introduced takes the place of only the ferric ions on B sites. Since the V^{3+} ion has two unpaired electrons, the magnetic moment of this ion will be roughly $2 \mu_B$. This means therefore that the total magnetic moment decreases as the vanadium content increases, the reason being that the V^{3+} ion has a smaller moment than the Fe^{3+} ion it replaces. At a particular vanadium concentration ($t \approx 0.75$) the magnetizations of A and B sites become identical, so that the total magnetic moment (at 0 °K) is zero (see fig. 4). At a vanadium concentration higher than this, the magnetization of the A lattice begins to predominate over that of the B lattice, as for example in $\text{Fe}[\text{NiV}]\text{O}_4$: $M = M_B - M_A = (2.3 \mu_B + 2 \mu_B) - 5 \mu_B = -0.7 \mu_B$ [5]. Since we take the magnetization of the B lattice as invariably positive, the magnetic moments

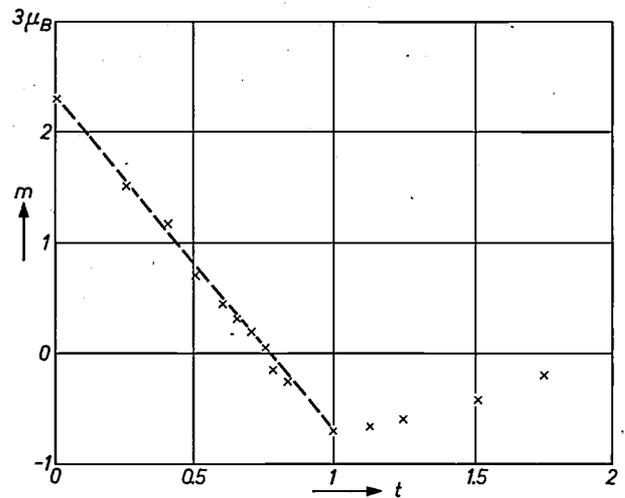


Fig. 4. Magnetic moment m (in Bohr magnetons) of the system nickel ferrite-vanadite ($\text{NiFe}_{2-t}\text{V}_t\text{O}_4$) as a function of vanadium content t . The crosses indicate values found experimentally. The straight line was calculated for the cation distribution $\text{Fe}[\text{NiFe}_{1-t}\text{V}_t]\text{O}_4$ on the basis of Néel's hypothesis and using the following values for the magnetic moments of the ions: $\text{Ni}^{2+} 2.3 \mu_B$; $\text{Fe}^{3+} 5 \mu_B$; $\text{V}^{3+} 2 \mu_B$.

for $t > 0.75$ are shown in the negative direction in fig. 4.

The decrease of the Curie temperature (fig. 5) with increasing vanadium content is also easily understood, the AB interaction between Fe^{3+} ions being much stronger than that between Fe^{3+} and V^{3+} ions, so that the average AB interaction decreases with increasing vanadium content. Since the AB interaction is the predominant magnetic interaction, this also determines, to a good approximation, the value of the Curie temperature.

At a higher vanadium concentration ($1 \leq t \leq 2$) the magnetization values differ considerably from those that can be calculated for any reasonable cation distribution on the basis of Néel's hypothesis. This means that the BB interaction is no longer weak compared with the AB interaction. An asymptotic Curie temperature equal to -750 °K has indeed been found for the spinel $\text{Mg}[\text{V}_2]\text{O}_4$, a compound in which the only possible magnetic interaction is BB interaction between V^{3+} ions. Evidently, therefore, this BB interaction must be strongly negative. Because of this

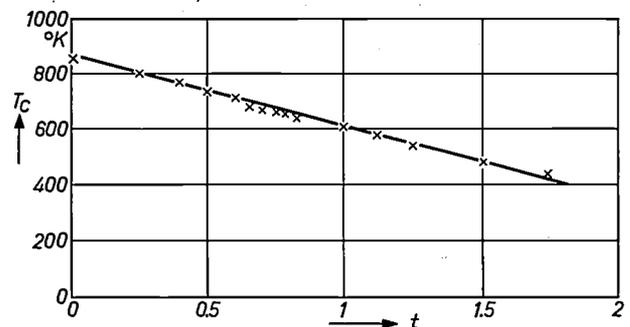


Fig. 5. The Curie temperature T_C for the system $\text{NiFe}_{2-t}\text{V}_t\text{O}_4$ as a function of the vanadium content t .

strongly negative BB interaction between V^{3+} ions the magnetic moments at B sites will no longer be parallel at a particular vanadium concentration (probably $t > 1$) but will align themselves at a particular angle to each other, thus making concessions to both the AB and the BB interaction. This may explain why the experimentally determined magnetizations differ so much from those predicted on the basis of Néel's hypothesis.

We have already stated that strong BB interactions would give rise to anomalous $M-T$ curves. These have in fact been found for certain values of t (see fig. 6). The $M-T$ curve for compositions with $t < 0.75$ is of type N , and for compositions with $t > 0.75$ it is of type P . This agrees well with the theory, since the magnetization of the B ions is greater than that of the A ions for $t < 0.75$, so that we have the situation of fig. 3a, while the magnetization of the A ions is greater than that of the B ions when $t > 0.75$, giving the situation in fig. 3b. The assumption here, then, is that the average BB interaction is strongly negative, and this is easily demonstrated by experiments on the spinel $Mg[V_2]O_4$ (see above).

We have succeeded in making a ferrite which has a low temperature-independent magnetization at about room temperature ($NiFeVO_4$, the composition with $t = 1$; see fig. 6). This property is particularly desirable for microwave ferrites. Unfortunately, the ferrite with this composition has certain other properties which make it less suitable for application, namely a relatively low electrical resistivity (about $100 \Omega cm$) and a high anisotropy.

By introducing a strongly negative BB interaction in this case we have made the normal $M-T$ curve of nickel ferrite change into anomalous curves. The curve with the compensation point, type N , was first obtained, and then the one with the maximum, type P . It may now be asked whether it is also possible to introduce a *positive* BB interaction into nickel ferrite, thus causing the normal $M-T$ curve to change first into a curve of type P and only afterwards into a curve of type N .

The Ni ferrite-antimonate system ($Ni_{1+2t}Fe_{2-3t}Sb_tO_4$)

The inverse sequence of $M-T$ curves (first type P and then type N) has been obtained by replacing an increasing number of the Fe^{3+} ions in nickel ferrite by a combination of Ni^{2+} and Sb^{5+} ions, in accordance with the rule $3Fe^{3+} \rightarrow 2Ni^{2+} + Sb^{5+}$. This results in materials that have the general composition $Ni_{1+2t}Fe_{2-3t}Sb_tO_4$. The maximum attainable value of

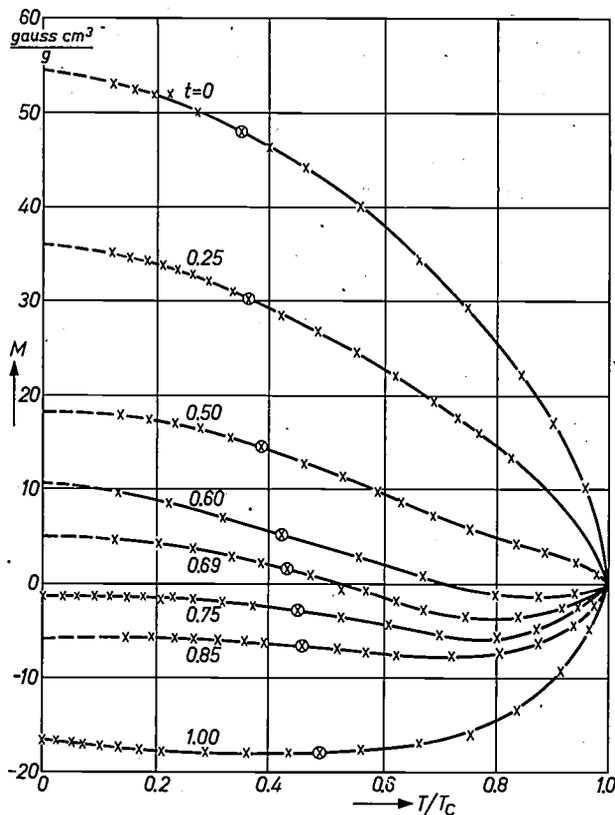


Fig. 6. The $M-T$ curve for the system $NiFe_{2-t}V_tO_4$ for various values of t ($0 \leq t \leq 1$). The vertical scale shows the magnetization M (in gauss $cm^3/gram$) and the horizontal scale shows the normalized temperature, i.e. the temperature T at which the measurement was carried out, divided by the Curie temperature T_C , which itself depends on t . The ringed crosses indicate measurements at room temperature.

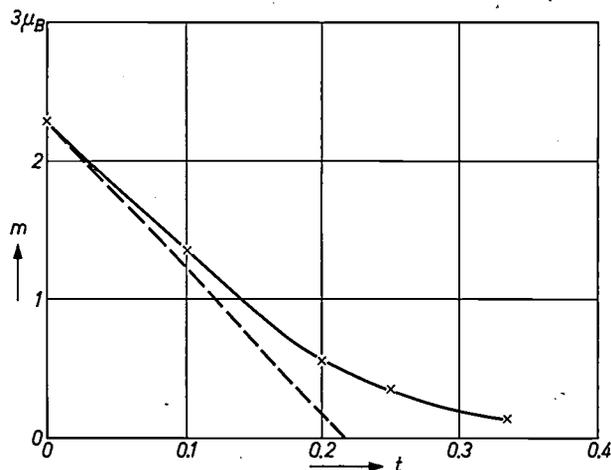


Fig. 7. Magnetic moment m for the system nickel ferrite-antimonate ($Ni_{1+2t}Fe_{2-3t}Sb_tO_4$). The crosses indicate experimentally determined values. The straight line was calculated for the cation distribution $Fe[Ni_{1+2t}Fe_{1-3t}Sb_t]O_4$, in other words with all Ni^{2+} and Sb^{5+} ions at B sites. In reality a small proportion of the nickel ions are to be found on A sites.

t is about $1/3$. At values of $t > 1/3$ single-phase spinels can no longer be obtained ($NiSb_2O_6$ forms a second phase). The magnetization and the $M-T$ curve as a function of t are shown in figs. 7 and 8.

[6] E. W. Gorter and J. A. Schulkes, Phys. Rev. 90, 487, 1953.

[7] L. R. Maxwell and S. J. Pickart, Phys. Rev. 92, 1120, 1953.

[8] The ways in which the cation distribution can be determined experimentally are discussed and commented upon in reference [2].

The cation distribution is seen to be as follows. The Sb^{5+} ions are located solely at B sites. The Ni^{2+} ions are for the greater part at B sites. The percentage of Ni^{2+} ions at A sites increases with increasing value of t .

By way of illustration we give the distribution for the two extreme compositions, i.e. for NiFe_2O_4 ($t = 0$) and for $\text{Ni}_{5/3}\text{FeSb}_{1/3}\text{O}_4$ ($t = 1/3$): these are $\text{Fe}[\text{NiFe}]_2\text{O}_4$ and $\text{Ni}_{1/3}\text{Fe}_{2/3}[\text{Ni}_{4/3}\text{Fe}_{1/3}\text{Sb}_{1/3}]_2\text{O}_4$ respectively.

It can be seen from fig. 8 that the M - T curve in this system changes from normal to type P . The type N is not found here. This is attributable to the fact that the state in which the magnetization of the A ions is greater than that of the B ions is not attained. The occurrence of a type P curve points to a positive BB interaction. Since the mutual BB interactions between Ni^{2+} ions and between Fe^{3+} ions are extremely weak, it can obviously be assumed that the BB interaction

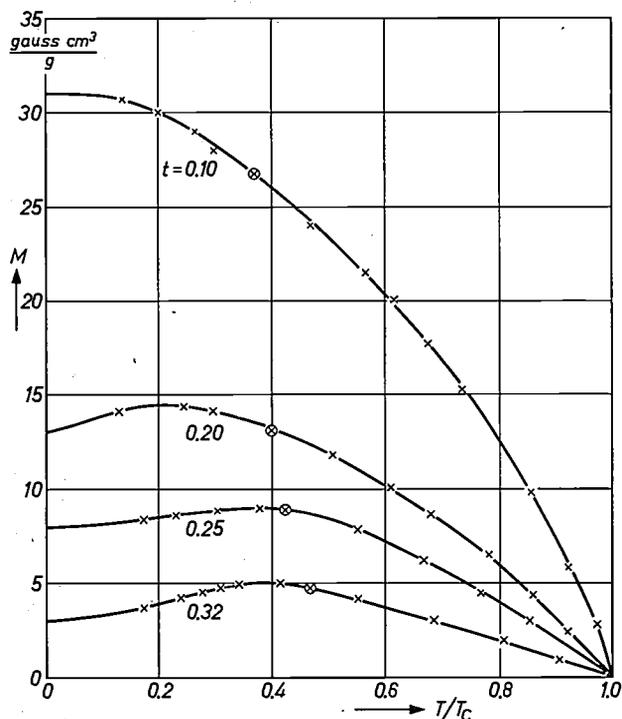


Fig. 8. The M - T curve for the system $\text{Ni}_{1+2t}\text{Fe}_{2-3t}\text{Sb}_t\text{O}_4$ for various values of t . The vertical scale shows the magnetization per gram, and the horizontal scale shows the normalized temperature. The ringed crosses indicate measurements at room temperature.

between Ni^{2+} and Fe^{3+} is positive. Further evidence supporting this assumption was later found independently.

As fig. 8 shows, the introduction of a positive BB interaction has enabled a P -type M - T curve to be obtained. The maximum in the curve, however, is between 200 and 250 °K, and this makes the material less suit-

able for application. Moreover, the shape of the curve at the maximum is rather sharp.

Nevertheless, the existence of positive BB interaction raises the question of whether it might be possible to make a ferromagnetic instead of a ferrimagnetic spinel. In a ferromagnetic spinel all the magnetic moments would be parallel, making it possible to achieve a very high magnetization.

Ferromagnetic oxides with spinel structure

Since it is very improbable on theoretical grounds that the AB interaction between any ions could ever be positive, a ferromagnetic spinel can only be obtained by locating diamagnetic cations at A sites and paramagnetic cations, between which there must be a positive interaction, at B sites. Materials have recently been found in which this situation occurs. An example is $\text{Cu}^+[\text{Mg}_{0.5}\text{Mn}_{1.5}^{4+}]\text{O}_4$ [9]. The Cu^+ ion ($3d^{10}$), like the Mg^{2+} ion, is diamagnetic. The Mn^{4+} ion is the only paramagnetic ion present. The magnetic interaction proves to be positive, so that the material becomes ferromagnetic. The Curie temperature is however rather low at 57 °K.

The nickel ferrite-rhodite and cobalt ferrite-rhodite systems ($\text{NiFe}_{2-t}\text{Rh}_t\text{O}_4$ and $\text{CoFe}_{2-t}\text{Rh}_t\text{O}_4$)

Finally, we shall show how a diamagnetic ion can affect the cation distribution of paramagnetic ions, and hence also the magnetic properties, in an entirely different manner. This we shall do by considering the replacement of Fe^{3+} ions by Rh^{3+} ions in nickel ferrite and cobalt ferrite.

The Rh^{3+} ion ($4d^6$) is in the low-spin state, i.e. the six d electrons are paired. This means that this ion is diamagnetic in spite of the presence of an incompletely filled d shell. (We disregard here a weak temperature independent paramagnetism.)

It proves to be possible to make spinels that have the composition $\text{NiFe}_{2-t}\text{Rh}_t\text{O}_4$ and $\text{CoFe}_{2-t}\text{Rh}_t\text{O}_4$. In figs. 9 and 10 the magnetic moment in these systems is shown as a function of composition.

The diamagnetic Rh^{3+} ions are found to occupy octahedral sites only. Whereas nickel and cobalt ferrite are inverse spinels ($\text{Fe}[\text{NiFe}]_2\text{O}_4$ and $\text{Fe}[\text{CoFe}]_2\text{O}_4$ respectively), nickel rhodite ($\text{Ni}[\text{Rh}_2]\text{O}_4$) and cobalt rhodite ($\text{Co}[\text{Rh}_2]\text{O}_4$) are normal. In the systems under consideration a transition must therefore take place from an inverse to a normal cation distribution.

In addition to the experimental magnetization values fig. 9 also shows the values calculated for the distribution $\text{Fe}[\text{NiFe}_{1-t}\text{Rh}_t]\text{O}_4$ on the basis of Néel's hypothesis. It can be seen that fairly good agreement exists in the concentration region $0 < t < 1$. The fact

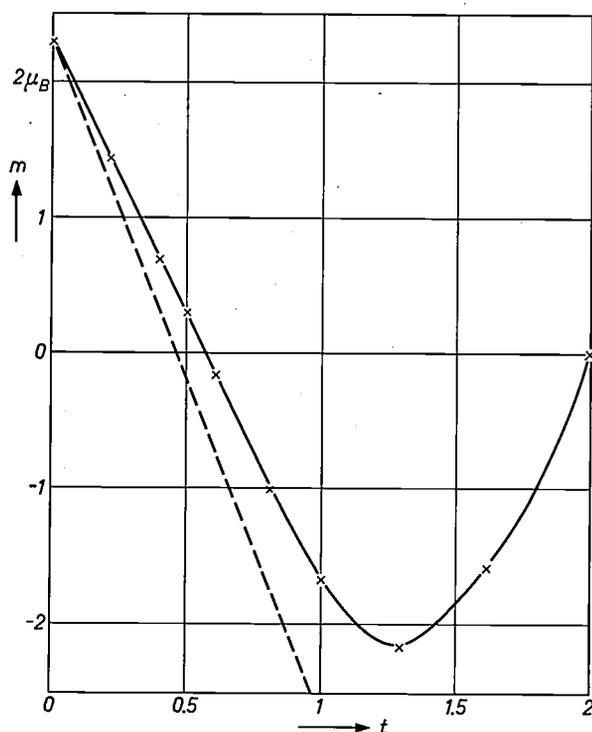


Fig. 9. Magnetic moment m of the system nickel ferrite-rhodite ($\text{NiFe}_{2-t}\text{Rh}_t\text{O}_4$). The crosses indicate experimentally determined values. The straight line was calculated for the distribution $\text{Fe}[\text{NiFe}_{1-t}\text{Rh}_t]\text{O}_4$, in other words with all Ni^{2+} and Rh^{3+} ions at B sites.

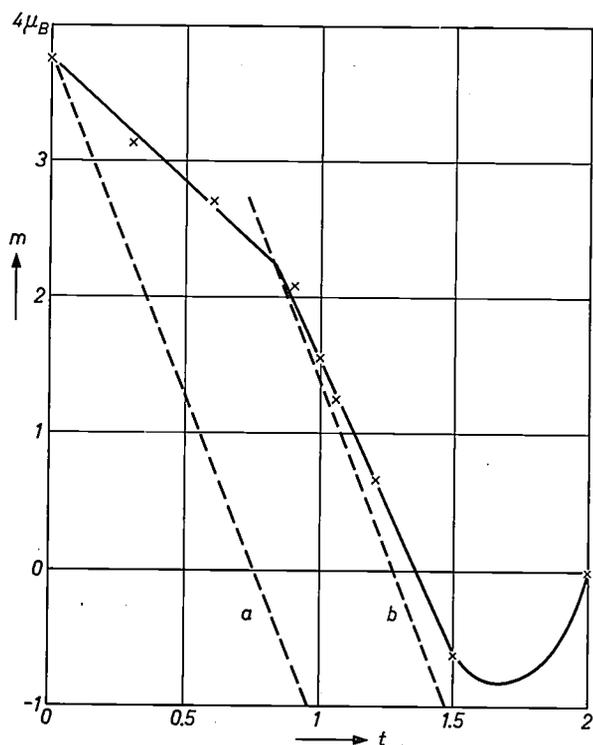


Fig. 10. Magnetic moment m of the system cobalt ferrite-rhodite ($\text{CoFe}_{2-t}\text{Rh}_t\text{O}_4$). The crosses indicate experimentally determined values. The straight line a was calculated for the distribution $\text{Fe}[\text{CoFe}_{1-t}\text{Rh}_t]\text{O}_4$ (Co^{2+} and Rh^{3+} on B sites), the line b for the distribution $\text{Co}[\text{Fe}_{2-t}\text{Rh}_t]\text{O}_4$ (all Co^{2+} ions at A sites). It is assumed that the magnetic moment of Co^{2+} at B sites is $3.75 \mu_B$ and that of Co^{2+} at A sites $3.6 \mu_B$.

that the experimental values are somewhat higher than the calculated ones may be explained in a simple way by assuming that a few of the nickel ions are located at tetrahedral sites. In the concentration region $1 < t < 2$ there must be a transition from inverse to normal distribution. Since there is then a very high concentration of diamagnetic ions at B sites, Néel's hypothesis should no longer be expected to apply: some of the paramagnetic ions on A sites are then surrounded by so many diamagnetic B ions that the A ion is no longer magnetically coupled to the other paramagnetic ions [10].

In the system $\text{CoFe}_{2-t}\text{Rh}_t\text{O}_4$ the transition from inverse to normal spinel takes place in an entirely different way. This can at once be seen by comparing the experimentally determined magnetizations with those calculated for the distribution $\text{Fe}[\text{CoFe}_{1-t}\text{Rh}_t]\text{O}_4$ (line a in fig. 10). We have also calculated the magnetizations to be expected for the distribution $\text{Co}[\text{Fe}_{2-t}\text{Rh}_t]\text{O}_4$, i.e. for a normal distribution; this calculation gives the straight line b in fig. 10. It can be seen that in quite a wide range of concentrations there is good agreement between these calculated values and those found in experiments. The conclusion can therefore be drawn that the inverse cobalt ferrite quickly becomes normal if rhodium ions are substituted for the ferric ions.

Summarizing, we arrive at the following result. The system $\text{NiFe}_{2-t}\text{Rh}_t\text{O}_4$ is inverse from $t = 0$ to $t = 1$ and becomes normal between $t = 1$ and $t = 2$. The system $\text{CoFe}_{2-t}\text{Rh}_t\text{O}_4$, on the other hand, becomes normal between $t = 0$ and $t = 1$ and remains so for $t > 1$. The Rh^{3+} ion therefore has a different effect on the $\text{Ni}^{2+}\text{-Fe}^{3+}$ cation distribution than on the $\text{Co}^{2+}\text{-Fe}^{3+}$ distribution. This demonstrates once again how complicated the question of cation distribution is.

We have observed a similar difference when the Fe^{3+} ion in nickel ferrite and cobalt ferrite was replaced by trivalent ions other than Rh^{3+} , namely Al^{3+} , V^{3+} or Cr^{3+} [2].

Not only is the variation of the magnetization as a function of composition (figs. 9 and 10) dependent on the manner in which the cation distribution goes from inverse to normal in these systems, but other physical properties also vary in different ways. In the system $\text{NiFe}_{2-t}\text{Rh}_t\text{O}_4$, for example, the Curie temperature for the composition with $t = 0$ (NiFe_2O_4) is 858°K , and for the composition with $t = 1$ (NiFeRhO_4) it is 540°K ; in the system $\text{CoFe}_{2-t}\text{Rh}_t\text{O}_4$ these values are 790 and 355°K respectively. A decrease in the Curie

[9] G. Blasse, J. Phys. Chem. Solids 27, 383, 1966 (No. 2).

[10] It can similarly be shown that Néel's hypothesis ceases to be valid at a given composition in the familiar and technically important system nickel-zinc ferrite, a transition occurring from ferrimagnetism (NiFe_2O_4) to weak antiferromagnetism (ZnFe_2O_4).

temperature is not of course unexpected, since a non-magnetic ion is substituted for a magnetic one. In the cobalt system, however, this drop is much greater than in the corresponding nickel system, which is attributable to the differences in the cation distribution.

Finally, we would like to point out that the compounds $\text{Ni}[\text{Rh}_2]\text{O}_4$ and $\text{Co}[\text{Rh}_2]\text{O}_4$ are ideal materials for study of the AA interaction in the spinel structure, containing as they do only paramagnetic ions at A sites and diamagnetic ions on B sites. The AA interaction in itself is found to be weak, but not so weak as expected [11]. Further investigations have shown that the interaction concerned is not a true AA interaction but a long range interaction in which the diamagnetic Rh^{3+} ion occupying the octahedral sites also takes part. We should therefore prefer to call it an ABA interaction. There is reason to believe that this ABA interaction has little if any significance in ferrites.

The examples discussed show that it is possible to

vary the magnetic properties of ferrites by chemical substitutions. In many cases it is possible to state that certain desired magnetic properties, such as specific values of magnetization and Curie temperature and a magnetization-temperature curve with the required shape, can be selectively imparted to a ferrite by "engineering" its composition.

Summary. The investigation of magnetic materials by the methods of crystal chemistry is illustrated by means of a number of examples relating purely to compounds with spinel structure. After a general introduction to the crystal chemistry (in particular the cation distribution) of spinels and a short account of Néel's theory of ferrimagnetic compounds, it is shown with reference to the systems $\text{NiFe}_{2-x}\text{V}_x\text{O}_4$ (nickel ferrite-vanadite) and $\text{Ni}_{1+2x}\text{Fe}_{2-3x}\text{Sb}_x\text{O}_4$ (nickel ferrite-antimonate) that it is possible to control the shape of the magnetization-temperature curve by chemical substitutions. Ferrites can be made whose magnetization is independent of temperature within a particular temperature range. This investigation also showed that ferromagnetic interactions are possible in spinels. On the basis of this result a ferromagnetic oxide with spinel structure was found. The Curie temperature of this material, however, is low (57 °K). Finally, by considering the systems $\text{NiFe}_{2-x}\text{Rh}_x\text{O}_4$ (nickel ferrite-rhodite) and $\text{CoFe}_{2-x}\text{Rh}_x\text{O}_4$ (cobalt ferrite-rhodite) it is demonstrated that a particular non-magnetic ion, Rh^{3+} , does not necessarily affect the cation distribution of different magnetic ions in an identical way.

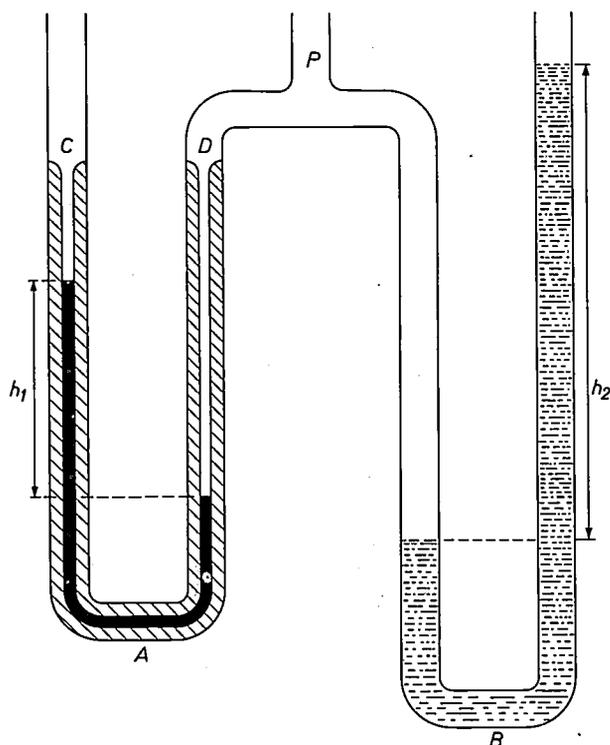
[11] G. Blasse, Philips Res. Repts. 18, 383, 1963.

Measurement of the density of small particles

In the study of solid materials a fairly accurate knowledge of the density is often necessary. For measuring the density of powders and large particles pycnometric methods are available. The density of a single small particle can be measured by immersing the particle in a liquid whose density is varied by mixing with a second liquid till the particle just remains suspended. Suitable liquids are:

$\text{CCl}_4 - \text{CBr}_4$	1.59–2.96	g/cm^3
$\text{CCl}_4 - \text{Cl}_4$	1.59–3.32	"
Thoulet's liquid (K_2HgI_4 in water)	1 –3.19	"
Rohrbach's liq. (BaHgI_4 in water)	1 –3.59	"
Clerici's liq. A (Tl malonate-formate 1 : 1 and water)	1 –4	"
Clerici's liq. B (Tl fluoride, formate and water)	1 –5.4	"

Fig. 1. Basic parts of the apparatus for density measurement. *A* is the capillary tube containing the particle and the liquid of adjustable density. The liquid is made homogeneous by bringing it into *C* or *D* by gentle suction or blowing at *P*. The U-tube *B* contains water. From h_1 and h_2 , the differences in height arising when the pressure at *P* is different from the atmospheric pressure, the density of the liquid is readily calculated.



temperature is not of course unexpected, since a non-magnetic ion is substituted for a magnetic one. In the cobalt system, however, this drop is much greater than in the corresponding nickel system, which is attributable to the differences in the cation distribution.

Finally, we would like to point out that the compounds $\text{Ni}[\text{Rh}_2]\text{O}_4$ and $\text{Co}[\text{Rh}_2]\text{O}_4$ are ideal materials for study of the AA interaction in the spinel structure, containing as they do only paramagnetic ions at A sites and diamagnetic ions on B sites. The AA interaction in itself is found to be weak, but not so weak as expected [11]. Further investigations have shown that the interaction concerned is not a true AA interaction but a long range interaction in which the diamagnetic Rh^{3+} ion occupying the octahedral sites also takes part. We should therefore prefer to call it an ABA interaction. There is reason to believe that this ABA interaction has little if any significance in ferrites.

The examples discussed show that it is possible to

vary the magnetic properties of ferrites by chemical substitutions. In many cases it is possible to state that certain desired magnetic properties, such as specific values of magnetization and Curie temperature and a magnetization-temperature curve with the required shape, can be selectively imparted to a ferrite by "engineering" its composition.

Summary. The investigation of magnetic materials by the methods of crystal chemistry is illustrated by means of a number of examples relating purely to compounds with spinel structure. After a general introduction to the crystal chemistry (in particular the cation distribution) of spinels and a short account of Néel's theory of ferrimagnetic compounds, it is shown with reference to the systems $\text{NiFe}_{2-x}\text{V}_x\text{O}_4$ (nickel ferrite-vanadite) and $\text{Ni}_{1+2x}\text{Fe}_{2-3x}\text{Sb}_x\text{O}_4$ (nickel ferrite-antimonate) that it is possible to control the shape of the magnetization-temperature curve by chemical substitutions. Ferrites can be made whose magnetization is independent of temperature within a particular temperature range. This investigation also showed that ferromagnetic interactions are possible in spinels. On the basis of this result a ferromagnetic oxide with spinel structure was found. The Curie temperature of this material, however, is low (57 °K). Finally, by considering the systems $\text{NiFe}_{2-x}\text{Rh}_x\text{O}_4$ (nickel ferrite-rhodite) and $\text{CoFe}_{2-x}\text{Rh}_x\text{O}_4$ (cobalt ferrite-rhodite) it is demonstrated that a particular non-magnetic ion, Rh^{3+} , does not necessarily affect the cation distribution of different magnetic ions in an identical way.

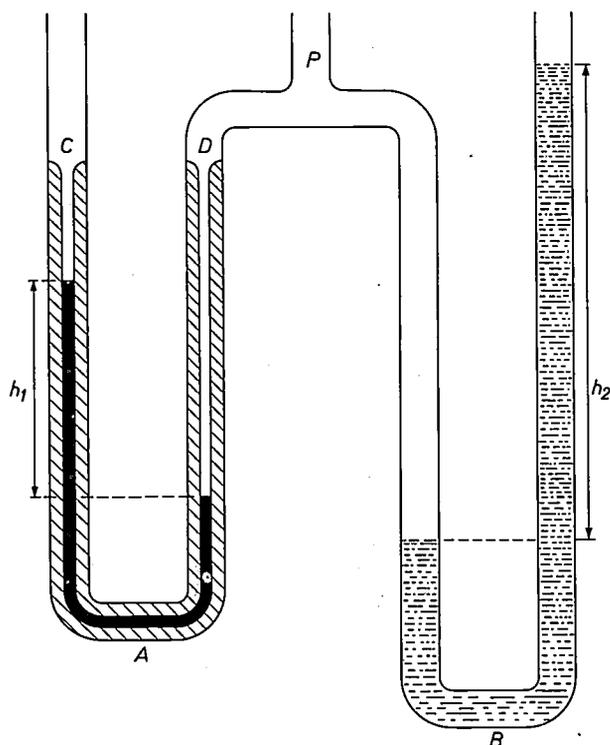
[11] G. Blasse, Philips Res. Repts. 18, 383, 1963.

Measurement of the density of small particles

In the study of solid materials a fairly accurate knowledge of the density is often necessary. For measuring the density of powders and large particles pycnometric methods are available. The density of a single small particle can be measured by immersing the particle in a liquid whose density is varied by mixing with a second liquid till the particle just remains suspended. Suitable liquids are:

$\text{CCl}_4 - \text{CBr}_4$	1.59–2.96	g/cm^3
$\text{CCl}_4 - \text{Cl}_4$	1.59–3.32	"
Thoulet's liquid (K_2HgI_4 in water)	1 –3.19	"
Rohrbach's liq. (BaHgI_4 in water)	1 –3.59	"
Clerici's liq. A (Tl malonate-formate 1 : 1 and water)	1 –4	"
Clerici's liq. B (Tl fluoride, formate and water)	1 –5.4	"

Fig. 1. Basic parts of the apparatus for density measurement. *A* is the capillary tube containing the particle and the liquid of adjustable density. The liquid is made homogeneous by bringing it into *C* or *D* by gentle suction or blowing at *P*. The U-tube *B* contains water. From h_1 and h_2 , the differences in height arising when the pressure at *P* is different from the atmospheric pressure, the density of the liquid is readily calculated.



It is seen that the method can be used for densities up to 5.4 g/cm^3 .

Although the principles of the method are well known there seems to be some reluctance in applying it, especially for the higher densities. As we have used the method quite successfully for many years almost as a routine method, it appears worth while to describe our version of the method [1].

For particles smaller than 0.2 mm , difficulties may arise from convection currents caused by temperature differences in the liquid. This problem has been solved by placing the particle and the liquid in a capillary tube of bore 1 mm or less. The use of such a capillary tube reduces the convection currents sufficiently. The particle is observed with a powerful magnifying glass or a microscope. The liquid is kept at the required temperature by surrounding the capillary tube with the vapour of a suitable boiling liquid. The density of the liquid is adjusted by addition of minute quantities of one of the components of the mixture till the particle neither rises nor sinks. Stirring is necessary for obtaining a homogeneous mixture; this is described below. The final density of the liquid can be determined very easily if a U-shaped capillary tube is used with one arm connected to a second U-tube containing water (*fig. 1*). If the pressure in the connection between the U's is made different from atmospheric pressure there will be a difference in height of the liquid level between the arms of each U. Elementary hydrostatics shows that the product of density and difference in height is equal for both U's. Hence the density is readily calculated from the observed height differences h_1 and h_2 .

Stirring of the liquid is done by repeatedly raising and lowering the pressure in the connecting tube: this brings the liquid out temporarily into the wider tubes at the ends of the capillary.

The complete apparatus is shown in *fig. 2*. Patience considerably increases the accuracy.

A simpler version of the method is used for larger particles. The whole process is performed in a small test tube (approx. 8 mm diameter) and the density of the liquid is determined by weighing it in a pipette of about 0.15 cm^3 . The pipette has to be filled rapidly in order to prevent crystallization. The liquid in the test tube is stirred with a thin glass rod, hooked round at one end, which is rotated rapidly between the fingertips.

[1] An extremely accurate but complicated version of the method has been described by P. Wulff and A. Heigl, *Z. phys. Chem. A* 153, 187, 1931.

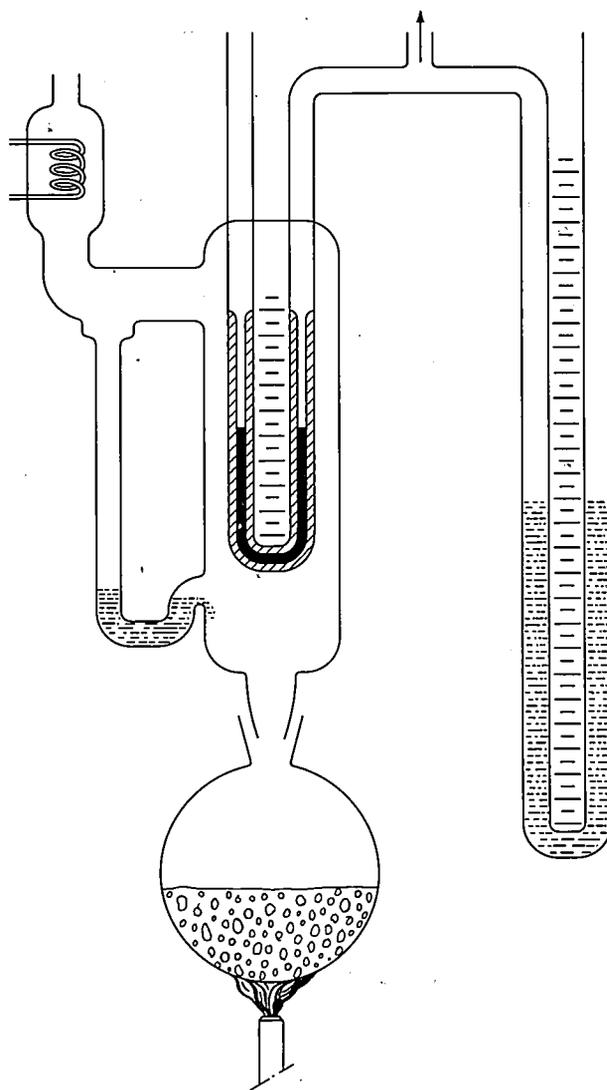


Fig. 2. Diagram of the complete apparatus for density measurement.

In both versions of the method large errors may arise from air bubbles adhering to the particle because of the poor wetting properties of the liquid. These are removed by vigorous stirring, applying vacuum, or by ultrasonic agitation.

The method has been successfully applied in measuring crystal densities used in the structure determination of ferroxdure and several other ferrites.

G. W. van Oosterhout

Drs. G. W. van Oosterhout is with Philips Research Laboratories, Eindhoven.

Recent scientific publications

These publications are contributed by staff of laboratories and plants which form part of or co-operate with enterprises of the Philips group of companies, particularly by staff of the following research laboratories:

Philips Research Laboratories, Eindhoven, Netherlands	<i>E</i>
Mullard Research Laboratories, Redhill (Surrey), England	<i>M</i>
Laboratoires d'Electronique et de Physique Appliquée, Limeil-Brévannes (S.O.), France	<i>L</i>
Philips Zentrallaboratorium GmbH, Aachen laboratory, Weisshausstrasse, 51 Aachen, Germany	<i>A</i>
Philips Zentrallaboratorium GmbH, Hamburg laboratory, Vogt-Kölln-Strasse 30, 2 Hamburg-Stellingen, Germany	<i>H</i>
MBLE Laboratoire de Recherche, 2 avenue Van Becelaere, Brussels 17 (Boitsfort), Belgium.	<i>B</i>

Reprints of most of these publications will be available in the near future. Requests for reprints should be addressed to the respective laboratories (see the code letter) or to Philips Research Laboratories, Eindhoven, Netherlands.

- N. W. H. Addink:** Suppression of halo on photographic plates used in mass spectrometry. *Nature* **211**, 1168-1169, 1966 (No. 5054). *E*
- K. H. Beckmann:** On the chemical composition of surface films produced on germanium in different etchants. *Surface Sci.* **5**, 187-196, 1966 (No. 2). *H*
- G. Blasse, A. Brill & W. C. Nieuwpoort:** On the Eu^{3+} fluorescence in mixed metal oxides, Part I. The crystal structure sensitivity of the intensity ratio of electric and magnetic dipole emission. *J. Phys. Chem. Solids* **27**, 1587-1592, 1966 (No. 10). *E*
- A. J. Burggraaf:** Sorption of argon by glass in a gas-discharge lamp. *J. Amer. Ceramic Soc.* **49**, 450-454, 1966 (No. 8).
- W. F. Druyvesteyn & C. A. A. J. Greebe:** Helicon-like resonances in superconducting indium plates. *Physics Letters* **22**, 17-18, 1966 (No. 1). *E*
- W. F. Druyvesteyn, A. K. Niessen & F. A. Staas:** The resistive behaviour of a type II superconductor on reversing the current. *Physics Letters* **22**, 127-128, 1966 (No. 2). *E*
- P. Eckerlin & J. Liebertz:** Darstellung und kristallographische Daten von $\text{Bi}_2\text{Al}_4\text{O}_9$ -Einkristallen. *Naturwiss.* **52**, 450, 1965 (No. 15). *A*
- P. J. Flanders:** Metamagnetic effects in hematite. *Phil. Mag.* **14**, 1-6, 1966 (No. 127). *M*
- R. W. Gibson:** Cadmium sulphide ultrasonic transducers. *Electronics Letters* **2**, 213, 1966 (No. 6). *M*
- A. A. van der Giessen:** The structure of iron (III) oxide-hydrate gels. *J. inorg. nucl. Chem.* **28**, 2155-2159, 1966 (No. 10). *E*
- T. Groot, P. C. M. N. Bruijs & J. H. T. C. Verbeek:** X-ray fluorescence analysis of heavy elements in a light matrix. *Nature* **211**, 1085, 1966 (No. 5053). *E*
- G. J. van Gorp:** Superconducting fluxline pinning by twin boundaries. *Phys. Stat. sol.* **17**, K 135-137, 1966 (No. 1). *E*
- M. K. McPhun:** Measurement of negative resistance using a Z-g diagram. *Proc. IEEE* **54**, 910-911, 1966 (No. 6). *M*
- J. Neirynek & Ph. van Bastelaer:** Distortion analysis of narrow-band carrier telegraph systems. *Rev. HF* **6**, 363-374, 1966 (No. 11). *B*
- A. van Oostrom:** Experimental evidence supporting chemisorption without surface rearrangement for nitrogen on tungsten at and below room temperature. *Physics Letters* **22**, 137-139, 1966 (No. 2). *E*
- E. Schwartz:** Wechsel der Bezugsimpedanzen bei Streu- oder Betriebsmatrizen. *Archiv elektr. Übertr.* **20**, 357-364, 1966 (No. 7). *A*
- P. T. Squire & J. W. Orton:** Relaxation of the Cr^{3+} ion in emerald. *Proc. Phys. Soc.* **88**, 649-657, 1966 (No. 3). *M*
- C. van de Stolpe:** Preparation of NiO single crystals by chemical transport. *J. Phys. Chem. Solids* **27**, 1952-1953, 1966 (No. 11/12). *E*
- J. van der Veen, J. Helfferich & L. K. H. van Beek:** Photo isomerization of methoxybenzene diazosulfonates. *Rec. Trav. chim. Pays-Bas* **85**, 895-898, 1966 (No. 8). *E*
- J. H. N. van Vucht:** Influence of radius ratio on the structure of intermetallic compounds of the AB_3 type. *J. less-common Met.* **11**, 308-322, 1966 (No. 5). *E*

The Mössbauer effect and its application in solid-state research

J. S. van Wieringen

Since Mössbauer's discovery ten years ago of the effect named after him, for which he received the Nobel prize, the main application of the effect has been in solid-state research. It has proved particularly useful for investigating the structure of magnetic materials that contain iron.

The Mössbauer effect is based on a resonance effect, whence the alternative name for this phenomenon: gamma-ray resonance absorption. Resonance is a familiar phenomenon in all fields of physics and technology involving vibrations and waves, such as sound waves, surface waves and electromagnetic waves. Use is made of electromagnetic resonance, for example, for tuning radio receivers, and resonance has also long been used in optics. It accounts for the presence of the Fraunhofer lines in the solar spectrum. The inner mantle of the sun emits a continuous spectrum. Atoms in the outer regions of the sun's atmosphere resonate with this radiation at their own wavelength and thus absorb the light emitted by deeper regions at that particular wavelength. In this case, the light source has a continuous spectrum and the absorber a line spectrum. The effect becomes even more striking, however, if both source and absorber have one and the same frequency. If, for example, one looks at a sodium lamp through a flame made yellow by sodium, the flame is seen to stand out darkly against the bright background of the lamp. When the radiation from an ultraviolet mercury lamp is directed over a mercury surface onto a fluorescent screen, dark shadows are seen on the screen of the vapour rising from the mercury surface. Both in the case of the sodium flame and that of the mercury vapour the effects observed are due to the fact that the flame and the vapour are able to absorb the incident light by resonance.

Since the concept of resonance was well known in the case of long-wave electromagnetic radiation, it is not surprising that attempts were also made to demonstrate the effect for short-wave gamma radiation (see

fig. 1a). The source used for this purpose was a group of radioactive nuclei which, in the last stage of their disintegration, changed from the excited state *A* to the ground state *G* of a stable nucleus, emitting gamma radiation in the process. When this radiation is incident on a substance containing the stable nucleus, one would expect at first sight that the radiation would be absorbed by a resonance effect in which the stable nucleus changes to the excited state *A* (see *fig. 1b*). This is not observed, however. Closer consideration shows

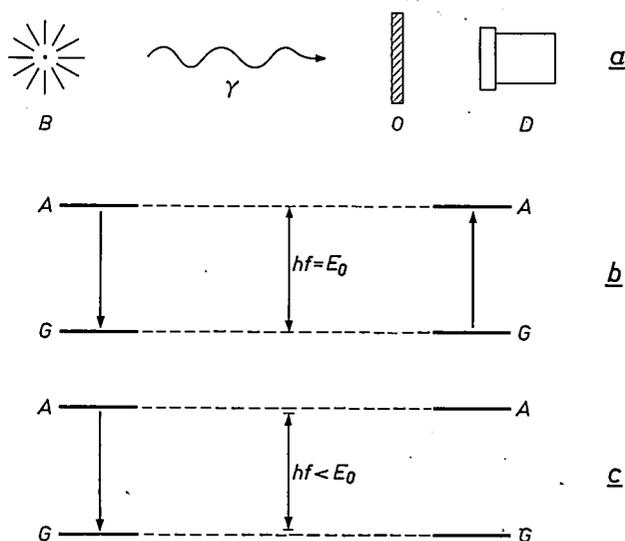


Fig. 1. *a*) Diagram illustrating the observation of gamma-ray resonance absorption. The source *B* emits gamma quanta γ , which impinge on an absorber *O*. The transmission is measured by detector *D*.

b) Phenomenon of resonance absorption. A nucleus in the source makes a transition from the excited level *A* to the ground level *G*, emitting a gamma quantum of energy $hf = E_0$. This is absorbed in the absorber *O* because a nucleus in it is excited from *G* to *A*. *c*) As in (*b*) but taking account of the recoil in *B* and in *O* due to the momentum of the gamma quantum. The energy of the quantum is now too low to be able to excite a nucleus in *O*. There is therefore no resonance absorption.

that this is quite understandable having regard to the changes of energy and momentum that occur when a gamma quantum is emitted by an atomic nucleus. In the initial state we have here an excited nucleus with energy E_0 and zero momentum. In the final state the energy of the emitted gamma quantum is hf ; this implies a momentum of hf/c (h being Planck's constant, f the frequency and c the velocity of light). Where the atoms and molecules are free, i.e. in a gaseous source, this momentum is balanced by the recoil of the emitting nucleus (mass M) which thereby acquires a velocity $v = hf/Mc$, representing an energy $\frac{1}{2}Mv^2 = h^2f^2/2Mc^2$. It follows from the law of conservation of energy:

$$hf = E_0 - (h^2f^2/2Mc^2). \quad (1)$$

The energy of the emitted gamma quantum hf is therefore lower than the energy difference E_0 between the ground state and excited state of the nucleus, part of this energy difference being used for imparting to the nucleus the recoil velocity $v = hf/Mc$. It is no wonder, then, that a quantum of this energy is unable to raise a nucleus in the absorber from the ground state G to the excited state A , particularly since in this excitation the momentum of the gamma quantum would again have to be transferred to the absorbing nucleus, once more at the expense of the available energy (see fig. 1c).

These considerations also apply to the optical region. There, however, the recoil is much weaker, owing to the much lower frequency f , and the resultant energy loss remains well within the normal width of the spectral line; see fig. 2a. Thus, although the spectral lines of both the source and the absorber are slightly displaced, they overlap each other so much that the net result is as shown in fig. 1b, and a readily observable resonance absorption is found. This should be compared with the relative situation of the spectral lines for the gamma radiation in fig. 2b.

In order to find resonance absorption for gamma radiation it is therefore necessary to avoid the energy loss due to the recoil. A similar idea is encountered in ballistics. Guns in permanent stations are often fixed rigidly to their heavy emplacement structure (see fig. 3). The recoil in this case is not taken up by the gun alone but by the much heavier mass of the emplacement so that, according to equation (1), the recoil energy decreases and the range of the projectile is increased slightly (by a few percent). In the same way, when gamma rays are emitted and absorbed by atomic nuclei the energy loss due to recoil can be eliminated or, rather, sufficiently reduced, by anchoring the nuclei to a heavy mass, viz, by incorporating them in the crystal lattice of a solid. In this way gamma-ray resonance absorption was found for the first time by R. L. Mössbauer in

1957 [1], between iridium nuclei in a solid source and a solid absorber; see fig. 2c.

Although the foregoing presents the essentials of gamma-ray resonance absorption, the situation in practice is more complicated because the nucleus is not perfectly rigidly bound to the crystal lattice. The bond is not so weak that the nucleus is dislodged by the recoil from its lattice site; the gamma quanta involved are too soft for that. The recoil momentum is therefore absorbed by the whole lattice. It is, however, possible that recoil energy is lost for the gamma radiation because vibrations are generated in the not completely rigid crystal lattice which are ultimately dissipated as heat.

The theoretical treatment of the effect referred to is not simple, because it involves the coupling of two quantized systems: the nucleus, which effects a transition between two energy levels while emitting a gamma quantum; and the lattice vibrations, which are excited during this emission via the mechanical coupling. The principal properties of the emitted or absorbed radia-

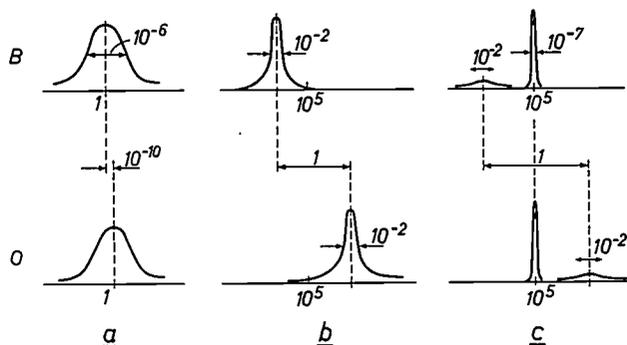


Fig. 2. Energy spectrum of the radiation emitted by the source B (above), and the spectrum required by absorber O for resonance absorption (below), sketched for three cases. The figures indicate the orders of magnitude of energies and energy differences in electron-volts.

a) Optical spectrum. The line width is much greater than the shift resulting from recoil. Resonance absorption is therefore always possible.

b) Gamma radiation in a gas. The energy of the quanta is about 10^6 times as large as in the optical case; the recoil energy therefore is about 10^{10} times larger. The line width is now so much smaller than the shift due to recoil that the resonance absorption is negligibly small.

c) Gamma radiation for solid-state source and absorber. Since a part of the transitions, called the Mössbauer fraction, is not shifted by recoil, distinct resonance absorption is possible.

tion can, however, also be derived in a relatively simple way from a classical model. The complete quantum-mechanical calculation leads to the same conclusions. We may summarize these results as follows:

- A fraction of the gamma transitions are resonant, that is to say for these transitions the energy E_0 is entirely converted into gamma radiation of frequency $f_0 = E_0/h$.
- The spectral line associated with these transitions

is not broadened by the thermal agitation of the nucleus, but its width is determined solely by the "natural line width", which is determined by the finite lifetime of the excited state A .

c) The relative intensity of the resonant transitions, i.e.

The complete calculation shows that $\overline{x^2}$ can be found by taking the average of x^2 over the ordinary thermal movement of a lattice point [2]. Carrying out this averaging for the Debye model of lattice vibrations [3] leads to a gamma spectrum as shown in fig. 2c.

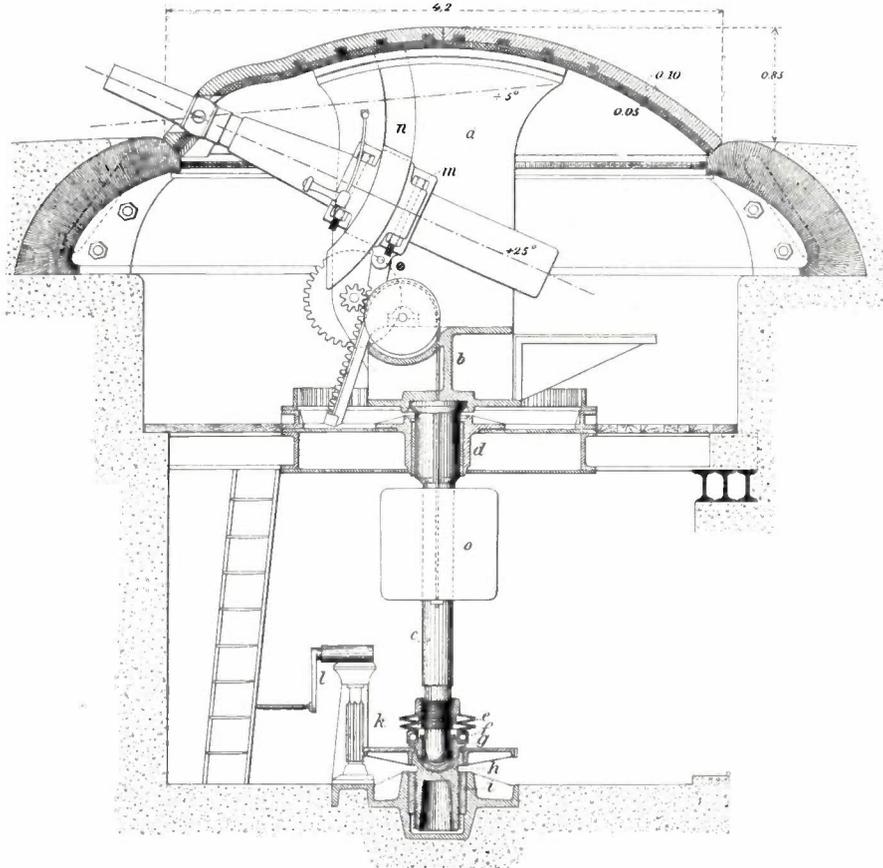


Fig. 3. Suppression of recoil losses by anchoring the source to a heavy mass. In ballistics this is well known for the case of fixed ordnance mountings, like this 15 cm gun in an armoured emplacement dating from 1900, one of which is still to be seen in the fort at Velsen, in the Netherlands. (From: L. J. Spanjaardt Speckman, *Duurzame versterkingskunst*, Breda 1934.)

the so-called Mössbauer fraction, is given by:

$$g = \exp(-4\pi^2\overline{x^2}/\lambda^2), \quad \dots \dots \dots (2)$$

where $\overline{x^2}$ represents the mean square displacement of the nucleus due to thermal motion. Hence, the intensity is greater the less the nucleus moves and the longer the wavelength λ of the gamma radiation.

d) The other transitions, where the gamma frequency differs from f_0 due to interaction with the lattice vibrations, are considerably broadened by the thermal movement of the nucleus.

A brief outline may be given here of the classical derivation of the above properties (a)-(d).

The nucleus is regarded as a radiation source of frequency f . Since the nucleus is not stationary but moves with a velocity $v(t)$ as a result of the lattice vibrations, the emitted electromagnetic wave shows a Doppler shift:

$$f(t) = f_0[1 + v(t)/c].$$

The time-dependence of the electromagnetic field is therefore not given by a factor

$$\exp(2\pi j f_0 t),$$

as for a stationary source, but by

$$h(t) = \exp[2\pi j \int_0^t f(t') dt'] = \exp(2\pi j f_0 t) \exp[2\pi j x(t)/\lambda],$$

where $x(t)$ is the displacement of the source from the equilibrium position and λ is the (Doppler-unshifted) wavelength of the gamma radiation.

[1] R. L. Mössbauer, *Z. Physik* **151**, 124-143, 1958.
 [2] This is not self-evident, since it is not *a priori* possible to say that the thermal movement of the nucleus in the lattice is unaffected by the emission or absorption of a gamma quantum.
 [3] See, for example, J. Volger, *Solid-state research at low temperatures*, III, Philips tech. Rev. **22**, 268-277, 1960/61.

We will first consider the case where the nucleus describes a simple harmonic oscillation of amplitude x_0 and frequency F :

$$x = x_0 \sin 2\pi Ft.$$

The time-dependence of the emitted electromagnetic wave is then:

$$h(t) = \exp(2\pi j f_0 t) \exp[2\pi j(x_0/\lambda) \sin 2\pi Ft].$$

We wish to know the frequency spectrum of this expression, in other words we wish to express $h(t)$ as a sum of simple harmonic terms. The second factor can be expanded in powers of $(x_0/\lambda) \sin 2\pi Ft$:

$$\exp[2\pi j(x_0/\lambda) \sin 2\pi Ft] = 1 + 2\pi j(x_0/\lambda) \sin 2\pi Ft - 2\pi^2(x_0/\lambda)^2 \sin^2 2\pi Ft + \dots$$

Consequently,

$$h(t) = \exp(2\pi j f_0 t) + \pi(x_0/\lambda) [\exp(2\pi j Ft) - \exp(-2\pi j Ft)] \exp(2\pi j f_0 t) - \frac{1}{2}\pi^2(x_0/\lambda)^2 [2 - \exp(4\pi j Ft) - \exp(-4\pi j Ft)] \exp(2\pi j f_0 t) + \dots = \exp(2\pi j f_0 t) [1 - \pi^2(x_0/\lambda)^2 + \dots] + \dots \quad (3)$$

It is seen, then, that, in addition to a term with frequency f_0 , sidebands also occur with frequencies $(f_0 \pm F)$, $(f_0 \pm 2F)$, and so on. The first is the one that gives rise to resonance absorption. As long as the amplitude of the thermal oscillation x_0 is small compared with the wavelength λ of the gamma radiation, the relative intensity of the central absorption line is found by taking the square of the coefficient of $\exp(2\pi j f_0 t)$ in (3), i.e. $[1 - (2\pi^2 x_0^2/\lambda^2)]$.

This treatment can easily be extended to the general case of random movement of the source. The position of the source can then be expressed as a Fourier series:

$$x = \sum_{n=1}^N x_n \sin(2\pi n Ft + \varphi_n),$$

where nF is the frequency and φ_n the phase of the lattice waves. The time-dependence of the electromagnetic wave is now:

$$h(t) = \exp(2\pi j f_0 t) \sum_{n=1}^N \exp[2\pi j(x_n/\lambda) \sin(2\pi n Ft + \varphi_n)].$$

If we again expand the second factor, we see first of all that the sidebands have become very much broader, owing to the occurrence of various frequencies nF instead of simply F , and because the bands themselves intermingle. The movement of the nucleus due to the lattice vibrations therefore results mainly in a broadening of the sidebands. The line broadening does not occur, however, for the term with frequency f_0 . This has now an intensity:

$$g \approx \prod_{n=1}^N [1 - (2\pi^2 x_n^2/\lambda^2)].$$

This can be simplified by writing:

$$\ln g \approx \sum_{n=1}^N \ln [1 - (2\pi^2 x_n^2/\lambda^2)] \approx -4\pi^2 \overline{x^2}/\lambda^2,$$

where

$$\overline{x^2} \equiv \frac{1}{2} \sum x_n^2$$

represents the mean nuclear displacement caused by the thermal agitation. The relative intensity of the resonance radiation is therefore given by eq. (2) above.

Conditions for observation of the Mössbauer effect

As we have seen, the Mössbauer effect is based on the fact that for some of the gamma transitions in the

nuclei of both source and absorber no lattice vibrations are involved. The probability g of these transitions is given by eq. (2). If the effect is to be observable, many such transitions must occur in both source and absorber. It has already been indicated that g is higher the less the thermal agitation — this implies holding source and absorber at low temperature — and the softer the gamma radiation, i.e. the longer its wavelength. The nuclear transitions for which the Mössbauer effect has hitherto been observed (see fig. 4) therefore all have a gamma energy E_0 of less than about 150 keV. In order to observe the effect at the higher energies in this range (above about 50 keV) it is necessary to resort to temperatures in the neighbourhood of the boiling point of helium.

The application of the Mössbauer effect in solid-state research to be discussed below is based on the fact that

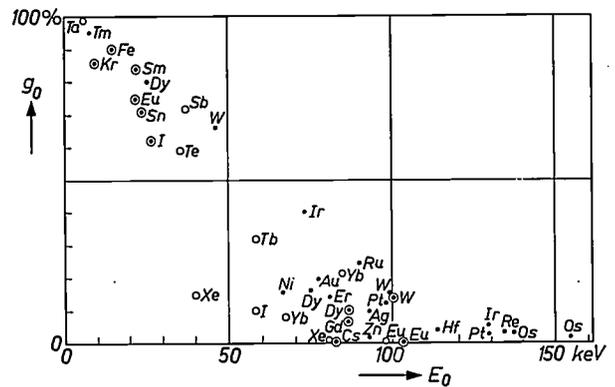


Fig. 4. Mössbauer fraction g_0 and quantum energy E_0 of the gamma radiation for the nuclides on which the Mössbauer effect has hitherto been found: the value of g_0 applies at 0 °K when the nuclide is incorporated in its own chemical element. The nuclides denoted by a circle are obtained in the excited state by the disintegration of artificial radioactive nuclei having a lifetime of at least 30 days; the nuclides denoted by a dotted circle have in addition a natural line width smaller than 1 mm/s. (The reason for giving the line width in mm/s is explained below.)

it provides the possibility of measuring the very fine splitting of nuclear energy levels caused by electric and magnetic fields at the site of a nucleus. In many cases the Mössbauer effect represents the only way in which this fine structure of energy levels can be measured. Nuclear-spin resonance can also be used for this type of investigation; the pros and cons of both methods will be discussed presently.

Gamma-ray resonance absorption as a function of frequency is known as the Mössbauer spectrum, by analogy with the optical case. If the relevant nuclear levels exhibit electrical or magnetic splitting, fine structure appears in the Mössbauer spectrum. This is observable only if the width of the resonance line (as mentioned, determined by the lifetime of the excited state) is not too great. This requirement limits the permissible

lifetime of the excited level; in practice it should not be shorter than about 10^{-9} seconds.

Finally, other practical limitations are imposed by the availability of a suitable source. The source should contain nuclei of the relevant nuclide in the excited state A . In some cases it has proved possible to produce these in sufficient numbers by exciting nuclei in the ground state by means of accelerated protons or other heavy particles [4]. As yet, however, this finds only limited application because of the elaborate equipment required. It is far more convenient to make use of artificial radioactive nuclei which upon disintegration undergo a transition to an excited state A followed by a transition to the ground state G of the required nuclide. The lifetime of the initial radioactive nucleus should not be too short. In a laboratory not in the immediate neighbourhood of a reactor or cyclotron producing the nuclei, a half-life of about 30 days is a desirable minimum. It should be remembered in this connection that the starting nuclide often has to undergo various chemical treatments before it can be used as a Mössbauer source. As appears from fig. 4, all these requirements are met by nuclides of the elements Eu, Fe, I, Kr, Sm and Sn, and — at liquid helium temperature — Cs, Dy, Eu, Gd, Yb and W.

Amongst these nuclides there is one which is well-nigh ideal for Mössbauer experiments, and moreover it is found in many materials of technological and scientific interest. This is ^{57}Fe , an isotope present to the extent of 2.2% in natural iron. The gamma energy is only 14.4 keV; according to (2) this leads to a relatively large g which makes measurements possible in a wide temperature range (up to about 1000 °C). The lifetime of the excited state is exceptionally long, being 9.8×10^{-8} s. The source employed is ^{57}Co , which is fairly easy to make in a cyclotron by the reaction $^{56}\text{Fe}(d,n)^{57}\text{Co}$ or $^{60}\text{Ni}(p,\alpha)^{57}\text{Co}$, and has the very favourable half-life of 267 days. The disintegration scheme is shown in fig. 5.

Apparatus for measuring Mössbauer spectra

Mössbauer spectra are measured by means of an absorption technique as commonly used in other forms of spectroscopy. The radiation transmitted through the sample containing ^{57}Fe nuclei is measured as a function of the frequency f of the incident radiation [5]. The source of the incident radiation is ^{57}Co , incorporated

in a substance such that a single narrow line — the Mössbauer transition — at 14.4 keV is obtained in the disintegration process. In order to be able to scan the spectrum it should be possible to vary the frequency of this gamma radiation continuously. Although at first sight it may seem difficult to vary the frequency of radioactive radiation, it can nevertheless be varied sufficiently for Mössbauer measurements, simply by *moving* the source. Denoting the speed of the source in the direction of the absorber by v , the frequency for an observer stationary with respect to the absorber is $f = f_0(1 + v/c)$ because of the Doppler effect. For the moderate speeds feasible in the laboratory the frequency variation attainable is of course limited but it is sufficient in view of the exceptionally small width of the Mössbauer lines (see fig. 2c). The fine structure to be measured is also very small. For ^{57}Fe , all spectra can be scanned with a speed range of 0-10 mm/s.

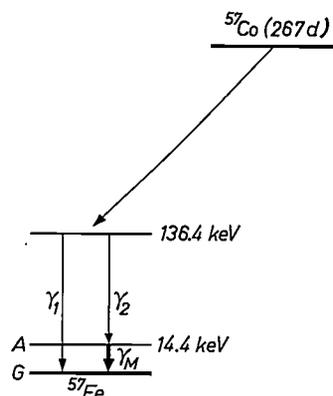


Fig. 5. Disintegration scheme of ^{57}Co , used as initial nuclide for Mössbauer measurements with ^{57}Fe . By capture of K-shell electrons by the Co nucleus (so-called K capture) the ^{57}Co changes to an excited nucleus of ^{57}Fe . Its energy level is 136.4 keV above the ground state G of ^{57}Fe . The transition to G takes place by the emission of gamma radiation, either γ_1 or $\gamma_2 + \gamma_M$. The second of these possibilities takes place via the lower excited state A and gives the gamma radiation γ_M of 14.4 keV required for the Mössbauer measurements.

Absorption due to the Mössbauer effect is very weak, and therefore prolonged measurement is necessary to achieve sufficient accuracy. During this time the source travels a long distance, and for practical reasons the distance is split up into numerous small intervals, the source being arranged to move to and fro. It can, for example, be moved at the uniform velocity $+v$ towards the absorber and then back again at the same velocity $(-v)$, and so on. The radiation transmitted by the absorber is measured and summed separately for the two directions of movement: in this way two points of the absorption spectrum corresponding to $+v$ and $-v$ respectively are determined simultaneously. The measurement is then repeated at a different speed v , and so on.

[4] S. L. Ruby and R. E. Holland, Phys. Rev. Letters 14, 591, 1965.

[5] Instead of absorption, emission too may be measured, but this is less frequently done because in that case every sample has to be radioactive and must be prepared by radiochemical means. It is also possible to measure the radiation scattered by the absorber, instead of the absorption. These measurements are less sensitive, however, and are seldom performed.

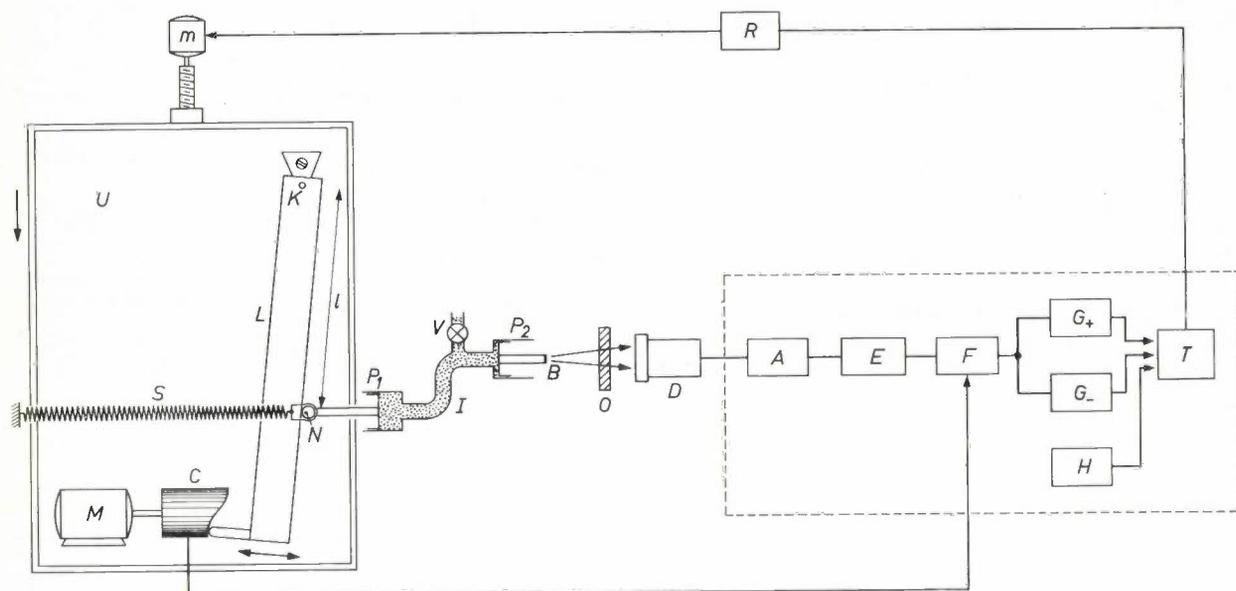


Fig. 6. Diagram illustrating the experimental set-up for Mössbauer measurements on ^{57}Fe . On the left is the mechanical part mounted on a carriage U . Via a cam C the motor M gives a reciprocating movement to the end of lever L , which pivots about point K . The pin N , at an adjustable distance l from K , is held against the lever L by the spring S and transmits its reciprocating movement to the source B by means of two pistons P_1 and P_2 and an oil line I , the latter being interposed to prevent vibrations being transmitted from M to B . The oil line I can be filled without air bubbles by means of the valve V . That part of the gamma radiation from B passing through the absorber O produces pulses in a counter tube D . These pulses are further processed in the electronic part of the equipment (dotted rectangle on the right). The pulses are amplified in A , and the 14.4 keV pulses are sorted in discriminator E and separated by circuit F , into pulses corresponding to positive and negative source speed. The counted numbers are stored in stores G_+ and G_- . After a preselected time, set by clock H , the numbers stored in G_+ and G_- are typed out by T . After this has been done, the measurement is automatically continued at a different source speed (i.e. a different value of l). For this purpose the carriage U is shifted in the direction of the arrow by the motor m , operated by circuit R .

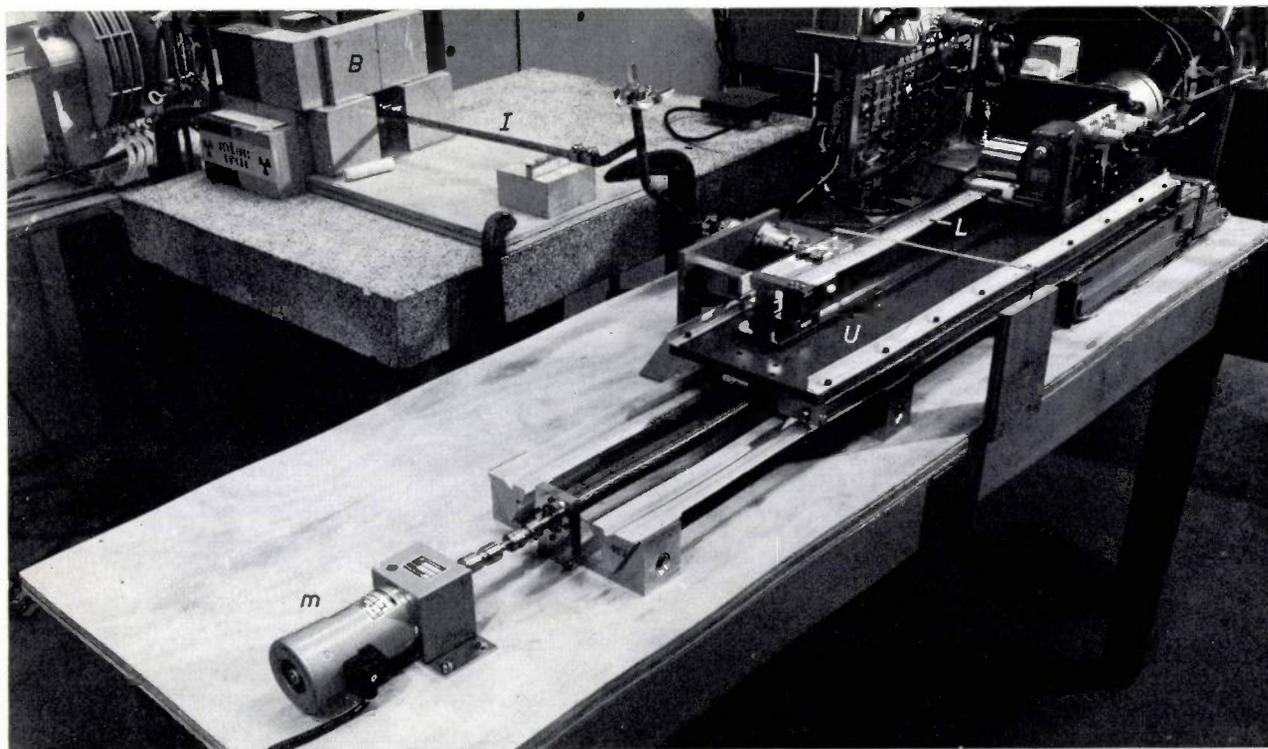


Fig. 7. View of the set-up for Mössbauer measurements in the Philips laboratory at Eindhoven [6]. The letters correspond to those in fig. 6.

In another frequently used method the source is moved to and fro not at uniform speed, but the speed is varied continuously between zero and a maximum value, for instance by giving the source a simple harmonic motion. The radiation intensity transmitted by the absorber is sorted according to speed in a multi-channel analyser (kick sorter), which also provides for the summation of the transmitted intensities for each speed and each direction of the source.

Both methods of measurement have their advantages. A drawback of the first method is the considerable time that elapses between the measurement at the first and last speeds: the method is therefore sensitive to any slow changes in the apparatus used to detect the radiation. The second method does not have this drawback, but it is dearer and does not have the facility that a particular part of the spectrum can be measured separately. Even if one is interested in only a limited part of the spectrum, the second method necessarily involves traversing the whole range of the spectrometer, which needlessly lengthens the time of the measurements.

In the spectrometer constructed in the Philips Research Laboratories at Eindhoven [6] the source is moved by a simple mechanism, as follows. A 40 cm long lever L is pivoted at K (see fig. 6 and 7). By means of a cam C and a synchronous motor M the other end of the lever is given a uniform motion to and fro over a distance of 13 mm, the speed being $+10$ mm/s and -10 mm/s respectively. The speed of any point N on the lever clearly lies between 0 and 10 mm/s, depending on its distance l from the pivot K . This movement is transmitted to the ^{57}Co source B by two pistons in cylinders P_1 and P_2 and an oil line I , whose purpose is to damp out mechanical vibrations. These vibrations would reduce the resolution of the spectrometer if they were to be transmitted to the source.

The pulses produced in the counter tube D by the gamma rays transmitted by the absorber O are further processed in the electronic circuits shown on the right in fig. 6. After amplification in the pre-amplifier A , the pulses are fed to the discriminator E which passes only those originating from the 14.4 keV γ_M radiation (see fig. 5), the pulses from the other gamma radiation (e.g. γ_1 and γ_2 in fig. 5) being suppressed. In the unit F , operated by contacts on the cam C , the γ_M pulses are sorted according to positive source speed $+v$ and negative source speed $-v$. In each channel the pulses are added and their number stored in stores G_+ and G_- until the timer clock H indicates that the measuring time for speed $\pm v$ is up. The numbers collected in G_+ and G_- are then read and printed out by a printer T .

For the following measurement at the speed $\pm v'$ the distance l has to be changed. This is done automatically, the whole cam-lever mechanism $M-C-L$ in fig. 6 being mounted on a carriage U which is displaced by a motor m in the direction of the arrow. When U is moved, N , I and B remain stationary.

Finally, the result of the measurement is obtained as

a series of pairs of numbers, indicating the intensity of the gamma radiation transmitted through the absorber as a function of the speeds $+v$ and $-v$. These numbers plotted versus v give the Mössbauer spectrum required. Because of this method, the positions and widths of the spectral lines are usually given in velocity units.

Mössbauer spectrum of iron nuclei

In ^{57}Fe the ground state G and the excited state A are, in fact, multiplet states. By virtue of the magnetic moment of this nucleus, the ground state splits into two levels under the influence of a magnetic field H (see fig. 8b). For the same reason the excited state

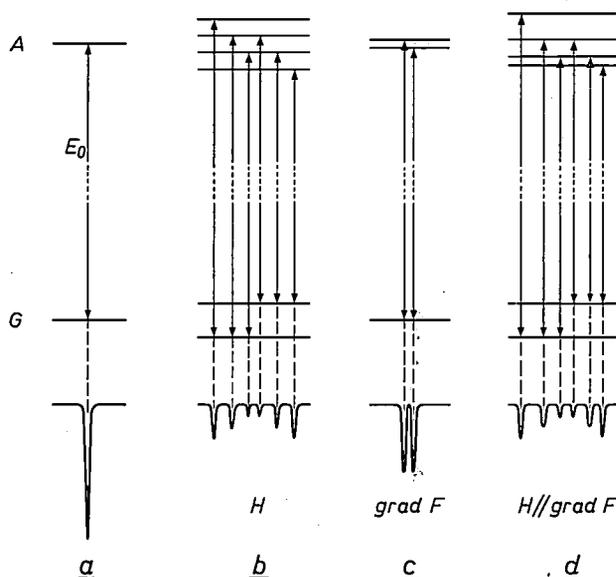


Fig. 8. Splitting of ground state G and excited state A of the nuclide ^{57}Fe when subjected to different magnetic and electric fields. Below: the corresponding spectra, i.e. transmission as a function of frequency (or source speed).

a) Degenerate states in the free nucleus; the energy difference between A and G is $E_0 = 14.4$ keV.

b) In a magnetic field H the ground state G splits into two levels and the excited state A into four levels at identical distances apart. In ferromagnetic substances local fields up to about 500 000 gauss occur. The corresponding splitting is of the order of 10^{-9} keV. The absorption spectrum in a polycrystalline sample consists of six lines and is symmetrical.

c) In a non-uniform electric field F the ground state G does not split, but A splits into two levels, proportional to $\text{grad } F$. In solids splitting up to about 10^{-10} keV may occur.

d) In the presence of both a magnetic field and a non-uniform electric field, the excited state A splits asymmetrically. In the case shown, H and $\text{grad } F$ are parallel. The spectrum then consists again of six lines, as in (b), but is now asymmetrical.

splits into four levels in a magnetic field. The nucleus in the state A also has an electric quadrupole moment; this will cause partial splitting into two levels in a non-uniform electric field F , that is to say an electric field with a gradient (fig. 8c). If the nucleus is subjected to

[6] This apparatus has been developed from that described by J. J. van Loef in Ned. T. Natuurk. 30, 293, 1964.

both a magnetic field and a non-uniform electric field, the splitting of A becomes more complex. The simplest case — when H and $\text{grad } F$ are parallel to one another — is represented in fig. 8*d*. The spectra corresponding to these level schemes are also indicated in fig. 8.

We have seen that in a magnetic field, G has two levels and A four and one would therefore expect eight transitions; in fact, however, there are only six in figs. 8*b* and *d*, the other two being “forbidden”. If the beam of gamma rays that excites the nuclei is parallel to the field H , a further two transitions are forbidden. All eight transitions can take place only when the nucleus is situated in a magnetic field H and in a nonuniform electric field F whose gradient is *not* parallel to H .

Apart from H and $\text{grad } F$, the spectrum is also sensitive to the presence of a space charge at the site of the nucleus. The energy difference E_0 of a given transition and hence the position of the absorption line (or, if a multiplet, the centre of gravity) depends to some extent on the space charge because in state A the nucleus is somewhat smaller than in state G and the energy of interaction with an electric space charge is therefore not entirely the same for A and G . The difference is proportional to the space-charge density which in turn depends on the substance in which the iron is incorporated.

Summarizing, we see that the Mössbauer spectra of ^{57}Fe can be influenced in the following way by the substance in which the iron is incorporated.

- 1) The centre of gravity of the spectrum depends on the effective charge density at the site of the iron nucleus; in ionic compounds of iron, for example, there is a distinct difference between bivalent and trivalent Fe.
- 2) A non-uniform electric field at the site of the iron core gives rise to quadrupole splitting; hence, the presence of such a splitting shows that the local symmetry is lower than cubic.
- 3) A magnetic field at the site of the nucleus also betrays its presence by splitting the absorption line into a multiplet.

In nuclear-spin resonance measurements^[7] only the splitting of the ground state G can be determined. This means in the case of ^{57}Fe that only a magnetic field H can influence the spectrum; no information is obtained about the charge density or $\text{grad } F$. Moreover, nuclear-spin resonance measurements are less sensitive and they are difficult to carry out if H (and hence the splitting of G) can assume widely different values. On the other hand, the resolution is not limited by the lifetime of the excited level A , as it is in Mössbauer effect measurements, so that the measurement can be more accurate. Furthermore, nuclear-spin resonance is applicable in the case of many more nuclides.

The Mössbauer spectra are usually measured on substances with a high iron content. There is then a strong exchange interaction between iron atoms on corresponding lattice sites. The consequence is that the spectrum is determined by the *average* state of all these atoms, not just those that happen to suffer resonance absorption. As far as the effects mentioned under (1) and (2) are concerned this is generally of no consequence since the electric charge and the electric field are identical for atoms on corresponding sites. This does not, however, apply to the magnetic field. The magnetic field at the site of the nucleus of an iron atom or ion is mainly governed by the magnetic moment of the electrons of the atom or ion. In case of strong interaction with iron on corresponding lattice sites this electron magnetic moment constantly changes direction. It follows from the foregoing, therefore, that the splitting measured is then proportional to the average magnetic moment of the group of corresponding atoms or ions. In some substances a group of this kind forms a magnetic sub-lattice, and the Mössbauer spectrum then gives the average magnetization of the sub-lattice. Because of this averaging effect, no magnetic splitting is found in the spectrum of paramagnetic substances, in which the electron magnetic moment is averaged away to zero in the absence of an external magnetic field. An exception is Fe in Al_2O_3 , where the interaction between the iron ions can be made sufficiently small by using low iron concentrations and low temperature^[8].

Some applications

“Ticonal”, a group of alloys for permanent magnets, undergoes the following treatments to achieve optimum magnetic properties:

- 1) homogenizing at 1260 °C, followed by quenching;
- 2) slow cooling from 900 to 600 °C in a magnetic field, followed by quenching;
- 3) annealing just below the Curie temperature and quenching.

X-ray diffraction measurements have shown that during all these treatments the geometry of the crystal lattice does not change. What does change is the distribution of the alloying constituents (Fe, Co, Ni, Al and Cu) over the lattice sites. This has been investigated with the electron microscope^[9]. It was found that in stage 2 the substance becomes divided into needle-shaped regions alternating in chemical composition. The periodicity, in the lateral direction is about 40 nm. If we assume that these two types of region also differ magnetically the one having a greater magnetization than the other, we then have a qualitative explanation for the permanent-magnet properties of “Ticonal”: A structure of this kind has a strong preference for magnetization parallel to the long axis, and it is difficult to

reverse the direction of magnetization; as a result the coercivity is high [10]. The heat treatment in stage 3 increases still further the coercivity but it produces no change in the electronmicroscopic picture (see *fig. 9*). The Mössbauer spectrum (*fig. 10*), however, does change: after stage 3 a peak appears corresponding to iron in a non-magnetic environment [11]. This indicates

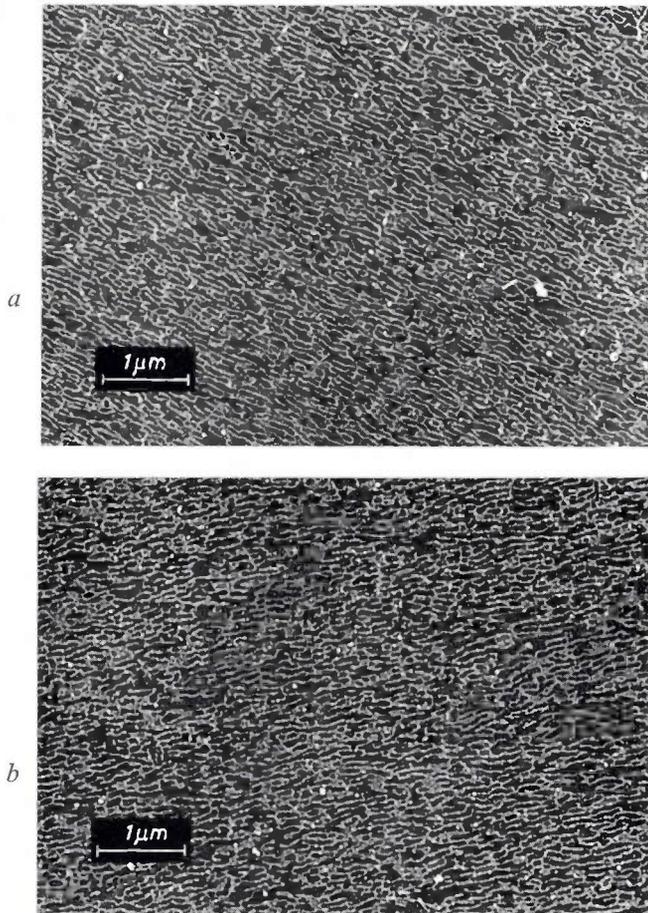


Fig. 9. Electron microscope photographs of "Ticonal".
a) After cooling in a magnetic field. Without annealing, needle-shaped regions have formed with a periodicity of about 40 nm.
b) Idem after annealing. Although the magnetic properties have changed as a result of annealing, the electron microscope photographs show no change.

that part of the material that was still magnetic after stage 2 becomes non-magnetic during stage 3. In the light of the foregoing it is clear that the coercivity must increase in this process.

A subtle change in the spectrum is that the shoulders on the inside of the outer peaks have become less prominent both after stage 2 and stage 3. These shoulders are due to those Fe atoms having many Al neighbours [12], and their rounding off indicates that during the heat treatments Fe and Al become physically separated.

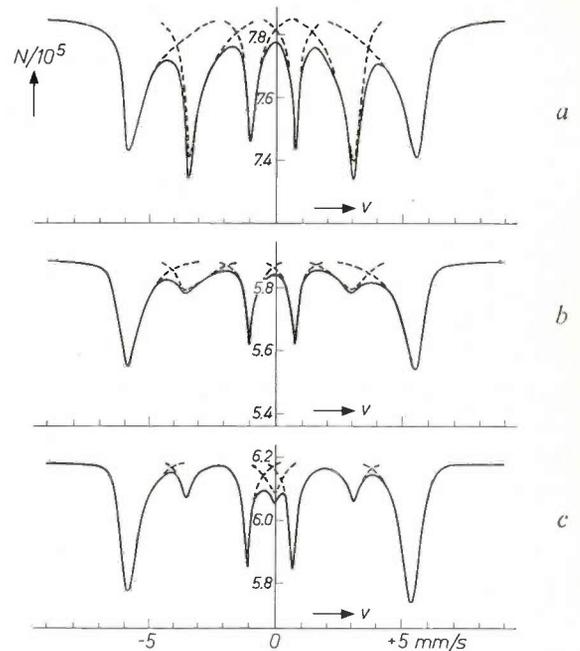


Fig. 10. Mössbauer spectra (number of pulses N , produced by the transmitted gamma radiation during a given measuring interval, as a function of source speed v of a "Ticonal" GG single-crystal in three stages of the heat treatment.

a) Homogenized by annealing at 1260 °C, followed by quenching. The cubic axes of the crystal lattice, along which the magnetization is oriented, are distributed randomly over three mutually perpendicular directions. Hence the six lines in the spectrum.

b) As (*a*), but after cooling from 900 to 600 °C in a magnetic field and quenching. The magnetization is now virtually in a single direction, which, during the measurement, was parallel to the beam of gamma rays: hence the marked attenuation of two of the six peaks.

c) As (*b*), after annealing at 585 °C. In addition to the six peaks of iron in a ferromagnetic environment, there is a small peak near $v = 0$ relating to iron in a paramagnetic environment.

In magnetic oxides the magnetic ions usually form various sub-lattices. As a rule their interaction is negative, in other words the sub-lattice magnetizations prefer to lie antiparallel. With two sub-lattices of crystallographically the same kind this results in antiferromagnetism, with dissimilar sub-lattices it leads to ferrimagnetism [13]. By macroscopic measurements it is possible

[7] D. J. Kroon, Nuclear magnetic resonance, Philips tech. Rev. **21**, 286-299, 1959/60.

[8] G. K. Wertheim and J. P. Remeika, Phys. Letters **10**, 14, 1964.

[9] K. J. de Vos, Philips Res. Repts. **18**, 405-412, 1963, and thesis, Eindhoven, 1966.

[10] E. C. Stoner and E. P. Wohlfarth, Phil. Trans. Roy. Soc. A **240**, 599, 1948.

[11] J. S. van Wieringen and J. G. Rensen, Z. angew. Phys. **21**, 69-70, 1966 (No. 2).

[12] G. K. Wertheim, V. Jaccarino, J. H. Wernick and D. N. E. Buchanan, Phys. Rev. Letters **12**, 24, 1964.

[13] Reference to these effects will be found in textbooks on magnetic material; see, for example, J. Smit and H. P. J. Wijn, Ferrites, Philips Technical Library, Eindhoven 1959, and also P. F. Bongers and G. Blasse, Crystal chemistry and magnetism of oxide materials, I. Principles and application of crystal field theory, II. Magnetic compounds with spinel structure, Philips tech. Rev. **28**, 13-30, 1967 (No. 1).

to determine only the resultant of the sub-lattice magnetizations. The Mössbauer effect makes it possible to determine the sub-lattice magnetizations separately.

An example is NiFe_2O_4 , an oxide related to the magnetic material known as ferroxcube IV. It was postulated that this has three sub-lattices for the magnetic ions, formed by Fe at tetrahedral sites, Fe at octahedral sites and Ni at octahedral sites. This picture is confirmed by the Mössbauer spectrum (fig. 11) [14]. The spectrum is built up of lines from each sub-lattice of iron, separated as a consequence of the difference in the average local magnetic field of each environment.

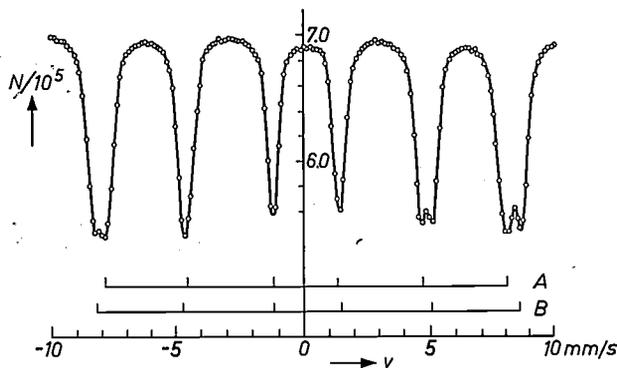


Fig. 11. Mössbauer spectrum of NiFe_2O_4 . Below in the figure it is shown how this spectrum is built up from two sub-spectra, originating from iron on tetrahedral sites (A) and on octahedral sites (B).

The permanent magnet material ferroxdure $\text{XFe}_{12}\text{O}_{19}$ where $X = \text{Ba}$ or Sr , contains five magnetic sub-lattices, formed by Fe ions on five different lattice sites. The spectrum is built up from five sub-spectra [11] [15] (fig. 12). Two of them coincide in measurements using no external magnetic field, as in fig. 12, but can be separated by putting the sample in an external field. By measuring Mössbauer spectra at different temperatures, the temperature-dependence of the sub-lattice magnetizations can be determined (fig. 13a). Since the direction of magnetization of the various sub-lattices is known, we find by addition the total magnetization as a function of temperature. This is found to agree well with macroscopic measurements (fig. 13b).

These Mössbauer measurements have thrown new

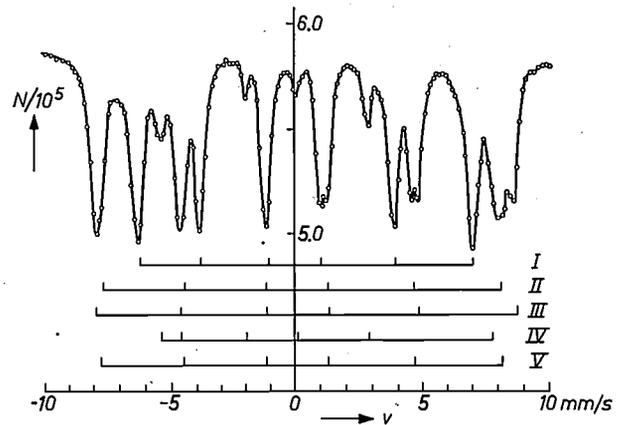


Fig. 12. Mössbauer spectrum of $\text{BaFe}_{12}\text{O}_{19}$. The breakdown of the spectrum into five sub-spectra of iron on five magnetic sub-lattices is indicated under the curve. From the relative intensities of the spectra and from the changes that result when the sample is placed in an external magnetic field, it is found that spectrum I belongs to a sub-lattice that contains six iron ions per molecule $\text{BaFe}_{12}\text{O}_{19}$, each surrounded by six oxygen ions; II to two iron ions each surrounded by four oxygen ions; III to two iron ions each surrounded by six oxygen ions; IV to one iron ion surrounded by five oxygen ions; and V to one iron ion surrounded by six oxygen ions.

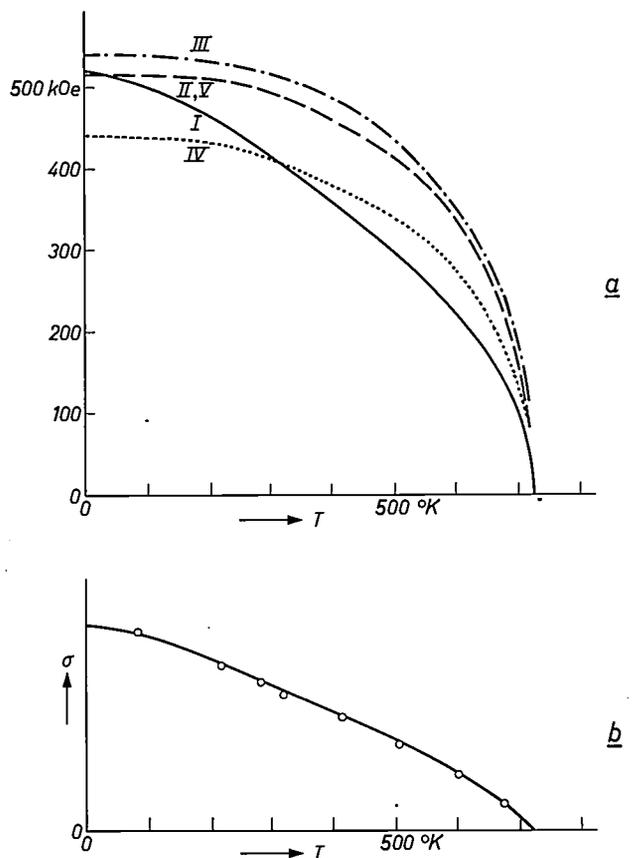


Fig. 13. a) Temperature-dependence of the five internal magnetic fields H proportional to the five sub-lattice magnetizations in $\text{BaFe}_{12}\text{O}_{19}$. The magnetic moments of the iron ions on the sub-lattices I, IV and V are parallel to the net magnetic moment and those on II and III antiparallel. b) Temperature-dependence of magnetization σ of $\text{BaFe}_{12}\text{O}_{19}$ derived from (a). The circles denote the experimental values.

[14] Measured in this laboratory by J. P. Morel; a paper on this work will appear shortly in J. Phys. Chem. Solids.

[15] J. J. van Loeff and A. Broese van Groenou, Proc. int. Conf. on Magnetism, Nottingham 1964, p. 646-649.

[16] J. Smit, J. Phys. Radium 20, 370-371, 1959.

[17] P. P. Craig and N. Sutin, Phys. Rev. Letters 11, 460-462, 1963.

[18] An extensive bibliography which is brought up to date at regular intervals is to be found in A. H. Muir, K. J. Ando and H. M. Coogan, Mössbauer Data Index, North-American Aviation Science Center.

light on the properties of ferroxdure in three respects:

- 1) The iron proves to be trivalent on all sub-lattices. The view that the large crystal anisotropy and hence the high coercivity can be explained on the assumption that the iron is partly bivalent is therefore incorrect.
- 2) All five sub-lattices show pronounced quadrupole splitting, the local symmetry thus differs considerably from cubic symmetry, and all five may be expected to contribute to the crystal anisotropy K . The assumption that K is entirely or almost entirely due to a single sub-lattice [16] (sub-lattice IV

in fig. 12, where 8% of the Fe atoms are located) is therefore incorrect.

- 3) The very rapid drop in the magnetization with increasing temperature (fig. 13*b*) is entirely due to the (50%) iron atoms constituting sub-lattice I in fig. 12, as is evident from curve I in fig. 13*a*.

All three examples given here relate to ferromagnetic or ferrimagnetic materials. The application of the Mössbauer effect is not limited to them however: for antiferromagnetic, paramagnetic and diamagnetic substances and even for viscous solutions [17] it has also proved to be an extremely useful tool of research [18].

Summary. The phenomenon of resonance absorption, familiar in the case of radio and light, occurs less readily with gamma radiation because the recoil caused by gamma quanta upon their emission and absorption is no longer negligible. The recoil energy can shift the natural frequency of source and absorber to such an extent as to destroy their resonance. In favourable cases (soft gamma rays, incorporation of emissive and absorptive nuclei in a sufficiently rigid crystal lattice) part of the radiation, the Mössbauer fraction, can nevertheless exhibit resonance absorption. These resonance lines are exceptionally sharp and can be used for investigating the very fine splitting of nuclear energy levels

caused by electric and magnetic fields at the site of the nucleus. This is illustrated for the case of iron nuclei in three magnetic materials: "Ticonal" alloys and the ferrimagnetic oxide materials $NiFe_2O_4$ and ferroxdure. In "Ticonal" the iron in the alloy was found to be partly in a ferromagnetic and partly in a paramagnetic environment. From this, and from the results of electron-microscopic investigations, an explanation is found for the high coercivity of this material. In $NiFe_2O_4$ the magnetization of the two iron sub-lattices was determined separately. In ferroxdure the same was done for all five sub-lattices over which the iron is distributed in this compound.

Digital circuit blocks

E. J. van Barneveld

Digital techniques are being applied today in a wide variety of different fields. These include not only computers of various types and sizes, and data transmission equipment, but also industrial control systems. Since digital equipment is mainly made up of large numbers of comparatively few basic circuits, the construction of this equipment can be considerably simplified if the basic circuits are available in the form of "ready-made" units. This article describes a system of such "circuit blocks", marketed by Philips.

Digital equipment is generally made up of a limited number of types of basic circuit, such as gate circuits and bistable circuits, a large number of each type being used. Of course, it would be extremely uneconomical to design completely new basic circuits for every new piece of equipment, and, in fact, in the vast majority of cases, all that is needed is a combination of existing circuits. The next stage is that the basic circuits, instead of being built by the makers of the equipment themselves, are supplied complete by the electronic component manufacturer, i.e. by the supplier of the usual resistors, transistors, etc., and in a form in which they can easily be combined with one another to form a complete equipment. As the name implies these basic circuits are generally in the form of small blocks, provided with a number of pins by which they can conveniently be attached to a printed wiring board.

Although a circuit block is, as a rule, a little more expensive than the components of which it consists, a piece of equipment constructed from such circuits is less expensive than a similar one made up from conventional components. Several factors are concerned here. First of all, much less time and specialized knowledge is required for the design of a piece of equipment made up from circuit blocks. Secondly, the circuits take up less space; when the circuit blocks described in this article are used, twice as many circuits can be included on one printed wiring board as when separate components are used. Fewer boards, connectors, etc., are therefore required. Finally, this system makes the assembly and testing of the individual circuits much simpler.

Series of circuit blocks

Circuit blocks are usually supplied in a series of different types of matched circuit. The digital equip-

ment from circuit blocks — mainly industrial measuring, regulating and control equipment — consists almost entirely of three main types of basic circuit, i.e. logic circuits, bistable circuits and trigger gates. In addition, there are various types of circuit such as input and output stages, delay circuits and clock pulse generators, which are used in only small quantities in each equipment. Although the number of different types is larger, these circuits do not usually amount to more than one-fifth of the total number of circuits. A series of circuit blocks should preferably contain all these types of circuit. If the series included only the most frequently occurring circuits — and this would at first sight seem an attractive proposition for the manufacturer — the makers of the equipment concerned would have to design all the other circuits themselves. In such a case, in spite of the use of circuit blocks, the design costs would be only very little lower than where *all* the circuits had to be designed by the equipment manufacturer.

A second reason why as complete as possible a range of circuit blocks is to be preferred is that the inputs and outputs of the circuits can then be matched, thus permitting the circuit blocks to be specified by a minimum of data and to be more readily combined.

This is the principle used in making up the series marketed by Philips. The oldest series, of which a few million units have already been made, is the "100 kHz series". The 10-series, with type numbers between 10 and 19, and the 20-series, with type numbers between 20 and 29, are of later design (see *figs. 1 and 2*). It is these latter two series that will be discussed here. They both contain the same basic circuits and are almost exactly alike in the construction of these circuits. *N-P-N* transistors are used in both ranges. The main difference between them is the semiconductor material used, which is germanium in the 10-series and silicon in the 20-series. Further, in the 20-series planar transistors are

Ir. E. J. van Barneveld is with the Philips Electronic Components and Materials Product Division, Eindhoven.

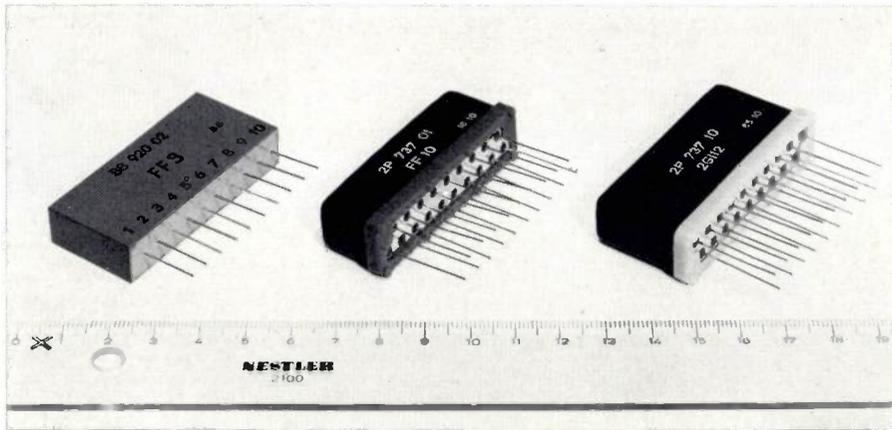


Fig. 1. The circuit blocks in the different series. On the left, a bistable circuit from the 100 kHz series, with beside it a bistable circuit from the 10-series and a block with two logic circuits from the 20-series. In the 10-series and 20-series a larger or smaller case is used according to the size of the circuit.

used, which have a higher cut-off frequency than the alloy transistors used in the 10-series. This permits a higher speed when using the 20-series blocks; some impression of the increased speed of operation may be gained from the fact that the maximum count rate of a bistable circuit is 1 MHz as against 30 kHz for the 10-series. The circuit blocks of the 20-series can also be used at higher ambient temperature; 85 °C is now permissible as against a maximum of 55 °C for the 10-series.

We shall now discuss various aspects of these series, paying particular attention to the measures taken to

make the combination of the circuit blocks as simple as possible [1].

Combination of the circuit blocks

The basic circuits included in this series can occupy only two stationary states which correspond to the conducting and non-conducting states of one or more transistors in the circuit. These states are characterized by the voltage at one or more points in the circuit (for

[1] An article by Ir. C. Slofstra will shortly appear in this journal describing (for some applications) the design of a circuit using circuit blocks.

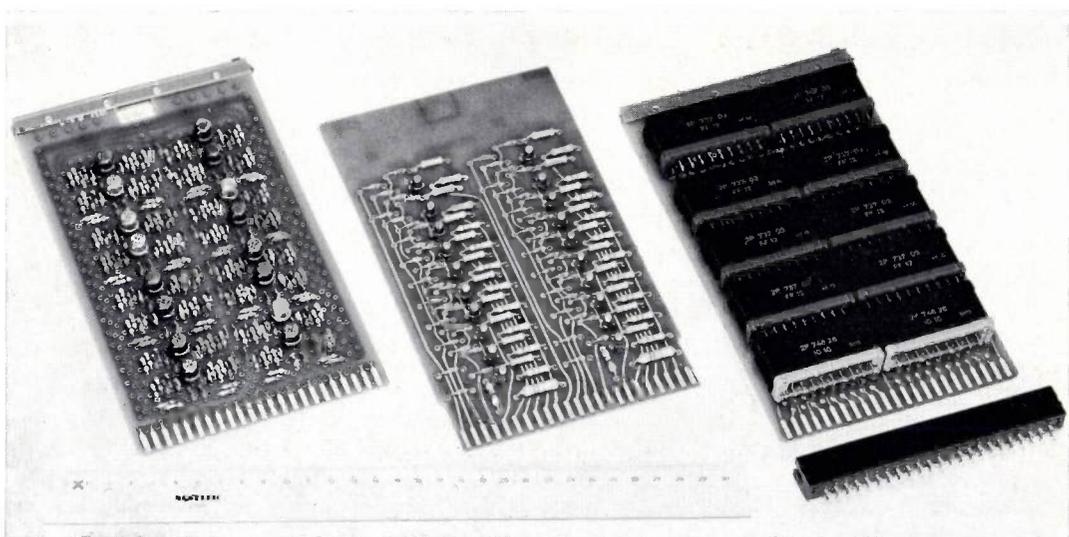


Fig. 2. On the right, some circuit blocks in the 20-series mounted on a printed wiring board. On the left, the same circuit made up from separate components, to illustrate the space-saving achieved with the blocks. With separate components, two printed circuit boards are required.

The boards are placed in a rack, where they fit into printed wiring connectors, one of which is visible in the foreground. The boards are connected together by means of wiring at the rear of the connectors.

instance the collector voltage of a transistor). In one state, indicated by H, this voltage is high (higher than V_H in fig. 3), and in the other, indicated by L, it is low (lower than V_L). Voltages between V_H and V_L cannot



Fig. 3. In digital circuits, in the steady state the output voltage is in one of the two ranges indicated by L and H. The signal temporarily enters the transition range only during the transition from L to H or vice versa.

occur continuously, but only during the transition from L to H or vice versa. In order to specify a circuit completely, V_L and V_H must therefore be given, together with the permitted output currents I_L and I_H for both states, and finally further details on the edges of the signal during the transition from L to H and vice versa. Six pieces of information are required for the output of each circuit and a further six corresponding ones for the input.

In combining logic circuits to form a larger unit, one must know the maximum number of inputs of other circuits which can be connected as a load at the output of one circuit. If a series contains only a few circuits, these numbers can be given in a table. *Table I* is the simplest possible table for an imaginary series of three circuits A, B and C. Where there is a mixed load, i.e. if it is desired to connect a number of different types of circuit to an output, this table does not give a definite answer. Therefore another table is often used, in which all allowed combinations are given. This next table is much larger (see *Table II*). Clearly, where the series contains a large number of circuits, the tables soon become unmanageably large, all the more so since some circuits have different kinds of inputs. In the present series, therefore, attempts have been made to limit the amount of information per input and per output so that a complex loading table is no longer required. It is intended rather that one should be able to calculate in a simple way from this data whether or not a certain load is permissible.

To limit the amount of information, the voltage levels V_L and V_H have first of all been made the same for all the circuit blocks. These need therefore only be given once and are of no further significance in combining the blocks. Furthermore, in this series the high-level current I_H is zero, apart from leakage currents, and only the slope of the negative-going edge of the signal in the transition from the high to the low level is of

importance, so that there now remain only two pieces of information to be given for each input and output. We shall first show how these simplifications have been achieved.

The load in the steady state

If we consider only the steady state, we can distinguish between two kinds of load corresponding to the following kinds of circuit.

- 1) Current-delivering circuits. These deliver current to the driving stage if it is at the L level, i.e. if the output of this stage is at the low voltage. (We see that in this situation the "load" — i.e. the stage following the driving stage — *delivers* current!) They do not deliver a current to the driving stage if it is at the H level.
- 2) Current-drawing circuits. These draw current from the driving stage if this is at the high level, but not if it is at the low level.

	A	B	C
A	6	3	6
B	5	2	8
C	1	1	0

Table I. Loading table for a hypothetical series of three circuit blocks A, B and C. The figures in each column show the maximum number of circuit blocks of a certain type that may be connected to the output of a circuit block indicated at the left.

	A	B	C
A	6 + 0 + 6		
	4 + 1 + 6		
	2 + 2 + 6		
	0 + 3 + 6		
B	5 + 0 + 8		
	3 + 1 + 8		
	1 + 2 + 8		
C	1 + 0 + 0		
	0 + 1 + 0		

Table II. An extension of Table I, in which mixed loading is taken into account. The maximum number of different inputs that can be simultaneously connected to one output are given.

Figs. 4 and *5* show examples of these circuits. The driving stage is represented by a transistor with a grounded emitter, as this form of output is used in all the circuit blocks in the series under discussion.

To obtain good loadability with current-drawing circuits, the collector resistance R_c must be low. This, however, means that, if the transistor is conducting, a large current passes through R_c , so that only a few current-delivering circuits can be connected. A higher R_c is better for current-delivering circuits but not for current-drawing circuits. The loadability is distributed over both types by fixing the value of R_c . This compromise will, of course, not be optimum for most

cases. Moreover, this distribution leads to complex loading tables like those mentioned above.

To avoid these difficulties, only current-delivering circuits are used in the circuit blocks described here; this means that the current at high level is always zero, as stated above. At first sight it would appear to be impossible to make up certain circuits with these circuit blocks. This is not so, however, since a current-drawing circuit can easily be converted into a current-delivering circuit by the addition of two resistors and a blocking diode (see fig. 6). In fig. 5 the current for the load is supplied via the collector resistance of the driving stage and in fig. 6 this current is supplied via the resistor R_1 , while diode D prevents current from being drawn from the driving stage. The new circuit thus provided can only deliver current to the driving stage and only if the latter is at the low level.

This simplification makes it sufficient to specify only one current per input and per output in the steady state, and this is the current which flows in the L state. This makes the loading table much simpler. The maximum current which may flow through it is given for each output, for each input the current it delivers is given. For the purposes of checking whether a given load is permissible, all that need be done is to add the currents of all the inputs connected to a certain output and see whether this exceeds the permissible output current. As the input currents of most of the circuit blocks are the same, all that usually has to be done is to count the number of inputs. This additive system of loading represents a great simplification, especially with mixed loading.

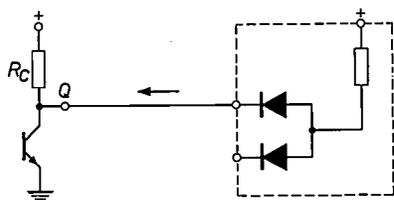


Fig. 4. Example of a current-delivering circuit (in the "box"). If the transistor in the driving stage is conducting and the voltage at the output Q is therefore at the low level, current flows to the left through the output. If the transistor is cut off (with the output at the H level), the current is zero.

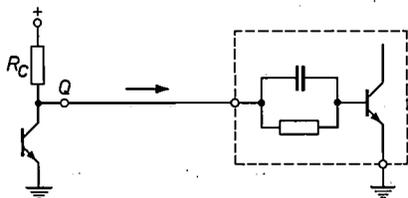


Fig. 5. Example of a current-drawing circuit (in the "box"). If the transistor in the driving stage is conducting (output Q at L level), no current flows through the output. If the transistor is blocked, current flows to the right.

Now that current-drawing circuits are not being taken into account, the value of the collector resistor R_C can be made so high that the maximum number of current-delivering circuits may be connected to an output.

There are two further advantages in the use of the blocking diode. The first is that there is an anti-noise threshold voltage across this diode. To see how this operates, let us consider the state where the driving transistor is non-conducting. The output Q is then at

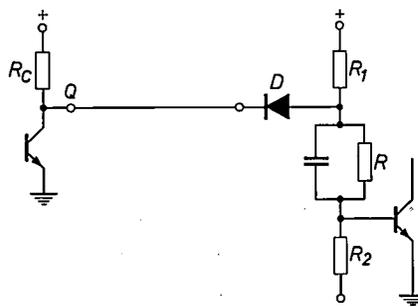


Fig. 6. The current-drawing circuit in fig. 5 can be converted into a current-delivering circuit by the addition of a blocking diode D and two resistors R_1 and R_2 . This also produces a threshold against interference since the left-hand side of D is at a higher voltage than the right-hand side.

the supply voltage and the driven transistor is conducting. If a negative noise pulse now occurs in the connection between the stages, it can block the driven transistor in the situation illustrated in fig. 5. In the situation given in fig. 6, the supply voltage is applied to the left-hand side of the diode D . The voltage on the right-hand side of the diode is lower, however, because of the voltage division due to resistors R_1 and R_2 . There is, therefore, a voltage in the reverse direction across the diode. Noise pulses can affect the second stage now only if they overcome the diode reverse bias. If however the driving transistor is conducting, it forms a low impedance to earth; the interfering pulses picked up on the connections are then small enough not to affect the next stage.

A second additional advantage of the use of a blocking diode is that by connecting a few more diodes in parallel a logic circuit (AND or OR function) can very easily be obtained. This will be discussed further later on.

To summarize, the most important advantages of limiting the units to current-delivering circuits are:

- 1) great loadability (fan-out);
- 2) only one piece of information is needed for each input and output for the stationary state, and hence the loading table is simple;
- 3) the loads are additive, so that it is a simple matter to check mixed loads;
- 4) the circuits are less sensitive to interference;
- 5) it is very easy to add a logic function.

Transient charges

The loading table must also provide information on the signal edges. Here we are faced with the same problem as with the steady state, i.e. that the table threatens to become complex, particularly with mixed loads. At the inputs of some circuits, particularly at the inputs of the trigger gates, the transition from the high to the low voltage must take place within a very short time. When different circuit blocks are combined, therefore, the question once more arises of whether the driving stage can provide a sufficiently steep edge, and the loading table should provide the answer. (In the transition from the low to the high voltage, the transition time does not critically affect the operation of the circuits.)

In assessing the edge steepness of the transition from the high to the low voltage, use is made of the concept of "transient charge" defined as follows. At the high voltage, no current flows through the output of the driving stage (see fig. 6 again), while current does flow at the low voltage. The load then supplies current to the driving stage. During the signal transition, the current through the output thus increases from zero to a certain final value. This means that, during this time, a given charge flows through the output of the driving stage. For each output this "transient" charge has a certain maximum which can be calculated from the transistor data.

The transient charge is absorbed by the load. If this is capacitive, it is clear that, during the transient, a *clearly-defined* charge will be required which must be supplied by the driving stage. The magnitude of this charge is CAV , where ΔV is the difference between the voltages across the capacitor before and after the transient. For a resistive load also the charge flowing during the transition from high to low voltage can be calculated from the final value of the current and the time taken by the voltage transition. Both kinds of load can therefore be characterized by one quantity, i.e. the charge displaced. The loads are thus also additive under non-steady-state conditions.

The procedure for checking a given combination of circuits is now the same as for the steady state. The transient charge is given for each output and input, the charges belonging to the inputs connected to one output are added together and a check is made to see whether the total charge is less than the output can supply. If so, the transition time will be shorter than needed for the trigger gate, and correct operation will be ensured.

There is one difference in checking the current load. The currents must always be checked, whereas the transient charges need to be checked only if inputs of trigger gates are connected to an output. If this is not

the case no special requirements are set for the transient.

The complete loading table thus consists only of figures giving one current and one transient charge for each input and output in the system. We give this table here for the circuit blocks of the 10-series; see *Table III*. (The symbols G , T and S indicate the different types of input; these will be dealt with later in this article. The outputs are indicated by Q , or by Q_1 and Q_2 if a circuit block has two.) Thirty-eight pieces of inform-

Table III. The loading table for the 10-series.

Type	Input terminal	Direct current mA	Transient charge nC	
FF11, FF12	} {	G	1.1	1.2
2.TG13, 2.TG14, 4.TG15		T	1.1	3.4
FF10, FF11, FF12		S	1.95	2.8
2.GI10, 2.GI11, 2.GI12		G	1.1	2.1
GA11		G	1.1	1.2
OS11	} {	G	1.1	1.2
		T	1.1	2.3
TU10, PD11	} {	G	1.1	1.2
		T	1.1	3.2
RD10		G	4.7	3.4
PA10		G	5.3	5.2

Type	Output terminal	Direct current mA	Transient charge nC	
FF10, FF11, FF12	Q_1, Q_2	8.2	27	
2.GI10, 2.GI11, 2.GI12	Q	8.2	9	
GA11	Q	62	75	
OS11	} {	Q_1	8.6	24
		Q_2	12.8	29
TU10	Q	32	27	
PD11	Q	100	185	
PS10	Q	10	39	

ation are required for the complete specification of nine basic circuits. The simplest loading table of the Table I type would require for this 88 pieces of information, since there are 11 different inputs and 8 outputs, and this would still give no definite indication on mixed loads, such as is given by Table III.

We shall now take a look at the various types of basic circuit included in the series. We shall see here the consequences of the measures discussed, in particular of the limitation to current-delivering circuits.

Logic circuits

It often happens in digital circuits that the voltage at a certain point must be high or low depending on whether there are certain combinations of voltages at a number of other points. These voltages are then applied to the inputs of a "logic circuit", i.e. a circuit designed

in such a way that the desired voltage appears at its output if the input voltages satisfy the appropriate conditions.

The name "logic circuits" is derived from the fact that the relationship between the input and output voltages of these circuits can be described by logic functions from Boolean algebra — the algebra which is used for calculations with elements that can operate in only two stationary states. A letter is allocated to each element in exactly the same way as in ordinary algebra, and the two states are indicated by "0" and "1", e.g. $A = "1"$ or $A = "0"$. The interdependence of different elements can also be expressed here by a function, e.g. $Z = AB + CD$. The significance of a function is however slightly different from its usual meaning. In this algebra all the functions, which are also called operations, can be broken down into three basic functions:

In breaking down an involved logic function, we can use the rules of De Morgan, which are given here without proof:

$$\overline{A + B + C + \dots} = \overline{A} \cdot \overline{B} \cdot \overline{C} \cdot \dots$$

and

$$\overline{A \cdot B \cdot C \cdot \dots} = \overline{A} + \overline{B} + \overline{C} + \dots$$

The commutative and distributive rules can also be used, just as in algebraic processes.

As logic functions can be broken down into the three basic functions, so can logic circuits be composed of three basic circuits, namely the AND, OR and NOT circuits. These basic circuits can be electrically constituted in a number of ways. One of the most usual is to use diodes and transistors (Diode-Transistor-Logic, or DTL). Figs. 7a, b and c show the three circuits. For purposes of simplicity, the AND and OR circuits are shown with only two inputs, although there can be more.

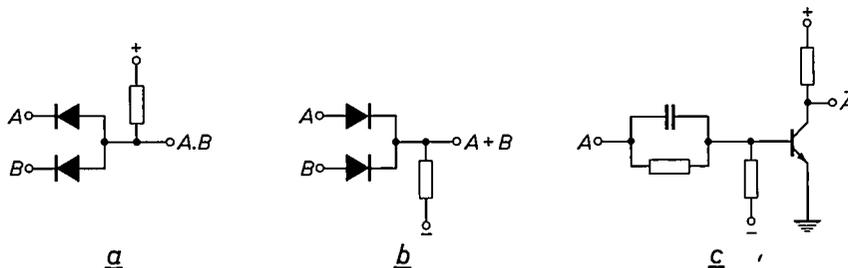


Fig. 7. a) The AND circuit. The output here is at a high voltage (state "1") only if the voltages at *A* and *B* are high. b) The OR circuit. Here, the output is at a high voltage if either *A* or *B* (or both) are at a high voltage. c) A transistor connected up as an inverting amplifier operates as a NOT circuit. If the input voltage is high, the transistor conducts and the output voltage is low.

"AND", "OR" and "NOT". In the AND function, designated by $Z = A.B.C. \dots$, Z is in state "1" if all the elements, i.e. *A* and *B* and *C* etc., are in state "1". For the OR function ($Z = A + B + C + \dots$), Z is in state "1" if one or more of the components, i.e. *A* or *B* or *C*, or *A* and *B*, etc., are in state "1". The NOT function $Z = \overline{A}$ gives an inversion, i.e. $Z = "1"$ if $A = "0"$ and $Z = "0"$ if $A = "1"$. The function $Z = A.B + C.D$ given above as an example therefore has the following significance: $Z = "1"$ if *A* and *B*, or *C* and *D*, or *A* and *B* and *C* and *D* are "1". The AND and OR functions can be given in the form of a "truth table". Table IV gives this table for two components *A* and *B*.

Logic functions are written in such a way as to provide as great a degree of correspondence as possible with ordinary arithmetical processes: if $A = "1"$ and $B = "0"$, then $A.B = "0"$ and $A + B = "1"$, just like ordinary multiplication and addition. The only exception is formed by the case $A = "1"$, $B = "1"$; now $A + B$ is not 2, as one would expect, but "1".

Table IV. Truth table for the AND function $A.B$ and the OR function $A + B$.

<i>A</i>	<i>B</i>	$A.B$	$A + B$
1	1	1	1
1	0	0	1
0	1	0	1
0	0	0	0

Because the AND and OR circuits may also be regarded as gate circuits, i.e. a voltage at one input is allowed to pass through to the output only if there is a certain voltage at the other input, the terms AND gate and OR gate have also come to be used.

"Positive logic" has been chosen for the examples in fig. 7; that is to say, state "1" is allocated to the high voltage level. This "logic convention" could equally

well have been reversed, i.e. state "1" could correspond to the low level. In such a case, the terminology as applied to the circuits would also be reversed. Fig. 7a is then an OR circuit, since the output is at the low level as soon as one of the inputs is negative, and fig. 7b is, then, an AND circuit. Table IV also shows that the AND and OR function can be interchanged in this way.

Regardless of the logic convention chosen, be it positive or negative logic, one of the two circuits, AND or OR, is always a current-delivering circuit (fig. 7a) and the other a current-drawing circuit (fig. 7b). This situation is not permissible in the logic circuits used in the series discussed here. The solution found to this problem amounts, in fact, to the use of mixed logic, i.e. positive or negative, as required, so that both the AND and the OR function can be obtained with one basic circuit. The current-delivering circuit is used here as the basic circuit (an AND circuit with positive logic) followed by an inverting amplifier (fig. 8a). This cir-

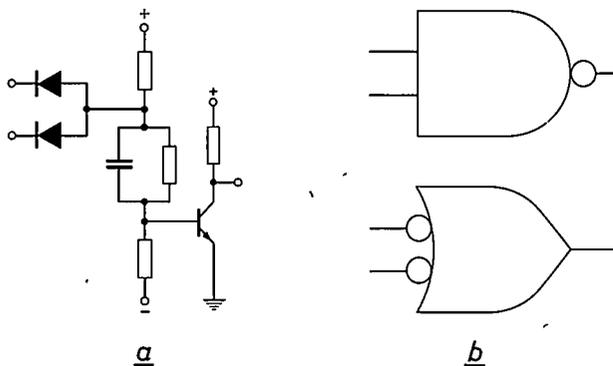


Fig. 8. a) The NAND circuit. b) The symbols indicating the NAND circuit as an AND function (upper symbol) and an OR function (lower symbol). A circle at an input or output indicates that the voltage at that point is low if the condition is fulfilled.

circuit is referred to as a NAND circuit, the term being a contraction of the words NOT and AND, just like NOR as the term for the combination of a NOT and an OR circuit. With this one basic circuit, all possible logic circuits can be constructed.

The method of working with this logic unit can most simply be described with the aid of a truth table using the voltage levels H and L at the inputs and at the output [2], instead of the states "0" and "1", which

change depending on whether positive or negative logic is used. Table V gives this truth table for the NAND circuit in fig. 8a. It will be seen from the table that the output voltage is low only if both input voltages are high. The NAND circuit therefore operates as an AND

Table V. The truth table for the NAND circuit in fig. 8a.

inputs		output
H	H	L
H	L	H
L	H	H
L	L	H

circuit at the *high* level — i.e. if voltages are applied to it that are high in the desired state — the output voltage being low. This is indicated by the upper symbol in fig. 8b, where the circle at the output indicates that this is at the low level if the AND condition is satisfied. The table also shows that the output voltage is high if either or both of the input voltages are low. The NAND circuit therefore operates as an OR circuit at the *low* level — i.e. if voltages are applied which are low in the desired state —, the output voltage being high. This is indicated by the lower symbol in fig. 8b, the circles at the inputs showing that this function operates at the low level, while the absence of a circle at the output shows that the output voltage is high if the input condition is satisfied.

The NAND circuit can also be separately used as a NOT circuit; the signal to be inverted must then be applied to one of the inputs. This is because an unconnected input behaves as if the high voltage is applied to it (in both situations the diode does not conduct). If we now apply a signal at one input and do not connect the other, this signal appears at the inverting circuit of the NAND (the AND gate is now always open) and the inverted signal appears at the output of the NAND.

The NAND circuit therefore enables all the basic logic circuits to be obtained, provided that the input signals can be supplied with the correct polarity. In some situations it may first be necessary to invert the input signals, but in practice this is seldom required. Signals of both polarities are usually available, especially if they originate from bistable devices.

An example of the construction of more complicated logic circuits is given by the various possible circuits for obtaining the function $A.B + C.D$. Fig. 9 shows this function obtained in the conventional way with two AND and one OR circuits and an inverting amplifier.

[2] This method of approach and the symbols that we shall use here have been taken from the American military standard MIL-STD-806B, described in: F. Flanagan, Standardization of logic diagrams, Computer Design 3, No. 7, 12-19, 1964; see also American Standard Graphical Symbols for Logic Diagrams, Y 32.14, American Standards Association.

[3] R. E. Burke and J. G. van Bosse, NAND-AND circuits, IEEE Trans. on electronic computers EC-14, 63-65, 1965.

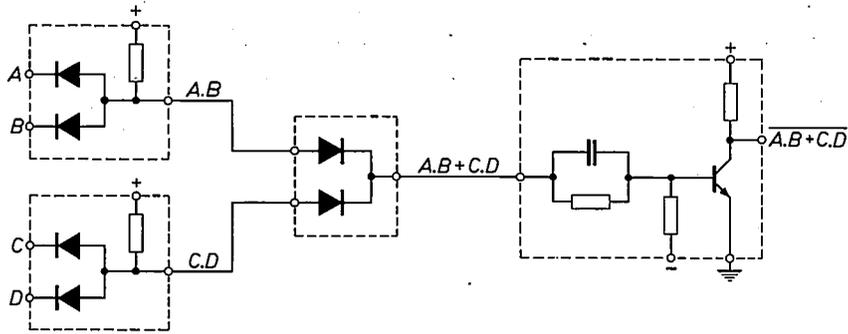


Fig. 9. The logic function $\overline{A.B + C.D}$ produced with AND and OR circuits and an inverting amplifier.

Fig. 10 illustrates the circuit made up from three NANDs. Here, however, the polarity of the output signal is incorrect, so that an inverting amplifier would be required. It is possible however to make two NANDs suffice by connecting the collectors of the output transistors in parallel (figs. 11a and b)^[3]. The two transistors now form an OR circuit. The common output has a low voltage if either of the transistors conducts. The disadvantage of this circuit is that the functions $A.B$ and $C.D$ are no longer available separately. We therefore see that the function $\overline{A.B + C.D}$, which requires four units with the conventional technique, can be made up from two NANDs.

A great advantage of the NAND system is that the voltage levels H and L are always the same. This is not

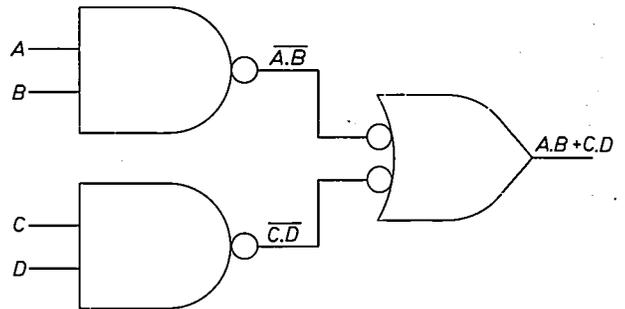


Fig. 10. The logic function $\overline{A.B + C.D}$ obtained with three NAND circuits. To make $\overline{A.B + C.D}$, an inverting amplifier must be added.

so in fig. 9. If the input voltages here are 10 V, the voltages at the inputs of the second stage are 4 V, for

example. As long as there is no amplifier in between, the difference between the high and low voltage will decrease with every stage. In the NAND system, however, there is further amplification in each stage, so that the same voltages occur throughout the equipment. This means that a voltage may be taken out anywhere in a system and used directly for another circuit; the tracing of faults in such a system is also much simpler.

One objection to the use of this method could be raised: this is that more transistors are necessary than in the normal system. Every stage, in fact, now contains an inverting amplifier, while otherwise ampli-

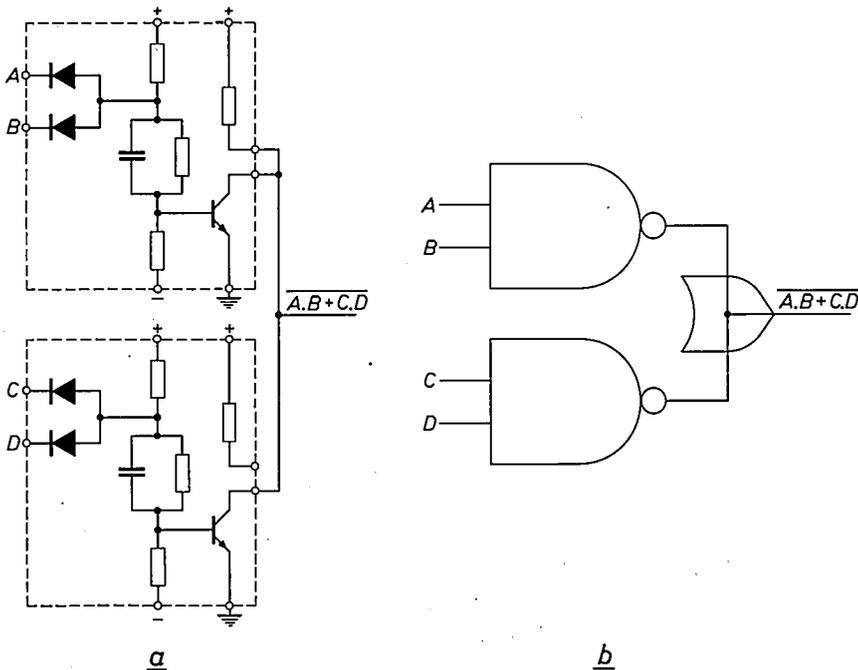


Fig. 11. a) The logic function $\overline{A.B + C.D}$ can be directly obtained with two NAND circuits by connecting the collectors of the output transistors in parallel. No separate circuit is then required for the OR function. Here the collector resistor of only one NAND-circuit is connected; this is done to get a lower current consumption and thus a higher loadability. b) The circuit in symbols.

fication often takes place only after every two stages (see fig. 9). The NAND technique could therefore be more expensive in use. This point has been investigated by comparing conventional and NAND designs in a few practical cases, including a digital computer [4]. The expected difference in cost was indeed found but it was too small to offset the considerable advantages of the NAND system.

We can recapitulate the most significant advantages of the use of NAND circuits as follows.

- 1) The advantages of the restriction to current-delivering circuits are also of course applicable here, the high degree of loadability being of particular importance.
- 2) All voltages in the system are standard voltages. Thus, every output voltage can be used at various points.
- 3) Fewer units are generally needed to make up a given circuit than when other basic circuits are chosen. This means that fewer circuit blocks, printed wiring boards, connectors, etc., are required.
- 4) Because of the restriction to one kind of circuit, not only are fewer units needed to make a piece of equipment but also fewer different types are required. This simplifies manufacture and servicing and reduces the number of spares required.

The last two points contribute most to the economic attraction of the NAND circuit.

The bistable circuit and the trigger gate

We should now like to consider further the bistable circuit and the trigger gates used in this series of circuit blocks. The bistable circuits (flip-flops) are of the "decoupled" type (see fig. 12). They consist in fact of two NAND circuits cross-coupled via diodes D_1 and D_2 . These diodes have two functions. First, there is a threshold voltage across each diode to counteract noise and secondly they facilitate the switch-over of the bistable circuit.

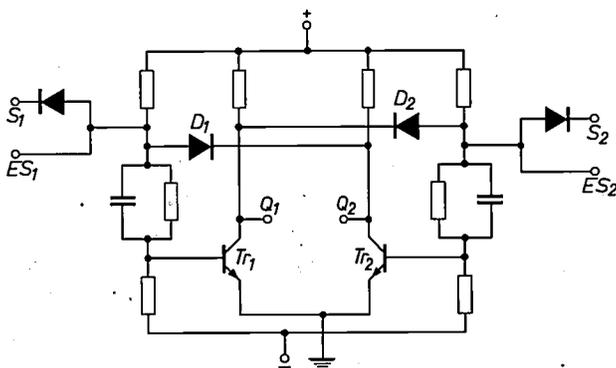


Fig. 12. A "decoupled" bistable circuit of the type included in the range (unit FF10 in the 10-series). The circuit consists of two cross-coupled NAND circuits. The circuit can be made to switch over by taking the S input of the conducting transistor to the low level.

The threshold voltage is the same as that discussed in the general treatment of the current-delivering circuit (page 47). Diodes D_1 and D_2 are, in fact, blocking diodes in current-delivering circuits. Here, the thresholds prevent a negative pulse at the output of the non-conducting transistor from arriving at the input of the conducting transistor, which could cause the bistable circuit to change its state. The name of this type of circuit comes from this "decoupling" effect provided by the diodes. The insensitivity to noise is very important in a bistable circuit. If it changes state as a result of a noise pulse, it will remain in this incorrect state even after the interference has ceased.

The bistable circuit can very easily be made to change state if the input S (S for "set") on the side of the conducting transistor, e.g. S_1 for Tr_1 , is taken to the low level. The base voltage of this transistor then becomes negative, the collector current becomes zero, and the output voltage Q_1 increases so that Tr_2 now begins to conduct and because of its low output voltage maintains Tr_1 in the cut-off state. This has thus brought the bistable circuit into its other stable state. If no decoupling diodes are used, the difficulty arises that, on changing state, not only point S but also the inputs of all circuits connected to Q_2 must be taken to earth potential. This is quite permissible for one bistable circuit, but if several are to be made to change over simultaneously — and this is often the case — an undesirably high current may be necessary and the speed of the change-over will be adversely affected. The S inputs show considerable similarity to the inputs of a NAND circuit. Supplementary diodes can be connected to the ES inputs (E for expander), enabling the number of inputs of a bistable circuit to be increased.

The bistable circuit is often used in combination with trigger gates, which ensure that a pulse can only cause the bistable circuit to change state if a certain condition is met.

Realization of the trigger gate

In its simplest form, the trigger gate consists of a capacitor, a resistor and a diode. Such a circuit is connected to the base of each of the transistors in a bistable circuit (fig. 13). Let us take Tr_1 to be in the conducting state and consider the trigger gate connected to this transistor. The base voltage of Tr_1 is close to earth potential. Let us now assume that point G_1 , the condition input, is also roughly at earth potential and that T_1 , the trigger input, is at the high level. The left-hand side of capacitor C_1 is now at a positive voltage and the right-hand side is at earth potential, and diode D_3 is reverse-biased. If now the voltage at T_1 falls ($H \rightarrow L$), the right-hand side of C_1 goes negative for a moment, the diode starts to conduct and thus takes

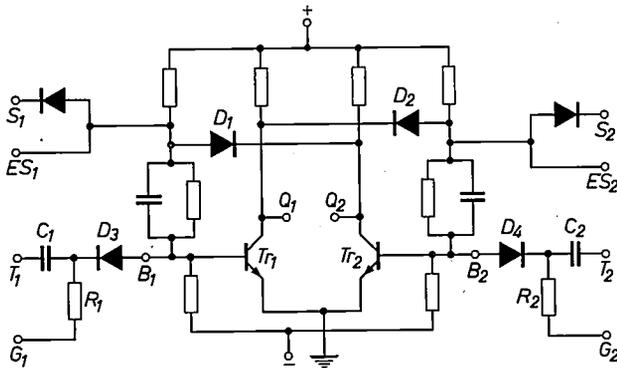


Fig. 13. The bistable circuit of fig. 12 with two simple trigger gates. G_1 and G_2 are the condition inputs, T_1 and T_2 the trigger inputs and B_1 and B_2 the base inputs of the bistable circuit.

up the base current of the transistor. There is then no transistor current for a short time and, if C_1 and R_1 are of the correct values, this period is sufficient to cause the bistable circuit to change state. If however we start from the situation where G_1 is at the high voltage, both sides of C_1 are at this high voltage. If the voltage at T_1 falls there will indeed be a fall in voltage at the right-hand side of the capacitor, but this does not go negative. The diode therefore remains reverse-biased and the bistable circuit does not change over.

Both the trigger inputs T_1 and T_2 of such a bistable circuit can be permanently connected to a clock pulse generator delivering pulses at a certain frequency to the inputs. The bistable circuit can then be set to a desired state by making the voltages high or low at the condition inputs G_1 and G_2 . The condition may also be derived directly from the bistable circuit, e.g. by connecting each condition input to the output of the transistor to which the trigger gate is connected (i.e. G_1 to Q_1 and G_2 to Q_2). If the two trigger inputs are now connected together, and if a series of pulses is applied to them, the bistable circuit is triggered by every pulse. The circuit now operates as a binary scaler. This is an important and often-used application.

Irrespective of the voltages at the inputs G and T , the bistable circuit can be set in a certain position by bringing one of the S inputs to earth potential. This is mainly used for setting the circuit in the zero position.

The simple trigger gate has several drawbacks which make it necessary to design a more complicated circuit for the units described here^[5]. In the first place, a voltage change at one of the inputs T or G produces a current pulse at the other input, and this current can flow in either direction. The effects of this on the circuit connected to this input can be undesirable, and are hard to prevent. Moreover, this bi-directional current does not suit our desire for a system with only one kind of load.

A second objection is that the response time of the circuit is very largely governed by previous conditions. If, for example, T has been at a low voltage and G at a high voltage for some time, capacitor C is charged. If now both voltages change, a long time will elapse before the charge state of the capacitor changes and a steady state holds once more. If however only one of the voltages changes, the response time is shorter. It is also very difficult to calculate the response times because of the effects of the capacitance of the wiring and the collector resistances of driving stages.

These objections can be partly overcome by requiring that, if the voltage at the condition input is to change, it must do so immediately upon the arrival of a trigger pulse at input T . This is known as synchronous operation. The likelihood of the following trigger pulse occurring before the circuit has returned to the steady state is then as small as possible. Such requirements however cannot always be made.

These drawbacks do not occur with the trigger gate circuit shown in fig. 14, which is used in our series.

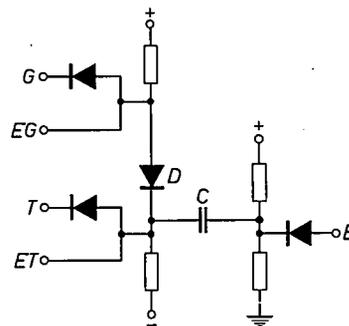


Fig. 14. The trigger gate circuit used in the circuit blocks. G is the condition input, T the trigger input, and B is connected to the base input of a bistable circuit.

The central point in this circuit is, once again, a capacitor C . The right-hand side of this capacitor is taken by the voltage divider to a voltage close to earth potential. If both G and T are at the high level, the left-hand side of the capacitor is at a high voltage. If the voltage at point T now falls sharply from high to low, the voltage at the right-hand side of C falls for a short time below earth potential, thus causing the bistable circuit to change over in the same way as in the simple circuit.

The capacitor is charged only when both G and T are at the high level. If G is at earth potential and T is at the high level, diode D is blocked so that the left-hand side of C is taken to approximately earth potential by the negative voltage. If the voltage at T is now also taken to earth potential, nothing happens, since the diode at T does not become conducting. The trigger gate is thus opened if G is at the high level and closed if G is at the low level.

[4] For this last case the research was undertaken by Dr. N. C. de Troye of Philips Research Laboratories, Eindhoven.

[5] See N. O. Sokal, Trouble-spots in circuits, *Electronic Design* 12, No. 23, p. 32, 1964.

In this circuit, a voltage change at one of the inputs T or G is clearly not noticeable at the other input. Moreover, the response time at a change in voltage at point G now depends solely on the value of the capacitance and the resistances of the circuit. These values may be so selected that the response times on opening and closing are equal. There need therefore be only one response time to be taken into account, and this can be calculated precisely and is independent of previous conditions.

Yet another facility is offered by the circuit of the trigger gate of fig. 14. It is possible to connect diodes at input EG in parallel with the diode at G , thus, in a very simple way, providing an AND circuit. The trigger gate is now opened only if there is a high voltage on *all* these diodes. An OR function can also be produced by connecting a few trigger gates in parallel. Such facilities can be very important in practice. The number of trigger inputs of a trigger gate can be increased in the same way as described for the G inputs and for the S inputs of a bistable circuit, i.e. by connection of additional diodes at point ET . The pulse gates are included in the series as independent circuit blocks, but bistable circuits are also available which already have two trigger gates (fig. 15).

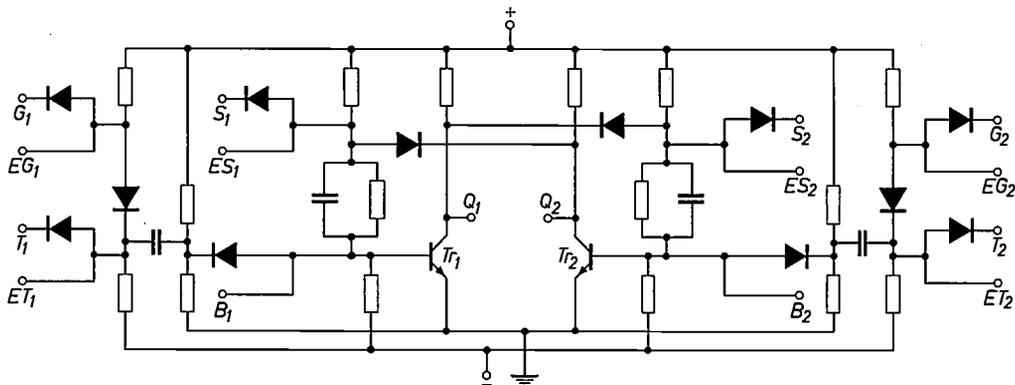


Fig. 15. A bistable circuit with trigger gates mounted in one circuit block (type FF12).

The other circuit blocks in the series

Table III gives a survey of the various types of circuit blocks in the 10-series. The inputs are classified in the manner used above: trigger inputs of trigger gates, T ; condition inputs and inputs of NAND circuits, G ; set inputs, S . The outputs of the circuits are called Q , or Q_1 and Q_2 if there are two.

The 10-series includes three circuit blocks with a bistable circuit, given the type indication FF (from "flip-flop"), one without and two with trigger gates. There are also three circuit blocks with two or four trigger gates (TG) and three with two NAND circuits (GI, gate inverters). We shall now briefly list the other circuit blocks in the 10-series.

An AND circuit with a non-inverting amplifier (GA, gate amplifier). The loadability of this circuit is much greater than that of the GI circuit.

Two circuits providing a time-delay. Both comprise a trigger gate and a monostable circuit. The type OS (one-shot multivibrator) gives a delay time of $4 \mu\text{s}$ to 30 ms, the type TU (timer unit) gives longer times, up to 60 s. Both circuits attain their longest delay time by the addition of an external capacitor.

A clock pulse generator (PD, pulse driver), consisting of a trigger gate followed by a monostable circuit.

A Schmitt trigger by means of which pulses can be properly shaped in slope and height (PS, pulse shaper).

Two different output amplifiers, one for low powers (RD, relay driver), and one for higher powers (PA, power amplifier).

A circuit that decodes the position of a decade of a binary counter and controls a decimal indicator tube (ID, indicator tube driver). As this circuit is only intended to be controlled by a counting decade, it is not given in Table III. Two of these circuits together with two decades each of four bistable circuits are shown on the printed wiring board in fig. 2.

The 20-series includes the same types, but there the

GA is replaced by a GI with a greater loadability. There are also circuit blocks for controlling and reading out from stores and for transmitting and receiving information via a cable.

Power supply equipment to suit the series is available, as are also a number of standard printed wiring boards with racks into which they can be fitted.

Construction of the circuit blocks

The internal arrangement of the circuit blocks is shown in figs. 16 and 17. The components are assembled on two printed wiring boards, which are connected by sturdy wire interconnections. The boards are then supplied with connection wires (pins) and folded to-

gether. A plastic strip with holes drilled in it keeps the two rows of pins in the right position. The complete unit is then inserted into a metal casing filled with a liquid potting compound. This compound hardens after a short time. A protecting layer of epoxy resin is then applied to prevent moisture from entering, and finally a plastic cover is fitted. The potting compound hardens to a rubbery consistency and prevents shock of vibration from causing relative movement of the com-

located at one end, with the *S* and *T* inputs next to them. These points can therefore be reached by conductors along the edge of the board. Next to the *T* inputs are the outputs, with the *G* inputs at the other end. Care has been taken to ensure that points that are most frequently connected will be close together after the circuit block has been placed on the board. These measures usually avoid the crossing of wires on the board, so that single-sided printed wiring boards can

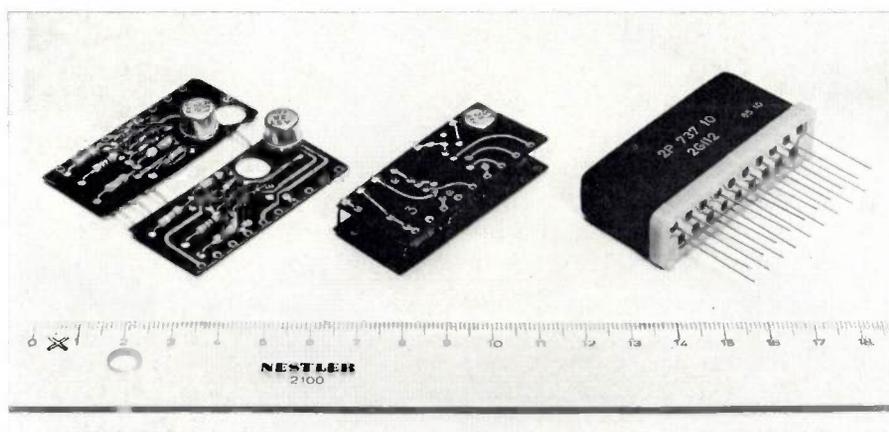


Fig. 16. The interior of the circuit blocks. The circuit is arranged on two printed wiring boards which are folded together and inserted into a metal casing filled with a rubbery compound.

ponents, which could lead to fracture of the connecting wires. Potting also ensures good transfer of heat to the wall of the unit. Since not all of the circuits are of the same size, there are two sizes of casing: the dimensions of a block are either $54.9 \times 14.7 \times 19.5$ mm or $54.9 \times 14.7 \times 27.0$ mm.

The system also includes a few types of printed circuit board on which the circuit blocks can be arranged, but the user will more often have to design these himself. Care has, however, been taken as far as possible to distribute the various connections among the pins on the units in such a way that these boards need not be excessively complicated. The connections are therefore arranged as follows: the power supply connections are

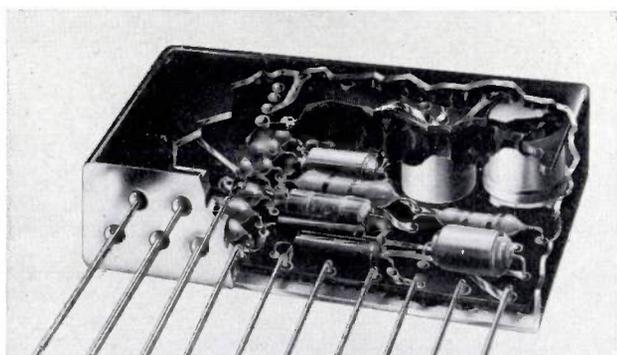


Fig. 17. Sectional view of a circuit block. (In the present design the pins come out through a cap of a different shape.)

be used. These apparently trivial considerations do nevertheless have considerable effect on the design costs of a piece of equipment.

Designing circuit blocks

Finally, we shall briefly examine the method used in designing the blocks. The designs must satisfy much more stringent requirements than those for the various circuits in say a radio receiver, if the reject rate in production is to be kept within reasonable bounds.

All the components used to construct a circuit deviate to a certain degree from their nominal values. The largest permissible deviation (tolerance) is determined for each component by several different factors. If it so happens that all the components in a circuit have values close to the tolerance limits and if these faults affect the result in the same sense, the result will be very poor. The likelihood of this in a circuit made up from a large number of components, e.g. a radio set, is very small, and in such a case no special action is required. The situation is quite different however in a digital equipment. A standard signal with a certain tolerance must be supplied to the output of each stage, and each stage must, therefore, be "good". Moreover, in such a stage there are sometimes only four or five components in the circuit, and the likelihood of the coincidence of the most unfavourable values for these components is far from small.

For these reasons, the "worst case" method is often followed in the design of digital circuits [6]. Here, the circuits are designed so that they will still operate properly even if the values of the components are at the least favourable values (but, of course, still within the limits of the tolerance). There will, therefore, be no rejects due to the inevitable spread during manufacture. One of the results of this method is that most circuits are better than calculated, and thus better than they need be, and, in nearly every case, much better. This does give the impression that the worst case method is not the most economical one.

There is another possible method, the "statistical" method, in which a small reject rate is accepted. It can only be used however if a few conditions are satisfied. First, there must be no correlation between the deviations in the various values. This condition is met by the actual values of different components, but when dealing for instance with the current amplification and the cut-off frequency of the same transistor, this condition does not necessarily apply.

Secondly, the nature of the distribution of the deviation must be known. Although a nominal value and two tolerance limits are always given, the distribution can still take many forms within these limits (fig. 18). The normal distribution in fig. 18a is seldom met with

in practice. The distribution in fig. 18b is often found, with a peak whose position changes during the production of the component. This arises because the production process is checked and readjusted only when the tolerance limits are exceeded. The distribution in fig. 18c is found with values that are difficult to measure. The tolerance limits taken by the manufacturer are then, of necessity, very wide, and inspection is made by sample testing to find out whether the product is still in tolerance. It often happens, too, that the manufacturer divides the product into two or more groups by setting an arbitrary limit. This gives two chopped normal distributions (fig. 18d). The position here becomes even more complicated if the peak of the distribution shifts in time or if the manufacturer modifies the selection limit in order to obtain smaller or larger quantities of one of the types.

It will be clear that, with such distribution of the components, it is extremely difficult to obtain suitable data for a statistical design. Furthermore, the method is laborious. Of course, the calculations can be made by an electronic computer, but the programming alone requires a great deal more work than ordinary "pencil and paper" calculation of the "worst case" design for the same unit.

The statistical method has been used only to a limited extent for the circuit blocks described in this article. The worst case method has usually been used. In this method, an extra margin of safety has also been included in the calculations to allow for the component ageing determined by life tests. In this approach we have endeavoured to ensure the maximum long-term reliability, something which, particularly in the industrial field, is of vital importance.

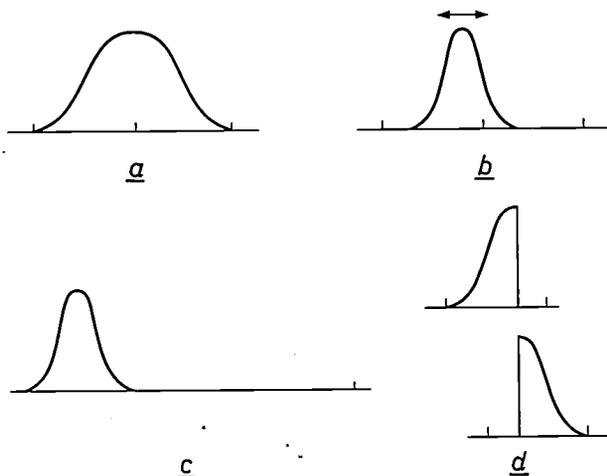


Fig. 18. The different distributions which the value of a component can have within its tolerance limits.

[6] A. I. Pressman, Design of transistorized circuits for digital computers, Rider, New York 1959.

W. Roehr and J. Kane, Transistor flip-flops — worst-case design is your best bet, *Electronic Design* 11: No. 23, p. 58; No. 24, p. 54; No. 25, p. 60; 1963.

W. D. Ashcraft and W. Hochwald, Design by worst-case analysis: a systematic method to approach specified reliability requirements, *IRE Trans. on reliability and quality control RQC-10*, No. 3, 15-21, 1961.

W. Bongenaar and N. C. de Troye, Worst-case considerations in designing logical circuits, *IEEE Trans. on electronic computers EC-14*, 590-599, 1965.

Summary. Digital equipment is made up of a large number of circuits of only a few basic types. These circuits are manufactured as "ready-made" circuit blocks which can easily be fitted to printed wiring boards. Both the design and assembly of digital equipment have been made much simpler and cheaper with these circuit blocks than they would be with conventional components. The article describes two series of circuit blocks, the 10-series and the 20-series, made by Philips. These series include virtually all the circuits normally found in digital equipment, i.e. not only bistable circuits, trigger gates and logic circuits, but also such less frequently occurring types as clock pulse generators, time-delay

units, output amplifiers, etc. The discussion includes the measures taken to combine these circuit blocks as simply as possible to form a complete equipment. To this end, the currents and voltages at the inputs and outputs of the circuit blocks, as well as the signal edges during the transitions from high to low voltage and vice versa, have been standardized to a high degree. The loadability of the circuit blocks can thus be described in very concise loading tables. It follows from the principles chosen for the logic circuits that only NAND circuits are used; this system is explained in detail. Finally, attention is paid to the design of the circuit blocks, based mainly on the "worst case" method.

A technique for depositing thin and thick films

The development of techniques for depositing metals or semiconductors by vacuum evaporation has not kept pace in recent years with the increased use of vacuum-evaporated films. With existing apparatus for deposition, uniform distribution of the material is possible only over relatively small surface areas. Moreover, the thickness of film that can be deposited in one working cycle is limited. To deposit a thick film the process has to be frequently interrupted, and since the evaporation always takes place in vacuum, this costs a great deal of time. For the vacuum deposition of integrated circuits^[1], for example, it would be very useful to have apparatus which could be used for depositing either very thin films (less than 1 μm for resistive elements) or thick films (25 to 100 μm for the contact strips) without having to interrupt the process. It would also be useful to be able to deposit large-area films continuously, e.g. for making rolled-foil capacitors.

We have developed a method in this laboratory which goes a long way towards meeting these requirements. In this method material is continuously supplied during the process to the place where it is evaporated, so that the deposition can continue as long as required; the efficiency of the method is high.

The technique makes use of an evaporator consisting of a long thin porous body *E* of high-melting-point material (*fig. 1*) about 30 cm in length and 1 to 5 mm in diameter. Fitted around the top of this evaporator is a crucible *C* in which the material to be evaporated is heated to slightly above its melting point (e.g. copper to 1100 °C). After the evaporator has also been heated to a high temperature — by passing a large current through it, the dissipation being 1000 to 2000 W — the evaporator takes up the molten material in its pores, and the material flows out of the crucible, assisted by gravity, via the evaporator. The material now evaporates while it sinks down the evaporator. This requires the temperature of the evaporator to be very high (e.g. 1700 °C for copper). Material not evaporated accumulates near the bottom on a collector plate *P*. The substrates on which the material is to be deposited are suspended around the evaporator in a cylindrical holder *H*. To ensure good adhesion of the deposited material the substrates also must be at a fairly high

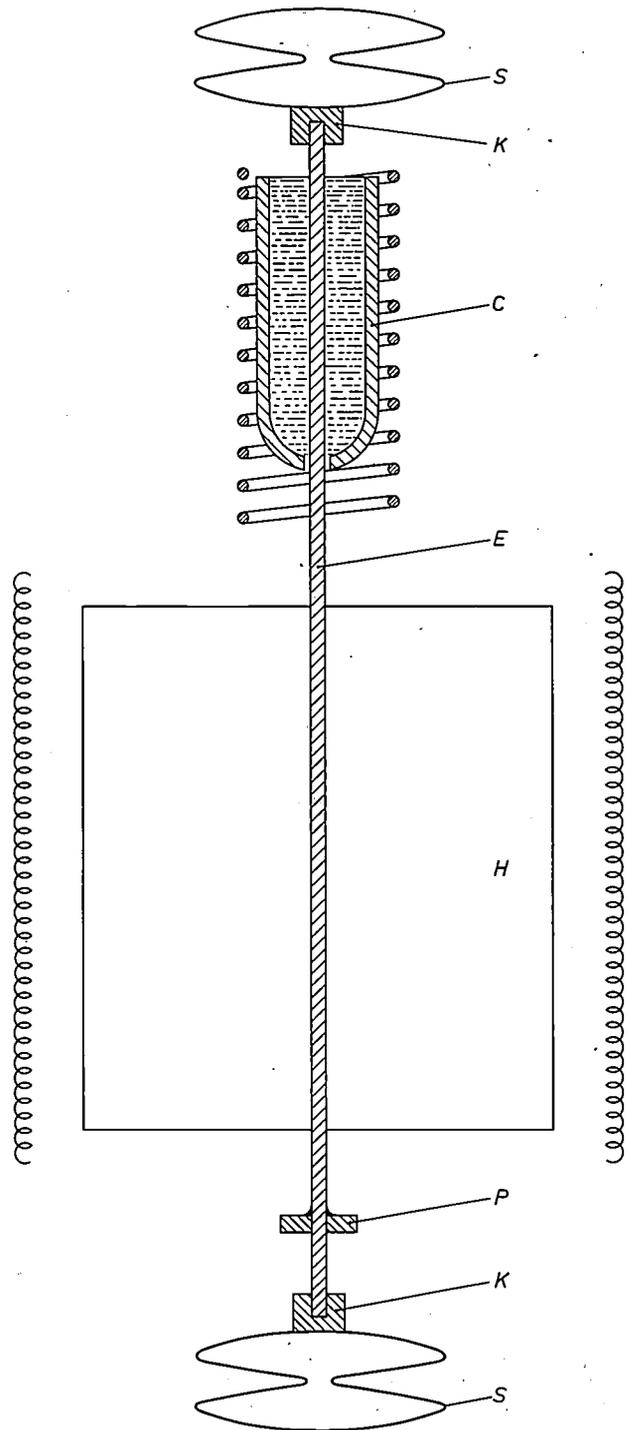


Fig. 1. Sketch of the vacuum-evaporation system. *E* porous evaporator (about 30 cm long). *C* crucible containing the material to be evaporated. The crucible is heated by a resistance element wound round it. *K* contacts through which current is supplied to the evaporator. *S* leaf-springs to keep the evaporator body taut. *H* substrate holder, with heater filaments round it. *P* plate for collecting unevaporated material.

[1] See E. C. Munk and A. Rademakers, *Integrated circuits with evaporated thin films*, Philips tech. Rev. 27, 182-191, 1966 (No. 7).

temperature (for Cu on glass about 300 °C), and for this purpose electric heating elements are disposed around the substrate holder.

If the various temperatures are properly chosen, it is possible in this way to achieve uniform deposition over a considerable area along the evaporator. The requirements to be met by the evaporator are very severe. It must be porous to the evaporated material at the operating temperature, but it must not react with it. It must also be electrically conducting and have a high melting point. An evaporator meeting these requirements can be made by twisting together a few wires of a suitable metal.

Fig. 2 shows a vacuum-evaporation system in which this principle is used, standing on a pump table. The evaporator and the crucible have been heated to a high temperature. At the back of the substrate holder a few glass substrates are arranged on which an aluminium film has been deposited; the evaporator can be seen reflected in the substrates. The electric heater elements around the substrate holder are not connected up. During the evaporation process the complete apparatus is further surrounded by a number of metal heat-screens, which have been left out of the photograph.

A large number of materials, including Cu, Al, Au, Ag, Sb, Bi, Ge and Si, can be vacuum-evaporated by means of this technique. The amount of material contained in the crucible is sufficient to deposit a film of the required thickness for most applications; if necessary a device can be added by means of which the material is fed to the crucible during the evaporation in the form of pieces of wire or grains.

Difficulties may be caused if the capillary action of the evaporator is so strong that it causes the material to flow too quickly from the crucible. Much of the material then has no chance to evaporate and accumulates at the bottom of the evaporator. The rate of evaporation can be increased slightly by raising the temperature of the evaporator, but this inevitably curtails the life of the wire. In such cases two modifications of the method are possible. First, the crucible can be located *below* the evaporator; the strong capillary action which was undesirable in the arrangement of fig. 1 now draws the material up the evaporator, so that it reaches a height enabling uniform evaporation over a large area to be obtained. Secondly, if the material is available in the form of wire, it can quite easily be fed to the evaporator without the use of a crucible. A reel of wire (*B* in fig. 3) is placed near the upper part of the evaporator, and the end of this wire is brought up against the previously heated evaporator. The material melts at this position and flows into the evaporator. If the wire is fed to the evaporator by means of a pair of small rollers *R*, the quantity of material entering the

evaporator can be completely controlled. The equipment can operate automatically for a long period if the roller is driven at the appropriate speed by a small motor. Obviously, all the equipment must be enclosed in the evacuated space.

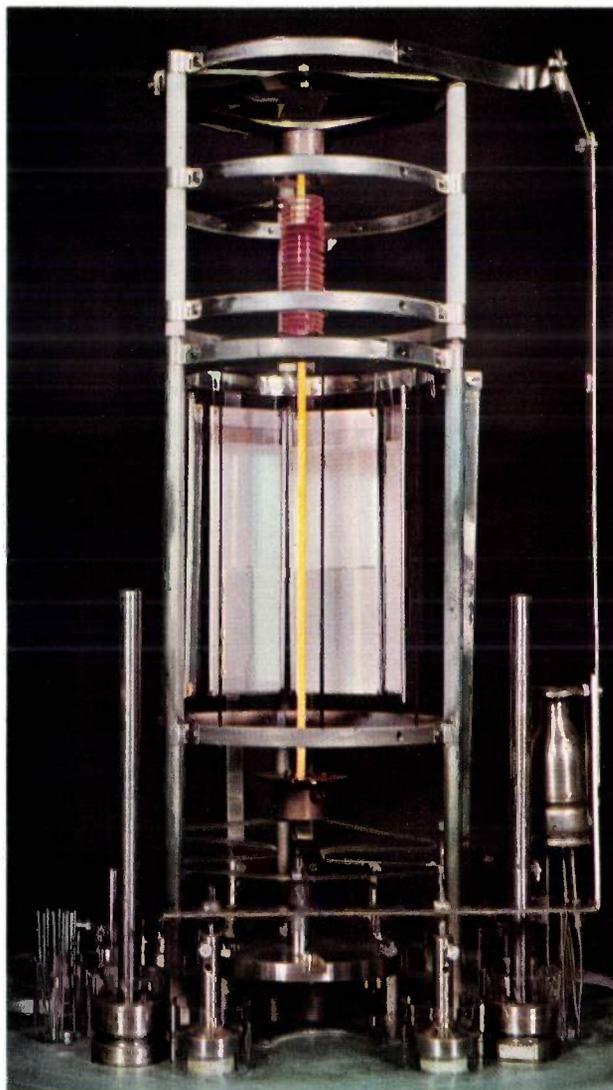


Fig. 2. Vacuum-evaporation apparatus on the pump table. When this photograph was taken the evaporator and crucible had been heated, but the apparatus was not in a vacuum and therefore no evaporation could take place.

In this last modification the evaporator does not have to be vertical; if necessary or more convenient, a horizontal arrangement is possible. The application of the method in this form is limited by the need to have the material available in the form of wire, and in most cases a crucible is therefore used.

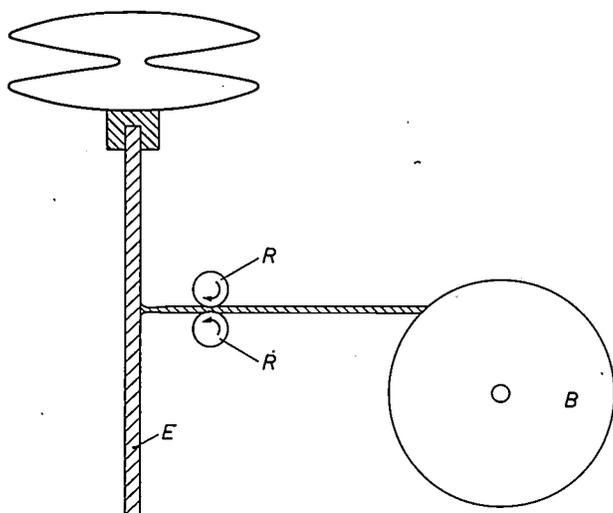


Fig. 3. Wire-feed device. *B* wire reel. *R* rollers for feeding the wire to the evaporator *E*.

Finally, we would also like to mention that when the evaporator has drawn up the material to be evaporated, it can be removed and used as the source in another evaporation equipment. The capacity is then of course limited.

J. J. A. Ploos van Amstel

J. J. A. Ploos van Amstel is with Philips Research Laboratories, Eindhoven.

Experimental electrostatically focused "Plumbicon" tubes

J. H. T. van Roosmalen

In "Plumbicon"^[] tubes developed up to now and also in most ordinary vidicons the electron beam is usually magnetically focused on to the photosensitive layer by means of a long coil encircling the tube. There are advantages to be gained by the use of electrostatic focusing. This article explains the methods used to meet the requirement that the deflected electrons should reassume their original direction so that they approach the photosensitive layer at right angles.*

Electrostatic focusing of the beam; the "landing" of the electrons on the photosensitive layer

In "Plumbicon" tubes developed up to now^[1] and also in most ordinary vidicons, the electron beam scanning the light-sensitive layer is focused magnetically. As well as the deflection coils, there is a much larger coil which ensures that the divergence at the beginning of the beam becomes a convergence, and that the diameter of the beam is as small as possible in the plane of the sensitive layer (target).

This focusing coil can be omitted and the power required for focusing reduced to very nearly zero by using electrostatic rather than electromagnetic focusing. This gives the additional advantage that the picture cannot be rotated, magnified, or reduced in size when the focus is adjusted. This is particularly attractive in colour television, where the camera contains not one, but three or four tubes. Also, in an electrostatically focused tube the deflection cannot affect the focusing.

Just as in magnetically focused tubes, it is a requirement that in the last part of their trajectory, where they pass through a retarding field, the electrons in the beam should move parallel to the axis of the tube. This means that the electrons must change direction again after deflection (*fig. 1*). If this requirement is not met, various undesired effects may appear. In a magnetically focused tube the desired shape for the trajectory can be obtained by locating the deflection coils in a clearly defined position and by very closely defining the shape of the fringe field. No additional auxiliaries are then

needed. In an electrostatically focused tube, on the other hand, a special lens has to be introduced^[2], the correction lens.

The undesired effects in question are chiefly the following^[3]. The least serious effect is that, with the target uniformly illuminated, local variations in the angle of incidence give rise to local non-uniformities in the output signals. This comes about because the potential assumed by a point on the target as the beam passes over it depends on the angle of incidence. A more serious effect is that oblique incidence results in a degree of discharge lag, which gives moving objects a "tail". Furthermore, the brightness

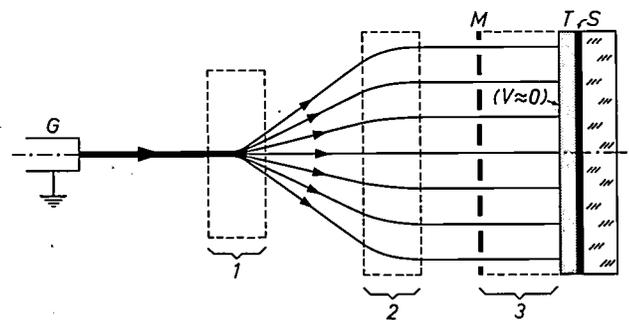


Fig. 1. Ideal electron trajectories in a vidicon or "Plumbicon" tube (diagram). G gun, M mesh, T target, and S signal plate. The potential V of the target surface is almost equal to that of the cathode of the gun. The potential of the mesh is several hundred volts higher.

Except for the divergence in the beam, the electrons first move along the axis of the tube. They then undergo a certain degree of deflection in space 1. In space 2, the trajectories have to become parallel again with the axis, so that, in space 3, the kinetic energy of the electrons can be completely used up in overcoming the retarding field prevailing there.

J. H. T. van Roosmalen is with Philips Research Laboratories, Eindhoven.

[*] Registered Trade Mark for television camera tubes.

of this tail often varies periodically rather than uniformly ("stern wave"). If the deviation from the direction of the axis is large, it can also happen that the potential of a part of the surface of the target does not establish itself at a fixed value but exhibits low-frequency oscillations. There is then a flicker on part of the picture at the receiver screen.

Fig. 2 shows a diagrammatic cross-section of the experimental electrostatically focused "Plumbicon" tubes used in our investigations. The focusing of these tubes is effected by means of a unipotential lens, i.e. a

assembly like that shown in fig. 2 depends to a considerable extent on the position of the deflection coils and on the configuration and potentials of the correction lens. A brief description of the calculations and experiments carried out on this subject is given in this article. These show that useful combinations of the electrically and geometrically relevant values can be found in which the effect of the correction lens is very satisfactory. A certain degree of distortion does occur, but it is anticipated that this can be sufficiently cor-

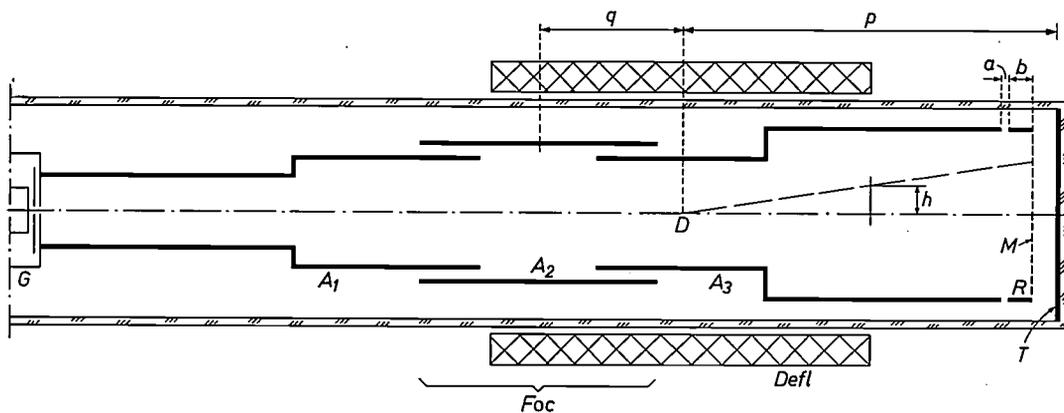


Fig. 2. Full-size diagrammatic cross-section of an experimental electrostatically focused "Plumbicon" tube. *G*, *M* and *T* have the same significance as in fig. 1. The focusing lens *Foc* consists of electrodes *A*₁ and *A*₃, which are at the same potential, and a wider electrode *A*₂ between them, which is at a much lower potential (a unipotential lens). *Defl* deflection coils, the position of which is variable. *D* is the "deflection point". Electrode *A*₃ (at 300 V) and ring *R* (width *b*) carrying the mesh (at a higher potential) form the correction lens.

cylinder in three parts with the centre section at a much lower potential than the outer two. These two outer sections may be regarded as the two separated halves of the anode cylinder of an ordinary "Plumbicon" tube. This electrostatic focusing lens takes up no extra space.

The correction lens also operates electrostatically and is formed by the third electrode of the focusing lens (*A*₃) and a ring *R*, carrying the mesh — this is found in all "Plumbicon" tubes (and other vidicons); it ensures that the retarding field is homogeneous. The potential V_M of ring and mesh is higher than that of the electrode *A*₃. The external dimensions of the experimental tubes as shown in fig. 2 are exactly the same as those of the magnetically focused "Plumbicon" tubes. The diameter of the mesh screen has been chosen at the largest possible in laboratory tubes (24 mm). The results of tests made on these experimental tubes are easy to convert for the mesh diameter usable in practice.

It has been found that the proper operation of an

rected by making use of the fact that the deflection coil can provide distortion of an opposite nature.

Calculations of electron trajectories

The beginning of our theoretical research consisted in calculating, with the aid of an electronic computer, the potential distributions in five hypothetical correction lenses and the trajectories of the electrons in the lenses [4]. In calculating the potential distributions, it

[1] For a description of this see E. F. de Haan, A. van der Drift and P. P. M. Schampers, The "Plumbicon", a new television camera tube, Philips tech. Rev. 25, 133-151, 1963/64.

[2] The first application of such a lens was in a magnetically focused vidicon, in order to improve the resolution, and was made by H. G. Lubszynski and J. Wardley (IEE paper 4006 E, June, 1963). An earlier application in an electrostatically focused tube is described in J. E. Kuehne and R. G. Neuhauser, J. SMPTE 71, 772, 1962.

[3] See L. J. van de Polder, Target stabilization effects in television pick-up tubes, to appear shortly in Philips Res. Repts.

[4] The programme for these calculations was drawn up by Ir. C. Weber of this laboratory.

was assumed that the mesh completely screens the target, so that its potential has no effect at all on the field in the lens. The calculations for each configuration were made for a number of different mesh voltages V_M . The anode voltage V_{A3} was taken as 300 V in all cases.

In each of these cases 48 electron trajectories were calculated consecutively by the machine from the potential distributions found. Both the lens and the deflection coils were taken as having a limited interaction region and the calculations related only to positions of the deflection coils in which the interaction regions were quite separate.

We have restricted ourselves to the trajectories of electrons that move along the axis before reaching the interaction region of the deflection coils. The divergence

of the correction lens. Finally it was assumed that all electrons had the same kinetic energy (eV_{A3}) on entering the interaction region of the lens, and therefore also at the mesh (eV_M).

For every case, the magnitude of the angle α between the beam axis and the direction of the trajectory at the mesh follows directly from the trajectory calculations. Once this angle is known, it is easy to calculate the velocity component v_n in the direction of the axis and from there, by converting the kinetic energy $\frac{1}{2}mv_n^2$ into electron volts, the potential V assumed by the target at the point at which the electrons in question strike it. This potential is, in fact, such that the electrons can only just reach the layer (fig. 4). The difference between the value found for V and the value for a beam for

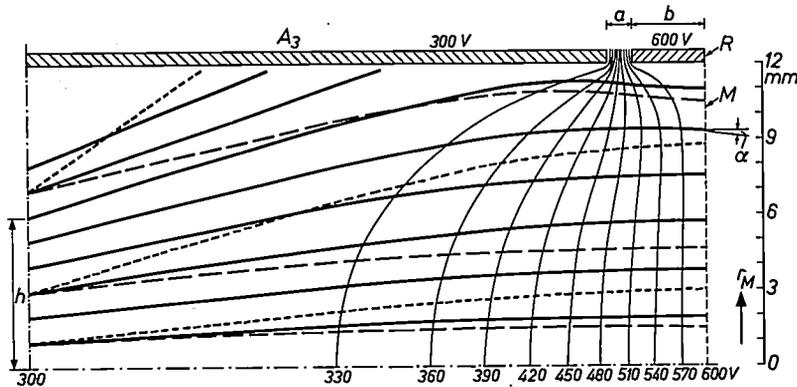


Fig. 3. Half-section of a correction lens ($a = 1.0$ mm and $b = 3.0$ mm) with a few calculated equipotential lines (see potentials inserted) and electron trajectories drawn in. These trajectories are followed if the distance p (fig. 2) is 50 mm and the mesh voltage V_M is 600 V. The deflection is characterized by the height h at which the electrons enter the interaction region of the lens. Electron trajectories with $p = 40$ mm (dotted lines) and $p = 60$ mm (broken lines) are shown for three deflections. r_M is the radial co-ordinate on the mesh and α the deviation in path direction.

of the beam — which, after all, is what makes focusing necessary — is small enough in the tubes in fig. 2 to be neglected. Only the central trajectory of each beam is taken into consideration. Here it was assumed that the whole length of this trajectory lies in a plane passing through the axis of the tube.

The position of the deflection coils was indicated by a point on the axis of the tube, the point D (fig. 2) from which the deflected electrons apparently originate. Each series of 48 trajectories consisted of three groups of 16. Calculations were made for three different positions of D and for 16 different deflections at each position. The deflection was indicated by the height h (fig. 3) at which a deflected electron on a trajectory as described above would enter the interaction region of

$$\begin{aligned}
 e\{V_M - V(r)\} &= \frac{1}{2}mv_n^2; \\
 e\{V(r) - V(0)\} &= \frac{1}{2}mv_t^2 = \\
 &= \frac{1}{2}mv^2 \sin^2 \alpha = \\
 &= eV_M \sin^2 \alpha.
 \end{aligned}$$

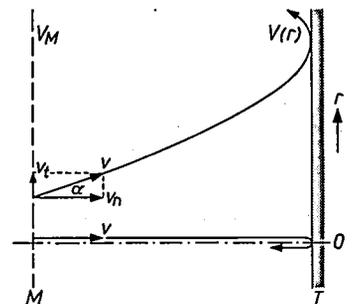


Fig. 4. The electrons describe a parabola in the retarding field between mesh and target. The potential $V(r)$ of a point on the target — r is the distance from the axis — is determined by the kinetic energy corresponding to the axial velocity component v_n of the electrons in question at the mesh. If there is a variation of the angle α with position, $V(r)$ will also vary ("landing error").

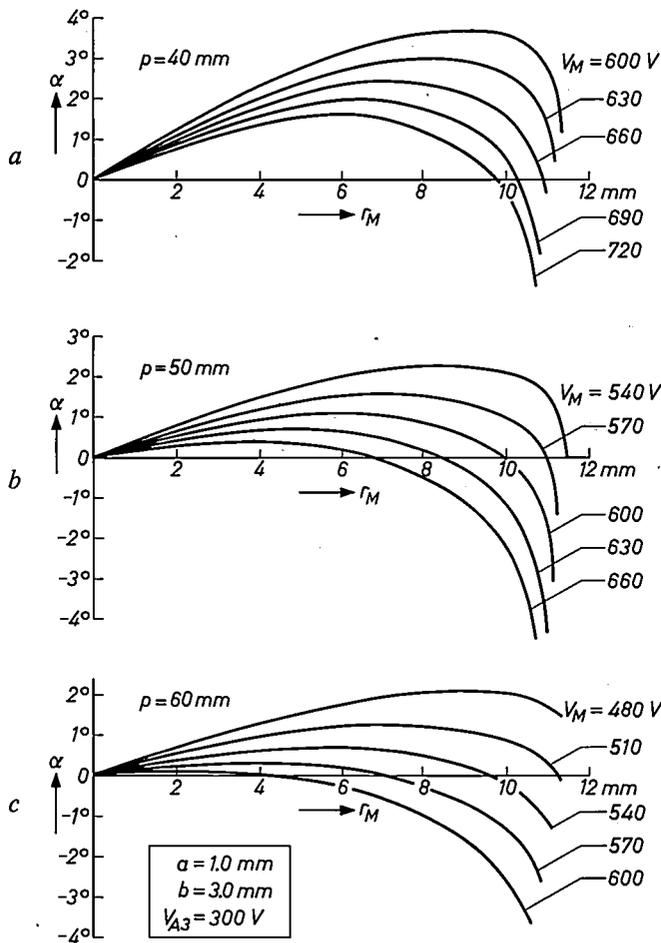


Fig. 5. The angle of trajectory α plotted against the position (r_M) on the mesh for a lens in which $a = 1.0$ mm and $b = 3.0$ mm and for three positions (p) of the deflection coils (cf. figs. 2 and 3). The curves apply to the values of the mesh voltage V_M shown. The angle α is taken as positive if the radial component of velocity of the electrons is directed outwards (fig. 4).

which $\alpha = 0$ determines the difference in the signals produced by the tube, and is referred to as the landing error [5]. It can easily be shown that this difference is equal to $V_M \sin^2 \alpha$.

Fig. 5 shows how α varies with location on the mesh, i.e. with the distance r_M from the point on the mesh to the axis of the tube, for some of the cases calculated. The three graphs relate to the same lens configuration (and anode voltage V_{A3}) but each of them applies to a different position (p) of the deflection point. The curves in each graph apply to different mesh voltages V_M .

Fig. 6 gives examples of the curves resulting from plotting the landing error, as defined above, against the co-ordinate r of the point on the layer at various mesh

voltages V_M . (Fig. 6b relates to the same situation as fig. 5b.)

In order to be able to assess the usefulness of a given combination, two aspects have to be considered: 1) the size of the area within which the landing error has an acceptable value, i.e. no larger than a few volts, and 2) the maximum landing error in this area. If, for instance, we consider fig. 6b, we see that for $V_M = 660$ V the landing error in the area within $r = 7$ mm is extremely small: its maximum value there is only 0.04 V. Nevertheless the situation at $V_M = 600$ V is more useful since the landing error here remains within reasonable limits over a much wider area.

As may be seen from fig. 6, the landing error is zero for $r = 0$ (i.e. on the axis) and also for one other value of r ; we refer to this as r_0 . Let the maximum landing error in the area $r < r_0$ be F . Since the landing error at the edge increases rapidly as r increases, the value r_F of r at which the landing error is again equal to F is a good measure of the radius of the useful portion of the target. The way in which r_F varies with b at $p = 50$ mm

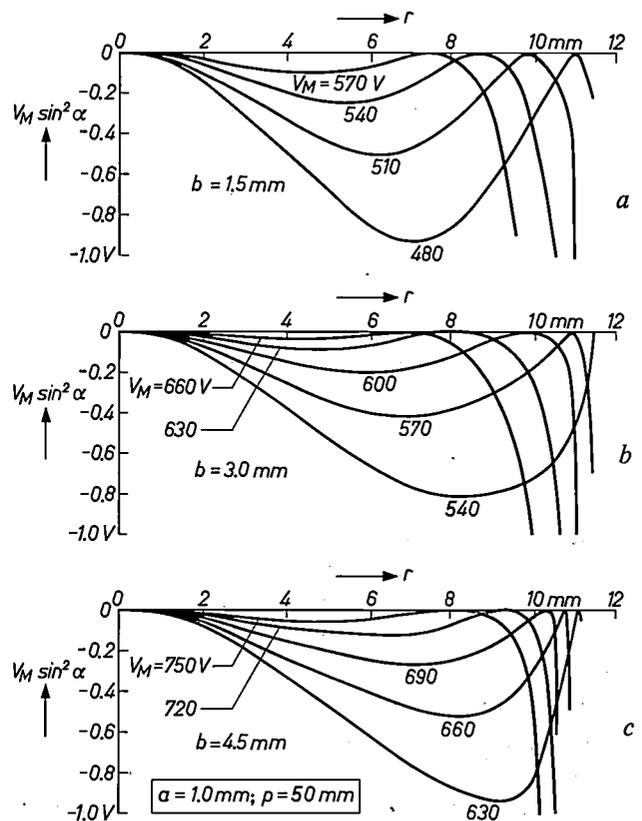


Fig. 6. Curves of the landing error $V_M \sin^2 \alpha$ (in volts) plotted against the position r on the layer. The curves apply for a deflection point distance p of 50 mm and to the lens parameters a, b shown. The various curves apply to the mesh voltages shown. There is a zone round the axis where the error is fairly small at an appropriate mesh voltage, falling to zero at one particular radius. Beyond it the error rapidly increases.

[5] For a more detailed treatment of the way in which the electron beam strikes the sensitive layer, see reference [3] or J. Castleberry and B. H. Vine, J. SMPTE 68, 226, 1959.

is shown in fig. 7 for five different values of F which are no greater than 0.5 V. For $0.5 \text{ V} \geq F > 0.1 \text{ V}$, there is an optimum at values of b around 3 mm.

There is also an optimum in this region for $p = 60 \text{ mm}$. Our calculations have in addition shown that, with the same landing error F , the radius r_F becomes larger as p increases.

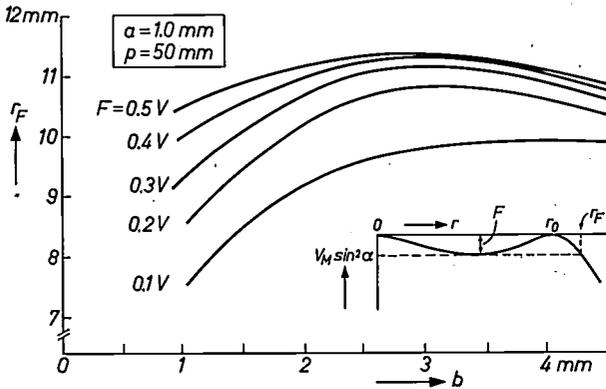


Fig. 7. The radius r_F of the circle at which the landing error is equal to the maximum F of the landing errors occurring at smaller circles (inset) as a function of the lens parameter b . The curves apply for the values of F shown. There is an optimum at $b \approx 3 \text{ mm}$.

Distortion

It has already been pointed out that the use of a correction lens results in a certain degree of distortion. This distortion, which is of the barrel type, naturally varies with the lens configuration, the voltages applied and the location of the deflection point. The way in which the relative distortion varies with r_1 , i.e. the value of r_M that would occur without distortion, has been calculated for all cases. The results of these calculations are summarized in fig. 8.

As may be seen, the distortion is smaller at greater values of p just like the landing error. With regard to the effect of b there is no similarity to the behaviour of the landing error: broadly speaking, the greater b is, the greater the distortion; there is no optimum. This can, however, initially be left out of consideration in assessing the lens configuration, since the distortion can, in principle, be corrected.

Experimental investigations

The landing errors occurring in a given case can be measured by uniformly illuminating the target and measuring the strength of the signals delivered by the various elements of the layer. In effect, the potential at each point on the layer is then measured. If there is no landing error, the signals are uniform everywhere

(uniformity measurements). Such measurements, the details of which will be outlined later, were made on electrostatically focused "Plumbicon" tubes with six different correction lens configurations. The tubes differed in the width b of the ring carrying the mesh (figs. 2 and 3), all other dimensions being the same. The width a of the gap between the anode and the ring was 1.0 mm, and b varied from 1.5 to 4.0 mm. Measurements were made with $p = 40 \text{ mm}$, 50 mm and 60 mm.

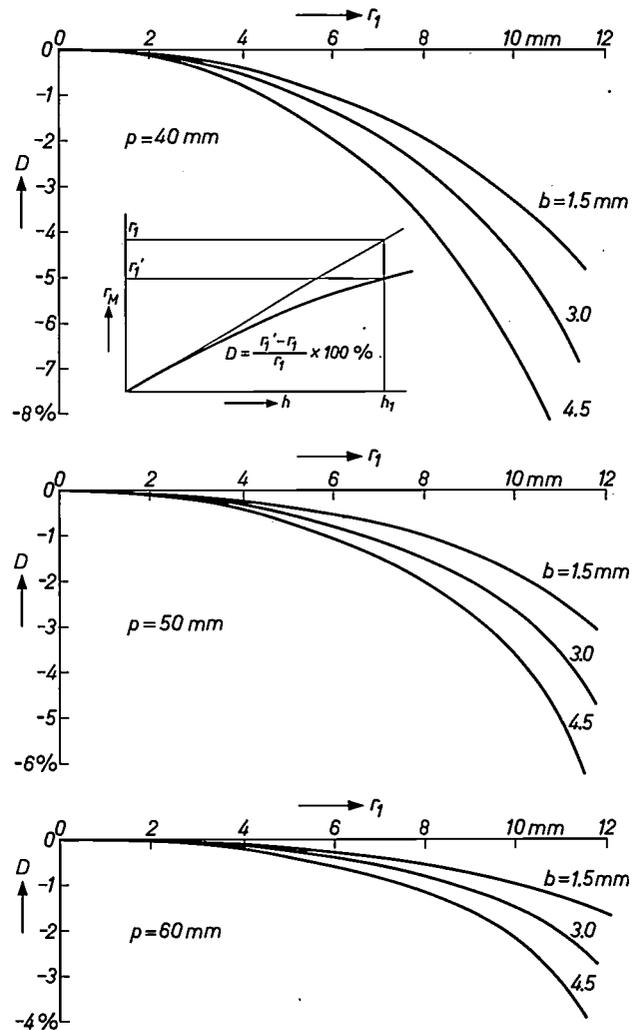


Fig. 8. The percentage distortion D in the plane of the mesh as a function of the radius r_1 at which the beam would pass through the mesh if there were no distortion (see inset; cf. fig. 3). The graphs apply for $p = 40, 50$ and 60 mm . The curves apply for the lens parameters shown and the optimum mesh voltage for these parameters.

A few more general tests preceded these measurements. For example, experiments were made to see whether there was any advantage to be gained by making the distance a larger or smaller than 1 mm. It was found that there was not.

The first test measurements also showed that the position of the deflection coils affected not only the uniformity but also the focus at the corners of the picture. To obtain good focus it was found necessary for the interaction regions of the focusing lens and the deflection coils to be quite separate. (As we mentioned above, this was already a basic assumption in the theoretical work.) To meet this condition, the distance q (fig. 2) between the centre of the lens and the deflection point had to be at least equal to the diameter of the lens.

In accordance with the theoretical result, it was found on the other hand that shifting the deflection coils towards the correction lens adversely affected the uniformity of the output signals and the distortion. These contradictory requirements — the deflection coils has to be well spaced from the focusing lens and also from the correction lens — could be met, without altering the length of the tube, by using a set of deflection coils slightly shorter than that used in magnetically focused "Plumbicon" tubes. When these coils were used (for their length, see fig. 2), the distance p could, indeed, be varied over a reasonably wide range around $p = 50$ mm without undesired effects.

Measuring methods and results

The location-dependence of the signal current i_t of the "Plumbicon" tubes used with these shorter coils was examined in the following manner. A grid of parallel wires was placed in front of the window outside the tube, their direction being perpendicular to the scanning direction. (The distance between the centres of the wires was 1 mm, and their thickness was 0.1 mm.) The signals obtained could thus be correlated with positions on the target quite simply. Except for the shadows of the wires, the window was uniformly illuminated. The signal plate voltage U was set to a much lower value than normal for the tube — 2 V as against 30-50 V — to make the difference in the potential $V(r)$ of the surface of the target (cf. fig. 4) provide the greatest possible relative difference in the signal current i_t .

Red light was used for illumination, because the tubes are most sensitive to this kind of light at very low signal plate voltages. The deflection amplitude of the electron beam was made about twice as great as the diameter of the target in order to reduce the effect of any non-linearity as far as possible. If the scanning electron beam does not pass over every part of the target at the same velocity, this can in itself form a source of non-uniformity in the signals. The only output signals measured were those delivered by the tube when the electron beam moved along a diameter of the screen. The mesh voltage was set to the same series of values used in the trajectory calculations.

An alternative method is the one in which the signal plate voltage is gradually allowed to increase from zero and the values are measured at which the various points (circles) of the screen of a connected picture tube start to give out light. The potential of the signal plate has then become just higher than the potential $V(r)$ assumed by the scanned side of the target at that point. This method of measurement eliminates the effect of the spread in the kinetic energy of the beam electrons. Only the fastest electrons in the beam reach the target at the signal plate voltage at which the picture tube just begins to give out light. But these electrons are the very ones that determine the potential $V(r)$ and thus the landing error. It has been found in practice that the method used, which is simpler, is very nearly as accurate.

By way of example, fig. 9 gives the results of a series of measurements at $p = 50$ mm. As may be seen, the area in which $i_t/i_t(0)$ remains close to 1.0 (the useful area) has the greatest diameter for $b = 3.0$ mm. For

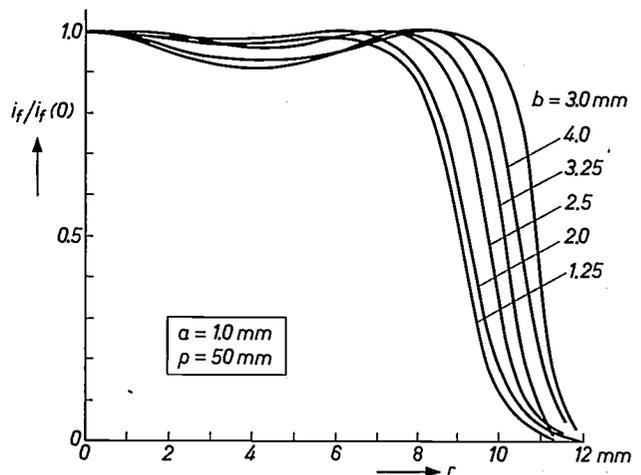


Fig. 9. The output signal i_t measured along a diameter (normalized at the value for $r = 0$), plotted against the position co-ordinate r on the target. The graphs apply to lenses with $a = 1.0$ mm and $p = 50$ mm. The diameter of the useful zone is largest when $b = 3.0$ mm.

$p = 60$ mm also, the configurations with b at about 3.0 mm are found to provide the largest useful area. This is in complete agreement with the theoretical results (cf. fig. 7).

In order to be able to make a further comparison with the theoretical results, we have converted the maximum ordinate variation of the horizontal part of the curves into a landing error in volts with the aid of the i_t-U characteristic. With $p = 50$ mm and $b = 3.0$ mm we then find a value of 0.3 V, and with $p = 60$ mm and $b = 3.0$ mm, about 0.15 V. Here also, there is close agreement with the theoretical values.

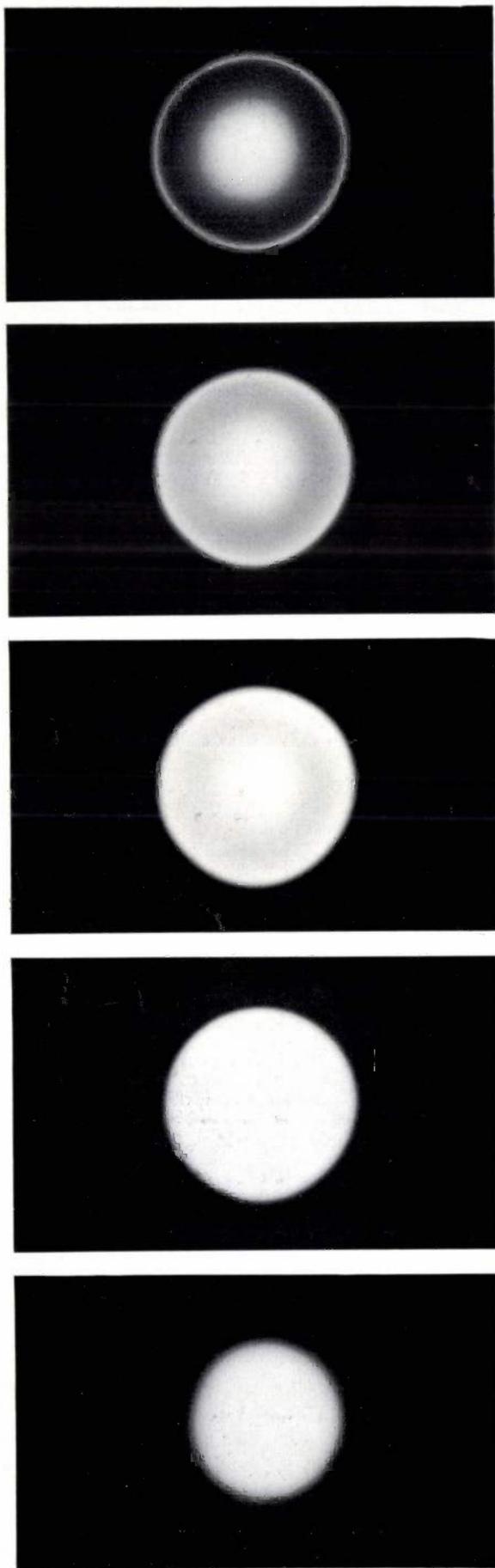


Fig. 10 gives a neat qualitative demonstration of the effect of the variation of the mesh voltage with the same correction lens. Calculations showed (cf. *fig. 6*) that in general the landing error is zero not only at the axis but also on a circle around it, and, in addition, that increasing the mesh voltage does lead to better uniformity in the picture, but in the long run decreases the useful part of the screen. This is in complete agreement with the results shown in the five photographs in *fig. 10*.

Measurements on distortion show that here also the calculations have provided reliable results.

Summary. In "Plumbicon" and also in ordinary vidicon tubes, the beam electrons are required to move in the axial direction of the tube in the homogeneous retarding field between the mesh and the target. The electrons therefore have to regain their original direction after deflection. This may be achieved in electrostatically focused tubes by mounting the mesh on a separate ring which, with the neighbouring electrode of the focusing lens, forms an additional lens. Calculations and experiments dealing with the best arrangement of this lens, i.e. with gap-width, ring-width and voltages, have been made on experimental electrostatically focused "Plumbicon" tubes with shortened deflection coils. The results agreed closely. There is a clear optimum. The ideal width of the ring with a mesh diameter of 24 mm, a deflection-point-to-mesh distance of 50 mm and a gap-width of 1 mm was 3 mm. The results can be converted for the practically useful diameter. The correction lens produces a certain degree of barrel distortion, which appears on the screen as pin-cushion distortion, but this can be corrected.

Fig. 10. Photographs of the screen of a picture tube connected to a "Plumbicon" tube under investigation, to show the effect of variations in the mesh voltage on the landing error; local differences in the landing error lead to non-uniformity of the picture. The signal plate voltage was deliberately made very low (2 V) to exaggerate the non-uniformity as compared with that found in normal use. The mesh voltages corresponding to the photographs are, in descending order, 540-570-600-660-720 V. Increasing the mesh voltage does give better uniformity, but also leads in the long run to a diminution in the diameter of the useful part of the target.

Recent scientific publications

These publications are contributed by staff of laboratories and plants which form part of or co-operate with enterprises of the Philips group of companies, particularly by staff of the following research laboratories:

Philips Research Laboratories, Eindhoven, Netherlands	<i>E</i>
Mullard Research Laboratories, Redhill (Surrey), England	<i>M</i>
Laboratoires d'Electronique et de Physique Appliquée, Limeil-Brevannes (S.O.), France	<i>L</i>
Philips Zentrallaboratorium GmbH, Aachen laboratory, Weisshausstrasse, 51 Aachen, Germany	<i>A</i>
Philips Zentrallaboratorium GmbH, Hamburg laboratory, Vogt-Kölln-Strasse 30, 2 Hamburg-Stellingen, Germany	<i>H</i>
MBLE Laboratoire de Recherche, 2 avenue Van Becelaere, Brussels 17 (Boitsfort), Belgium.	<i>B</i>

Reprints of most of these publications will be available in the near future. Requests for reprints should be addressed to the respective laboratories (see the code letter) or to Philips Research Laboratories, Eindhoven, Netherlands.

75 jaar Philips.

59 short papers, compiled by H. E. Kater.
Polytechn. T. A/E **21**, 842A-891A/691E-740E, 1966 (Nos. 20).

J. Adams & B. W. Manley: The mechanism of channel electron multiplication.
IEEE Trans. on nuclear science NS-13, No. 3, 88-99, 1966. *M*

R. Astor & J. Cayzac: Equipement pour liaison troposphérique utilisant la modulation d'amplitude à porteuse supprimée.
Acta electronica **9**, 241-254, 1965 (No. 3). *L*

Y. Beurel & G. E. Zaaijer: Equipements UHF de réception à faible bruit pour transmissions à grande distance.
Acta electronica **9**, 255-268, 1965 (No. 3). *L*

A. L. Biermasz: Enige trends in de ontwikkeling van elektronische meetinstrumenten.
T. Ned. Elektronica- en Radiogen. **31**, 129-136, 1966 (No. 6).

G. Blasse: Sodium lanthanide oxides NaLnO_2 .
J. inorg. nucl. Chem. **28**, 2444-2445, 1966 (No. 10). *E*

G. Blasse & A. Brill: On the Eu^{3+} fluorescence in mixed metal oxides, II. The 5D_0 - 7F_0 emission.
Philips Res. Repts. **21**, 368-378, 1966 (No. 5). *E*

P. F. Bongers & U.ENZ: A bistable resistor on the basis of vanadium oxide.
Philips Res. Repts. **21**, 387-389, 1966 (No. 5). *E*

J. C. Brice, R. E. Hunt, G. D. King & H. C. Wright: Inhomogeneities in the electrical properties of gallium arsenide crystals.
Solid-State Electronics **9**, 853-857, 1966 (No. 9). *M*

H. Bruining: Überblick über Ergebnisse und Ziele der Forschung auf dem Gebiet der Elektronenröhren.
Nachrichtentechn. Z. **19**, 381-385, 1966 (No. 7). *A*

K. H. J. Buschow: The crystal structures of the rare-earth compounds of the form R_2Ni_{17} , R_2Co_{17} and R_2Fe_{17} .
J. less-common Met. **11**, 204-208, 1966 (No. 3). *E*

K. H. J. Buschow: Das Zustandsbild Erbium-Kobalt.
Z. Metallk. **57**, 728-731, 1966 (No. 10). *E*

J. R. Chamberlain, D. H. Paxman & J. L. Page: The visible fluorescence of Er^{3+} in yttrium gallium garnet.
Proc. Phys. Soc. **89**, 143-151, 1966 (No. 1). *M*

J. A. Cundall & A. P. King: Comparison of magnetic measurements and theoretical predictions for nickel-iron films over a wide composition range.
Phys. Stat. sol. **16**, 613-619, 1966 (No. 2). *M*

J. A. Cundall & A. P. King: The angular lag of the mean magnetization direction due to magnetization ripple in nickel-iron films.
Brit. J. appl. Phys. **17**, 1105-1107, 1966 (No. 8). *M*

J. R. Dale: Alloyed semiconductor heterojunctions.
Phys. Stat. sol. **16**, 351-387, 1966 (No. 2). *M*

W. F. Druyvesteyn, G. J. van Gorp & C. A. A. J. Greebe: Helicon-like resonances in superconducting niobium.
Physics Letters **22**, 248-249, 1966 (No. 3). *E*

C. Ducot: Les télécommunications à grande distance par faisceaux dirigés. Quelques aspects de leur évolution.
Acta electronica **9**, 211-214, 1965 (No. 3). *L*

C. Ducot: Problèmes de bruit liés à la transmission de signaux aux radiofréquences très élevées et aux fréquences optiques.
Acta electronica **9**, 215-226, 1965 (No. 3). *L*

C. Ducot: Systèmes de modulation utilisables dans les liaisons hertziennes à grande distance.
Acta electronica **9**, 227-240, 1965 (No. 3). *L*

- P. Eckerlin:** Eine einfache und genaue Methode zur Bestimmung kleiner Winkel zwischen ebenen Oberflächen und kristallographischen Netzebenen von Einkristallen.
Z. Instrumentenk. **74**, 319-322, 1966 (No. 10). *A*
- G. Engelsma:** The influence of light of different spectral regions on the synthesis of phenolic compounds in gherkin seedlings in relation to photomorphogenesis, III. Hydroxylation of cinnamic acid.
Acta bot. neerl. **15**, 394-405, 1966 (No. 2). *E*
- U. Enz:** Magnetische voorkeursrichtingen in hexagonale oxiden.
Ned. T. Natuurk. **32**, 241-248, 1966 (No. 8). *E*
- F. C. Eversteijn:** Low-temperature deposition of alumina-silica films.
Philips Res. Repts. **21**, 379-386, 1966 (No. 5). *E*
- C. A. A. J. Greebe, W. F. Druyvesteyn & A. J. Smets:** Helicon resonances in flat boxes of pure indium.
Physics Letters **22**, 246-247, 1966 (No. 3). *E*
- W. Haidinger, J. C. Courvoisier, P. J. W. Jochems & L. J. Tummers:** Controlled doping of germanium layers made by the evaporation-condensation method.
Solid-State Electronics **9**, 689-693, 1966 (No. 7). *E*
- K. H. Härdtl & P. Gerthsen:** Explanation of the Hall effect of LaCoO_3 in the polaron picture.
Solid State Comm. **3**, 283-284, 1965 (No. 9). *A*
- K. E. Johnson:** A subjective investigation of some errors in the chrominance signal decoding circuits of colour television receivers.
Radio and electronic Engr. **31**, 345-357, 1966 (No. 6). *M*
- C. Kooy & J. M. Nieuwenhuizen:** Structural effects in thin films observed by electron microscopy of the film cross-section.
Basic problems in thin film physics, Proc. int. Symp., Clausthal-Göttingen 1965, p. 181-187; Vandenhoeck & Ruprecht, Göttingen 1966. *E*
- H. de Lang:** Derivation of the relation between two weakly coupled nonlinear optical oscillators.
Appl. Phys. Letters **9**, 205-207, 1966 (No. 5). *E*
- R. Memming & G. Schwandt:** Potential distribution and formation of surface states at the silicon-electrolyte interface.
Surface Sci. **5**, 97-110, 1966 (No. 1). *H*
- A. Meyer:** The use of the convolution theorem and the generalized sampling theorem in evaluating arbitrary arrays.
IEEE Trans. on antennas and propagation **AP-14**, 503-505, 1966 (No. 4). *E*
- H. Mooijweer:** Comment on "analysis of the amplification by means of a negative impedance".
T. Ned. Elektronica- en Radiogen. **31**, 77-79, 1966 (No. 4). *E*
- A. W. Moore:** Pyrolytic graphite.
Ned. T. Natuurk. **32**, 221-232, 1966 (No. 7). *E*
- J. Neirynek:** Maximally flat attenuation and delay characteristics.
Electronics Letters **2**, 351, 1966 (No. 9). *B*
- D. J. van Ooijen & G. J. van Gorp:** Motion and pinning of flux in superconducting vanadium foils, studied by means of noise.
Philips Res. Repts. **21**, 343-367, 1966 (No. 5). *E*
- L. J. van der Pauw:** The planar transducer — a new type of transducer for exciting longitudinal acoustic waves.
Appl. Phys. Letters **9**, 129-131, 1966 (No. 3). *E*
- E. Scharrer, K. Böke, J. Schnell, H. Polnitzky, K. Hannig & H. Schildknecht:** Peltier-Kühlbatterien in der chemischen Apparatechnik.
Chemie-Ing.-Technik **37**, 1039-1046, 1965 (No. 10). *A*
- P. J. Severin:** The influence of an electric field along the cathode of a cold-cathode glow discharge.
Philips Res. Repts. **21**, 325-342, 1966 (No. 5). *E*
- M. J. Sparnaay:** Water structure and interionic interaction.
J. Colloid and Interface Sci. **22**, 23-31, 1966 (No. 1). *E*
- A. J. A. van Stratum:** Reactivation of impregnated and "L" cathodes after exposure to air.
Rev. sci. Instr. **37**, 1080, 1966 (No. 8). *E*
- T. L. Tansley:** Heterojunctions applied.
New Scientist **31**, 316-318, 1966 (No. 508). *M*
- N. C. de Troye:** A generalized two-input flip-flop and its realization.
Philips Res. Repts. **21**, 390-409, 1966 (No. 5). *E*
- A. G. van Vijfeijken:** Analogies of the Ettingshausen and Peltier effects in the mixed state of type II superconductors.
Physics Letters **23**, 65-67, 1966 (No. 1). *E*
- M. T. Vlaardingerbroek & G. Wierda:** Theory of Cerenkov coupling to beam-plasma systems.
Electronics Letters **2**, 368-370, 1966 (No. 10). *E*
- K. Walther:** Quantenresonanzen der Ultraschallverstärkung in Wismut.
Z. Naturf. **21a**, 1443-1462, 1966 (No. 9). *H*
- K. R. U. Weimer & H. Bodt:** Cerenkov coupling to beam-plasma systems.
Electronics Letters **2**, 368, 1966 (No. 10). *E*
- J. S. van Wieringen & J. G. Rensen:** Mössbauer measurements in permanent magnets.
Z. angew. Physik **21**, 69-70, 1966 (No. 2). *E*
- H. C. Wright & G. A. Allen:** Thermally stimulated current analysis.
Brit. J. appl. Phys. **17**, 1181-1185, 1966 (No. 9). *M*
- G. E. Zaaijer:** Contribution à la théorie des amplificateurs paramétriques à résistance négative.
Acta electronica **9**, 269-291, 1965 (No. 3). *L*

PHILIPS TECHNICAL REVIEW

VOLUME 28, 1967, No. 3/4



In the countries where the Philips group of companies operates, the group is thought of as an electrical engineering enterprise. It has sometimes seemed a little surprising that the main activity of one of the subsidiary companies, N.V. Philips-Duphar, is the manufacture of pharmaceutical products. The explanation for this apparently incongruous development was simple: the lamp company, investigating the possible applications of the radiation emitted by its products, became interested in the production of vitamin D by ultra-violet irradiation of provitamin and after making a discovery which brought considerable advances, the company decided to undertake the large-scale manufacture of vitamin D themselves. This was in 1936. Wide-ranging research in the fields of organic chemistry and biochemistry and application research accompanying this new activity subsequently led to the manufacture of other groups of substances, and in a few decades the subsidiary enterprise had become not only one of the biggest producers of vitamins but also a leading manufacturer of pesticides, pharmaceutical specialities, etc.

Work at Philips-Duphar is described in this issue of Philips Technical Review. The articles fall roughly into three groups in accordance with the three main areas of activity — the care of man, animal and plant. The first article reports on the development of the retro-steroids, a new class of substances with hormone action;

rather unexpectedly Duphar's original product, vitamin D, has led to research in the (anti-)fertility problem. The next article deals with the cultivation of virus vaccines, which are becoming increasingly important in human and animal medicine. The transition to the agrarian sector is marked by the third article, which deals with the use of vitamins in animal feeds. Then follows an article which throws light on a very general problem concerning the effect of chemical compounds on the living cell: the problem of the "availability"; this plays an essential role in the action of pesticides. Next, by way of an interlude there is a description of the work being done by the biochemical group of the Philips Research Laboratories at Eindhoven on the formative effect of light on plants

(photomorphogenesis); one of the main objects of this work is to learn more about the regulating mechanisms involved in this process — and therefore also in plant protection with chemical agents. In the sixth article we arrive at the complex problem of the "formulation" of pesticides, i.e. the problem of diluting the relevant chemical compounds, selected for their specific action, in forms suitable for practical applications, such as spraying from aircraft.

The last article of the series draws on some instructive and also quaint facets of the history of man's struggle against agricultural pests and diseases through the ages — and thus may help to bring into perspective the work being done today in this field.

Gaps in the picture of this industry are filled in with a few full-page illustrations, depicting chemical mass production, the intricate art of handling radioactive isotopes — which also come within the Philips-Duphar sphere of operations [*] — and the elaborate field tests on crops which Philips-Duphar is carrying out in countries all over the world.

[*] This part of Philips-Duphar's work has been dealt with fairly extensively in: A. H. W. Aten and J. Halberstadt, The production of radio-isotopes, *Philips tech. Rev.* **16**, 1-12, 1954/55. See also *Philips tech. Rev.* **21**, 361, 1959/60.

Retro-steroids

A new class of compounds with sex-hormone action

O. A. de Bruin, H. F. L. Schöler and J. N. Walop

Introduction

The interdependence of the various organs, tissues and cells of an organism requires the presence of a communication system by means of which the different parts can act upon one another. If a rapid reaction is required the human body generally makes use of the nervous system, while for less rapid but longer-lasting effects the hormone system comes into play. Both systems are very closely co-ordinated.

The hormones, which bring about their effects via the bloodstream, are produced in glands with internal secretion, or endocrine glands. The sex hormones are produced in the pituitary gland, in the ovaries of the female, and in the testes of the male. The pituitary gland is a tiny gland lying at the base of the brain and closely connected with the nervous system. In man and woman it produces identical hormones, which have a stimulating action on the production centres in the sex glands and are therefore called "gonadotropic hormones" (gonad = sexual gland, tropic = directed towards). The hormones produced in the ovaries and in the testes are different; the most important ones are the estrogens and progesterone in the ovaries and testosterone in the testes. In addition to their specific sex-hormone action in male and female, these affect in their turn the production of gonadotropic hormones. Control of the production of hormones is thus obtained by means of feedback.

The gonadotropic hormones and the hormones produced in the sex glands belong to two different classes of chemical compounds, the polypeptides (proteins) and the steroids respectively.

The steroid group of sex hormones began to receive considerable attention in about 1930 from many research workers, both in scientific and industrial laboratories. This was not because the steroid hormones are medically more important than the others, but because it has been found possible to produce steroid hormones and variants of them synthetically. This is not yet possible for the gonadotropic hormones.

Steroids are abundantly represented in living organisms, both animal and vegetable, and they are also interesting because, although structurally closely re-

lated, they exhibit widely different physiological action.

Digitalis, a drug which is prepared from the foxglove, has an action on the human heart and has been used since 1785 in the treatment of heart disorders. Cholesterol is found in the blood and in all tissues of the animal organism; it plays an important part in the building up of cell membranes. Bile acids play a part in the digestion of fats. Provitamin D is converted by ultra-violet irradiation into vitamin D, which prevents rickets. Finally, the steroids also include the cortical adrenal hormones, which fulfil such an essential function in the body that the removal of the adrenal gland results in death within a few days.

In about 1930, at the time when the chemical structure of the sex hormones was established, intensive investigations were started with the object of synthesizing naturally-occurring steroids. Later on, efforts were made, by introducing all kinds of structural variations, to prepare substances possessing a more powerful or better, i.e. more selective, action than the natural hormones. These efforts succeeded remarkably well. Steroid research received a fresh stimulus in 1949, when cortisone, an adrenal cortical hormone, was found to alleviate rheumatoid arthritis. The latest powerful incentive to further research on steroids was the discovery of the contraceptive action of some steroids: the invention of "the pill", about ten years ago.

The interest of Philips-Duphar in steroids is as old as this company itself. As long ago as 1938 an article appeared in this journal, describing how Philips embarked upon vitamin D research as a result of efforts to find applications for ultra-violet lamps [1].

Chemistry of the steroids

The steroids are defined chemically as a class of organic compounds with a polycyclic structure, consisting of four linked carbon rings: three rings of six carbon atoms and one of five. The carbon skeleton, with the conventional numbering of the carbon atoms and with the letters denoting the rings, is given in *fig. 1a*; *fig. 1b* gives a spatial representation of the carbon skeleton.

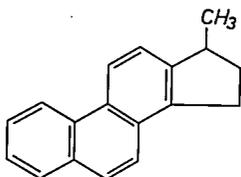
Dr. O. A. de Bruin, Drs. H. F. L. Schöler and Dr. J. N. Walop are with N.V. Philips-Duphar, Research Laboratories, Weesp, Netherlands.

[1] A. van Wijk, Philips tech. Rev. 3, 33, 1938.

[2] See L. F. Fieser and M. Fieser, Steroids, Reinhold, New York 1959.

Fig. 2 shows the structural formulae of several steroid compounds which are mentioned in this article, in particular the female sex hormones estradiol and progesterone, the male sex hormone testosterone, and ergosterol (a compound related to cholesterol).

The work of analysing these structures has been long and laborious; for an account of this the reader is referred to the literature [2]. We shall only mention here that treatment of different compounds with selenium led to the formation of one and the same product, methylcyclopentenophenanthrene:



This proved clearly and elegantly that the above compounds were indeed related.

The main function of the selenium is that it removes hydrogen from the six-membered-carbon rings of the steroid skeleton, so that these become unsaturated. This aromatization, as it is called, is accompanied by the removal of fragments from the carbon skeleton such as the methyl groups 18 and 19. The living organism is also capable of aromatization, but in this case it generally remains limited to ring A. We shall return later in more detail to this physiologically extremely important reaction in the living organism.

The saturated steroid skeleton contains various *asymmetric carbon atoms*, that is to say, carbon atoms which are linked by their four valency bonds to four

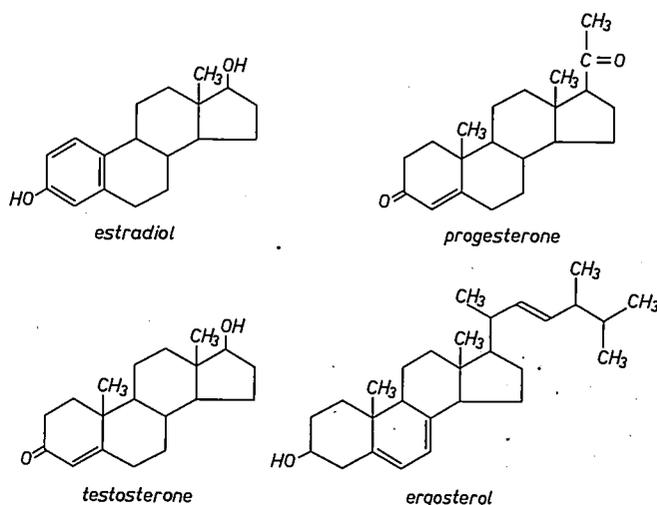


Fig. 2. Structural formulae of the female sex hormones estradiol and progesterone, of the male sex hormone testosterone, and of ergosterol. Ergosterol yields vitamin D₂ under ultra-violet irradiation. A by-product of this reaction is lumisterol₂, the original starting material from which Philips-Duphar synthesized retro-steroids; see figs. 3, 4 and 5.

different atoms or groups of atoms. Two spatial arrangements are possible for an asymmetric carbon atom, and each is the mirror image of the other.

As can be seen in fig. 2 testosterone and progesterone have six asymmetric carbon atoms; estradiol has one less and ergosterol has as many as eight. The presence of *n* asymmetric carbon atoms means that it is possible to have 2ⁿ compounds with the same chemical structure but with different spatial configuration (stereoisomers). We have already mentioned that steroid compounds, with their close structural relationship, show considerable differences in physiological properties. Their physiological action is extremely sensitive to changes in chemical structure, and this may also be expected to apply for changes in their spatial structure. It has been found, however, that in nature, in both vegetable and animal organisms, mainly one stereochemical structure is used for the widely different tasks which the steroid compounds have to fulfil. Indeed, this preference for one particular spatial configuration is the rule rather than the exception in nature. It is a well-known fact, for example, that the amino acids, from which proteins are built up, occur exclusively with an L-configuration, and never with its mirror image, the D-configuration.

As indicated by the heading, this article is concerned with retro-steroids. These are steroids in which the natural spatial configuration has been artificially changed at the two carbon atoms 9 and 10 to that of the mirror image.

Chemistry of the retro-steroids

The retro-structure can be obtained by ultra-violet irradiation of "provitamin D", which consists of ster-

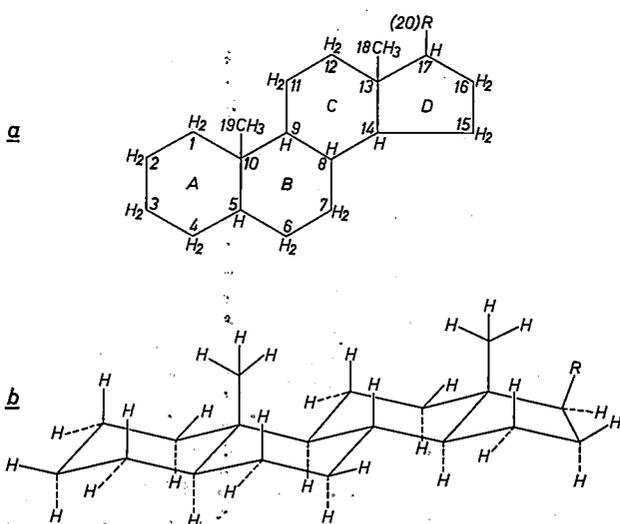


Fig. 1. a) The steroid skeleton. The carbon atoms are numbered and the rings are designated A, B, C and D. b) Spatial view of the skeleton. The solid and dashed lines indicate that the substituents are situated respectively above or below a plane passing approximately through the carbon atoms of a ring. R is a side-chain.

oids having two unsaturated bonds between carbon atoms 5 and 6 and between 7 and 8. A great deal of work has been done in this field at the Central Laboratory of Philips-Duphar, at Weesp, Netherlands [3], in co-operation with research workers of the Laboratory for Organic Chemistry, Leyden University [4]. The decision to start synthesis of retro-steroids was taken in 1956 by Reerink, under whose supervision the irradiation research at Weesp took place. By this time it was already possible to give some indications of the chemical reactions occurring during irradiation. These are shown in *fig. 3*.

As can be seen in the diagram, irradiation has the effect of breaking the bond between the carbon atoms 9 and 10. This bond can be reconstituted by further irradiation under the right conditions. In principle the mirror-image configurations at carbon atoms 9 and 10 can occur when this ring is formed again.

The different configurations are denoted by the letters α and β . In this notation the steroids are characterized by a $9\alpha,10\beta$ -configuration; and the retro-steroids by a $9\beta,10\alpha$ -configuration.

The starting material used in the investigations at Weesp was ergosterol (see *fig. 2*), a compound which is related to cholesterol and is easy to isolate from vegetable sources. The retro-form of this compound — at first just a by-product of the production of vitamin D₂ — was to play a leading part in the research at Weesp under the name of lumisterol₂.

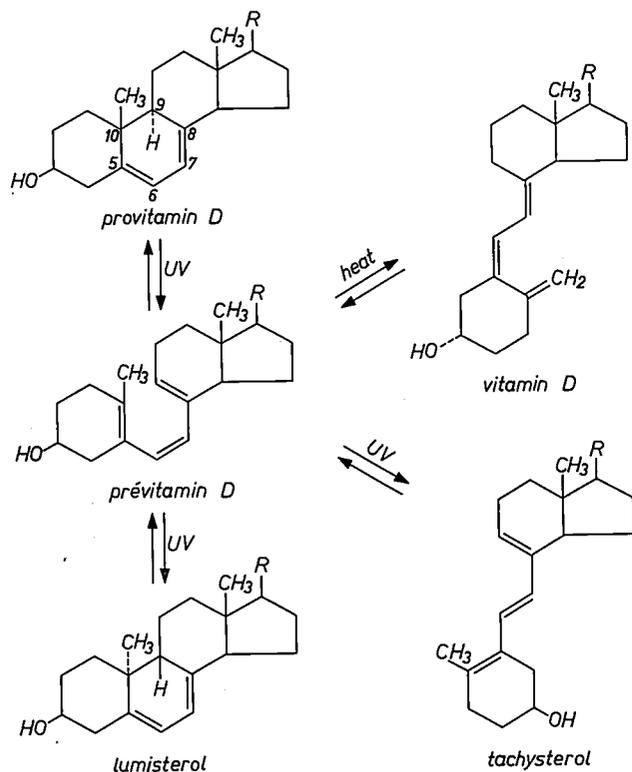


Fig. 3. Diagram showing the reactions that take place when a provitamin D is irradiated. R is an alkyl group.

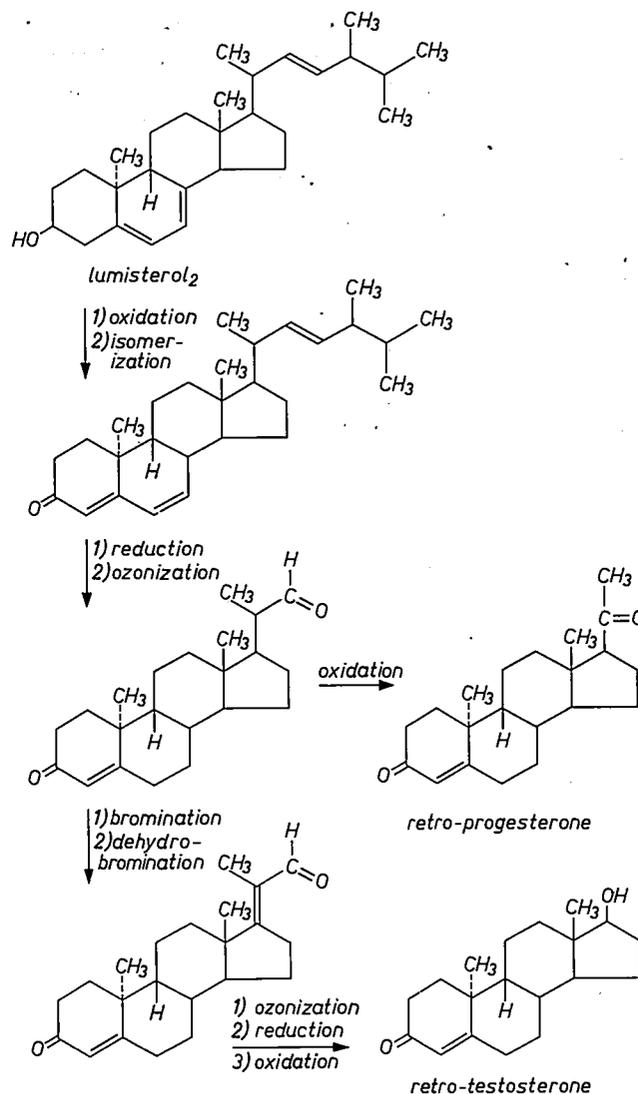


Fig. 4. Reaction diagram for the preparation of retro-progesterone and retro-testosterone from lumisterol₂.

The diagram of *fig. 4* shows the manner in which the retro-configurations of the sex hormones testosterone and progesterone can be synthesized from lumisterol₂. As well as this we have prepared many other compounds with interesting biological activity [5]. Three of these are selected for further discussion:

- 1) the 6-dehydro-retro-progesterone, or dydrogesterone, which is now marketed under the name of "Duphaston" [6];
- 2) the 16α -ethylether of dydrogesterone;
- 3) the 17α -(2'-methallyl)-retro-testosterone.

The preparation of these compounds is indicated in the diagram of *fig. 5*. A special method was used for the second compound. Not all carbon atoms of the steroids are equally accessible for chemical substitution, and indeed for the last ten years or so derivatives have been made with the aid of micro-organisms, which sometimes have a highly specific hydroxylating action. Research in this direction indicated that micro-

organisms can also convert retro-steroids [6]. A case in point was the successful use made of the mould *Sepedonium ampullosporum* for the conversion of dydrogesterone into its 16 α -hydroxy derivative, which was further converted into the corresponding 16 α -ethylether by means of the usual chemical methods.

To conclude this section we show in *fig. 6* the marked change that occurs in the spatial structure as a result of conversion into the retro-form [7].

Pharmacological action

Work on synthetic substitutes of the sex hormones generally has the following aims in view:

- to reinforce or weaken certain effects of natural hormone activity, with the aim of increasing its specificity;
- to obtain activity after oral administration — most natural sex hormones are inactive after oral administration.

In what follows we shall show some of the results we have been able to achieve with retro-steroids, taking the three compounds mentioned as examples. First, however, it is necessary to know a few elementary facts about the menstrual cycle in women and about the changes that take place at the beginning of pregnancy.

After puberty there is regular menstrual bleeding. This is caused by the shedding of the inner lining of the womb, the endometrium, which has been subjected to the action of the estrogens and progesterone pro-

duced in the ovaries. The estrogens are mainly produced in the Graafian follicle, the vesicle containing a ripening egg cell, or ovum. After the ovum has ripened sufficiently, the Graafian follicle bursts open (ovula-

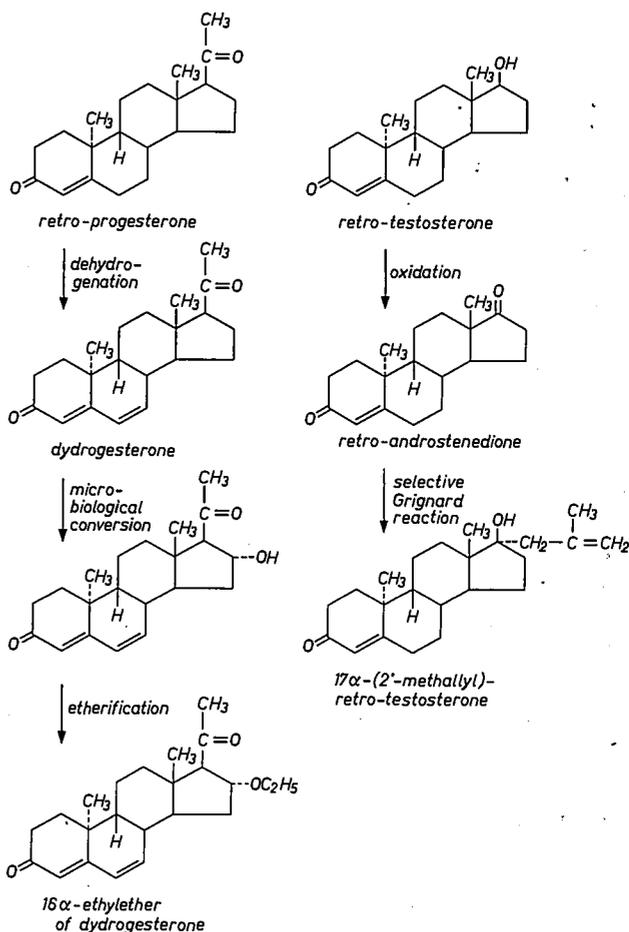


Fig. 5. Reaction diagram for the preparation of dydrogesterone, the 16 α -ethylether of dydrogesterone and 17 α -(2'-methyl)-retro-testosterone.

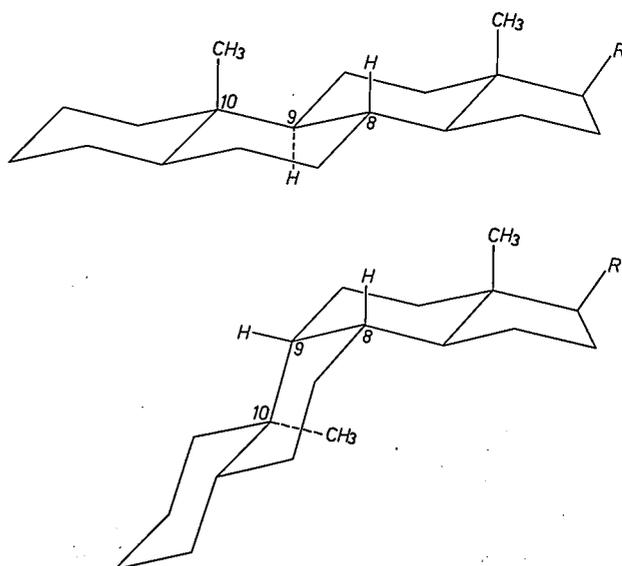


Fig. 6. The spatial structure of the normal steroid skeleton (above) and of the retro-steroid skeleton (below).

[*] Registered trade mark, N.V. Philips-Duphar.

[3] A. G. Boer, J. van Niekerk, E. H. Reerink and A. van Wijk, Proc. Kon. Ned. Akad. Wetensch. **39**, 622, 1936; U.S. patent 2 163 659 (1939) and 2 216 719 (1940); J. van der Vliet, Rec. Trav. chim. Pays-Bas **67**, 246 and 266, 1948.

[4] E. H. Reerink and A. van Wijk, Biochem. J. **23**, 1294, 1929; A. L. Koevoet, A. Verloop and E. Havinga, Rec. Trav. chim. Pays-Bas **74**, 788, 1955; E. Havinga, A. L. Koevoet and A. Verloop, Rec. Trav. chim. Pays-Bas **74**, 1230, 1955; P. Westerhof and J. A. Keveling Buisman, Rec. Trav. chim. Pays-Bas **75**, 1243, 1956; M. P. Rappoldt, P. Westerhof, K. H. Hanewald and J. A. Keveling Buisman, Rec. Trav. chim. Pays-Bas **77**, 241, 1958; M. P. Rappoldt, J. A. Keveling Buisman and E. Havinga, Rec. Trav. chim. Pays-Bas **77**, 327, 1958; M. P. Rappoldt and E. Havinga, Rec. Trav. chim. Pays-Bas **79**, 369, 1960; M. P. Rappoldt, Rec. Trav. chim. Pays-Bas **79**, 392 and 1012, 1960.

[5] P. Westerhof and E. H. Reerink, Rec. Trav. chim. Pays-Bas **79**, 771, 794 and 1118, 1960; P. Westerhof and A. Smit, Rec. Trav. chim. Pays-Bas **80**, 1048, 1961; A. Smit and P. Westerhof, Rec. Trav. chim. Pays-Bas **82**, 1107, 1963; P. Westerhof, Rec. Trav. chim. Pays-Bas **83**, 1070, 1964; R. van Moorselaar, S. J. Halkes and E. Havinga, Rec. Trav. chim. Pays-Bas **84**, 841, 1965; P. Westerhof, J. Hartog and S. J. Halkes, Rec. Trav. chim. Pays-Bas **84**, 864, 1965; S. J. Halkes and E. Havinga, Rec. Trav. chim. Pays-Bas **84**, 890, 1965; H. van Kamp and S. J. Halkes, Rec. Trav. chim. Pays-Bas **84**, 904, 1965; P. Westerhof and J. Hartog, Rec. Trav. chim. Pays-Bas **84**, 918, 1965.

[6] In co-operation with the Royal Netherlands Fermentation Industries, Ltd., Delft. See J. de Flines, D. van der Sijde and W. F. van der Waard, Rec. Trav. chim. Pays-Bas **85**, 701, 712 and 721, 1966.

[7] C. Romers, E. van Heykoop, B. Hesper and H. J. V. H. Geise, Rec. Trav. chim. Pays-Bas **84**, 885, 1965.

tion) and the cells in which the ovum was embedded develop into a gland with internal secretion, the corpus luteum, which, in addition to estrogens, now produces progesterone. The progesterone is thus produced only in the second half of the cycle, while the estrogens are active during the whole of the cycle (fig. 7). Since progesterone is formed and starts to work only after the ovum has been discharged from the ovary, it is

because estrogens and progesterone can in turn inhibit the production of FSH and LH.

If fertilization of the ovum does take place, a hormonal stimulus (HCG, see the pregnancy diagram in fig. 9) goes to the ovary from the fertilized ovum implanted in the endometrium. This ensures that the corpus luteum does not degenerate and that the production of progesterone is therefore maintained.

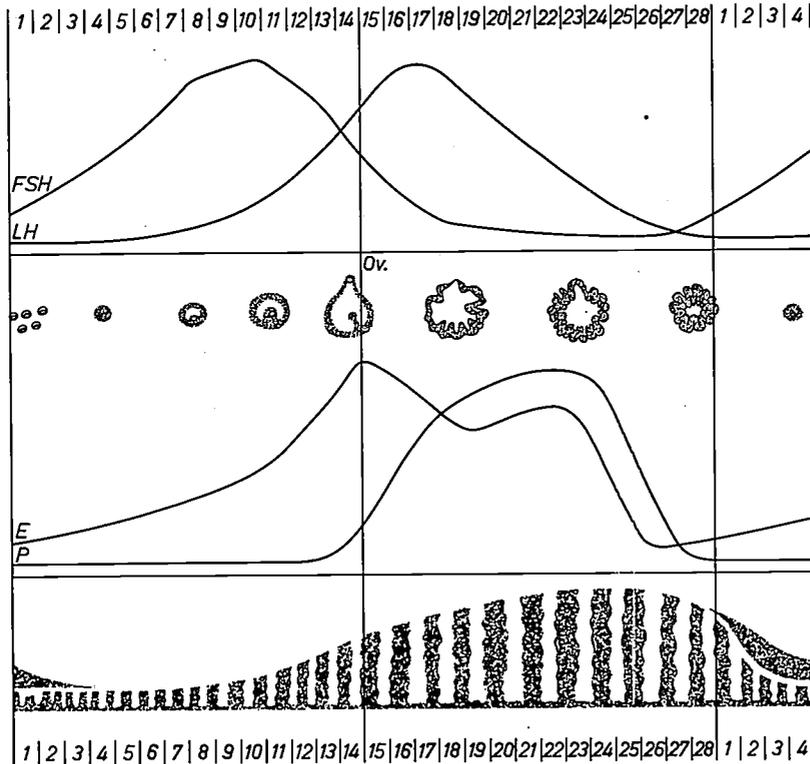


Fig. 7. Interaction of the gonadotropic hormones FSH and LH from the pituitary gland and the estrogens (E) and progesterone (P) from the ovary, and their effects on the ovary and the endometrium (the inner lining of the womb) during a normal menstrual cycle.

easy to understand that the principal task of this hormone is to ensure that the ovum, if it is fertilized, will be able to become implanted in the inner lining of the womb, which has been prepared for this by the progesterone (fig. 7 and fig. 8). If no fertilization takes place, the corpus luteum degenerates. The production of hormone ceases, and as a result of this the whole endometrium is shed; this manifests itself as menstrual bleeding. A renewed ripening of a Graafian follicle then starts and the process starts all over again. The whole process takes about 28 days.

The production of estrogens and also of progesterone during the cycle is stimulated by the hormones which are produced in the pituitary gland. Putting it very simply, we may say that the gonadotropic hormone FSH (follicle-stimulating hormone) stimulates the production of estrogen as well as the growth of the Graafian follicle, while the inducing of ovulation (bursting of the follicle) and the production of progesterone are stimulated by the gonadotropic hormone LH (luteinizing hormone). This is shown in the cyclic menstrual diagram of fig. 7. The cyclic behaviour results

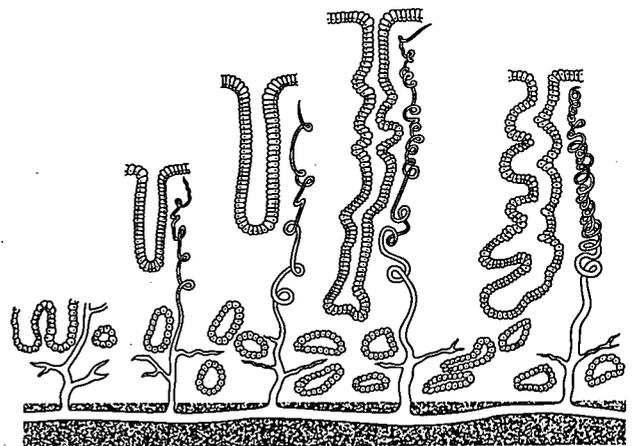


Fig. 8. Further details of the changes taking place in the endometrium during the menstrual cycle. The first phase in the build-up, called the proliferative phase, is controlled by the estrogens. Glands, blood vessels and other cell tissues are formed. In the second, or secretory phase, which is controlled by progesterone, the blood-vessels and glands take on a rather more convoluted appearance, and the glands also begin to secrete substances capable of nourishing a fertilized ovum. At the climax of this phase the endometrium is about 7mm thick. In the pre-menstrual phase the growth of the endometrium diminishes due to the reduced supply of blood, the termination of the glandular secretions and loss of water.

Furthermore, the fertilized ovum itself starts to make progressively more and more progesterone. As a result the endometrium is preserved and no menstruation occurs during pregnancy. The production of gonadotropic hormones in the pituitary is inhibited during pregnancy, so that there is no renewed ripening of follicles and therefore no further ovulation. All this can be considered as a protection for the embryo: if there were any further ovulations, the released ova could also be fertilized and would endanger the originally fertilized ovum.

progesterone in pregnancy is to be seen. This brief account mentions only a few of the many activities of progesterone.

Dydrogesterone ("Duphaston")

Cases occur in which the body produces insufficient progesterone or none at all, with the result that the life of the foetus is endangered and miscarriage may result. If the treatment requires the administration of a substitute for progesterone, one must be sure that the substitute chosen has all the actions of progesterone

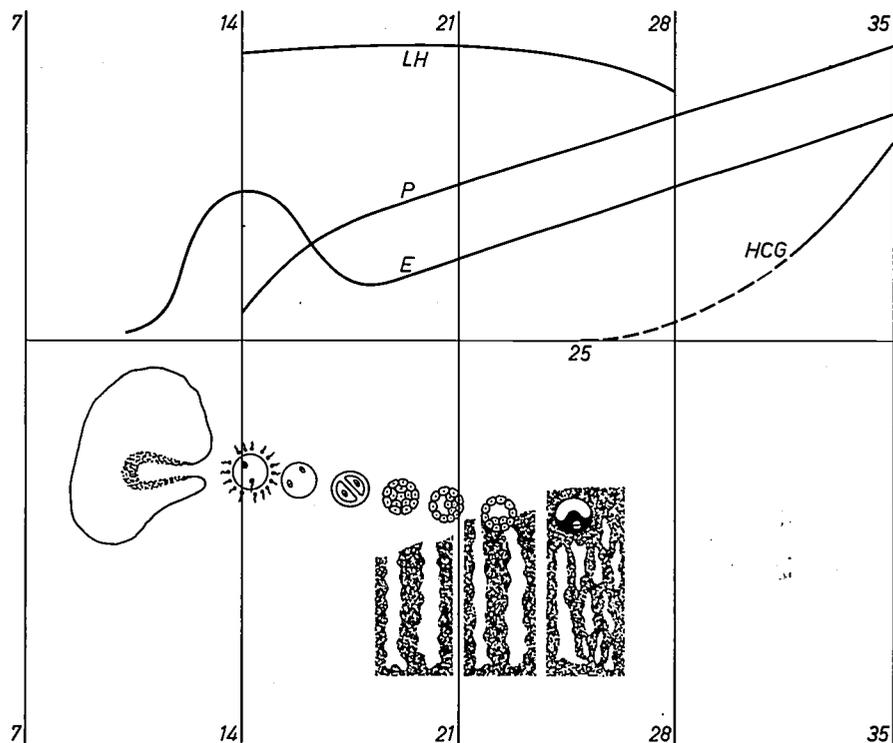


Fig. 9. Diagram showing the implantation of a fertilized ovum in the endometrium, and the accompanying changes in hormone concentration. The diagram shows a continuation of the production of estrogens and progesterone compared with the production in a normal menstrual cycle (fig. 7). From about the 25th day of the cycle a hormonal stimulus HCG goes out from the fertilized ovum. The dashed part of the curve HCG was obtained by extrapolation.

A second action which progesterone has is directly connected with this protective function. During ovulation the cervix uteri (the external orifice of the womb) is subjected to the influence of the estrogens alone. These make certain that the glands, present here in great abundance, give off a thin secretion which greatly facilitates the penetration of spermatozoa, so that fertilization is able to take place. The progesterone, which is produced after ovulation and is also present in high concentration during pregnancy, exerts a marked thickening action on the cervical mucus so that the passage of spermatozoa is made very difficult or even impossible. Here again, the protective action of

which we have mentioned. The developing embryo in the womb is extremely sensitive to all kinds of hormonal action, and therefore any substitute for progesterone must not possess any hormonal side-effects. It is not sufficient for a synthetic product to show progestational action in appropriate tests; it must also be thoroughly investigated to see if the substance has any undesirable side-effects on the foetus.

In the early stages of the search for progestational compounds that would be active after oral administration, work was largely concentrated on male hormone substances which have certain activities in common with progesterone. When this progestational action was in-

tensified by bringing about certain changes in structure the male action did not as a rule completely disappear.

In experiments on animals [8] the drawback of administering such substances during pregnancy was therefore particularly apparent with female embryos. These types of compound could not therefore be considered for use in maintaining a human pregnancy.

At the time when we started work on retro-progesterone there were very few, if any, progestational substances for oral administration that did not have one or more hormonal side-effects. When it had been found that oral administration of retro-progesterone gave progestational action and that 6-dehydro-retro-progesterone or dydrogesterone gave even greater action, these compounds were subjected to the appropriate experiments, in which all kinds of progestational action could be determined as well as other possible hormonal side-effects. Dydrogesterone was found to have no action not found in natural progesterone [9].

It does not however have *all* the actions of the natural hormone. For example, this retro-steroid does not have the thermogenetic effect of the natural hormone. If the temperature of an adult non-pregnant woman is recorded daily (before she gets up in the morning) it is found that in the second half of her cycle, i.e. after ovulation has taken place, her temperature is 0.5 °C to 1 °C higher than in the first half of her cycle. The body temperature is also slightly higher during pregnancy than it was in the period preceding the ovulation that led to fertilization. This makes it possible to find out, by taking the temperature of a patient being treated with dydrogesterone, whether the body has started to produce its own progesterone. The breakdown of this retro-steroid in the body also takes a different course from that of the natural hormone (see last section), so that the excretion products of the body's own progesterone can be determined in addition to those of dydrogesterone. This has distinct advantages for diagnosis and therapy. During treatment with dydrogesterone it is thus possible to find out whether the body's own production during pregnancy, for example, has become high enough after a certain period of time (since the implanted ovum is also going to make progesterone itself) for treatment with the pregnancy-maintaining drug to be stopped or adapted. We are in fact able to say that dydrogesterone is a substance with distinct advantages for use as a substitute for progesterone [10]. These advantages are:

- a) its great similarity to the natural hormone progesterone, as proved in progestational tests;
- b) its oral activity;
- c) the total absence of hormonal side-effects (no virilizing effect);
- d) It does not interfere with diagnosis.

One other property in which dydrogesterone differs from progesterone has not yet been mentioned. As we have said, the action of progesterone prevents ovulation from occurring after pregnancy has begun. Dydrogesterone does not have this ovulation-inhibiting action of progesterone. Even in very large doses the substance has no inhibiting effect on the mechanism causing ovulation in the normal cycle [10].

Because of this combination of properties, dydrogesterone in certain cases shows not only a pregnancy-maintaining effect, but can also promote pregnancy. Where infertility is due to the body's insufficient production of progesterone, for example, treatment with dydrogesterone can bring the endometrium to an optimum state without reducing the chance of ovulation, thus increasing the chances of pregnancy.

The 16 α -ethylether of dydrogesterone

We shall now deal with a second substance which has a progestational action, but which is specialized to a very much greater extent in its activity. This substance, 16 α -ethylether of dydrogesterone, exhibits only very slight progestational action on the endometrium and like dydrogesterone, it does not inhibit ovulation. The only indication of its progestational action is the marked thickening of the cervical mucus that occurs after administration. This thickened mucus acts, as we have seen, as a barrier to sperm cells.

If this substance is administered in the first half of the cycle, i.e. before ovulation, the cervical mucus becomes very viscous. Normally at this time the cervical mucus is only subject to the effects of estrogens and is a rather thin fluid. Although at this time the cervical mucus should promote fertilization by allowing the rapid and unobstructed passage of sperm cells, their passage is hampered as a result of the administration of this substance (and hence also of other progestationally active compounds). The substance might therefore act as an impediment to pregnancy, without bringing about other progestational effects. This brings us to the interesting question of whether this action could be taken as the basis of a contraceptive preparation.

Presumably, the principal action on which most such preparations depend is the one which inhibits the gonadotropic substances from the pituitary gland. The underlying mechanism is based on an imitation of the processes during pregnancy which prevent the occurrence of ovulation. Nevertheless, there are indications that a number of women using such contraceptive measures do still have ovulations, with no resulting pregnancy. This absence of pregnancy is probably due to other (local) effects of the contraceptive preparation, i.e. on the oviducts, the womb and the cervical mucus,

reducing the probability of fertilization or hindering the process of implantation.

In all probability, the proportion of estrogens to progesterone in the various preparations has an important bearing on their different kinds of action, but it would take us too far afield to go into this subject here.

The search for substances with a contraceptive action is still in progress, with particular emphasis on these local effects. The action which has been found for 16 α -ethylether of dydrogesterone is in this connection highly significant.

17 α -(2'-methyl)-retro-testosterone

Not every cycle is accompanied by an ovulation. In such cases the endometrium is therefore not affected by the progesterone. This may sometimes give rise to menstrual irregularities, resulting in fairly prolonged bleeding.

We have not yet touched on the fact that the action of progesterone must in general be preceded by the action of the estrogens. In a few cases high doses of progesterone itself can bring about an effect, but the addition of a small amount of estrogen increases the action of progesterone considerably (synergistic action).

If enough endometrium activated by estrogen is still present the progestational substance is generally able to stop excessive bleeding like that mentioned above, since progestational changes can still be brought about in the endometrium. The treatment artificially reproduces the pattern of an ovulatory cycle. If treatment with the progestational agent is then stopped, normal menstruation will usually follow.

In cases of very persistent bleeding, in which progestational agents do not help, we can still stop the bleeding by giving an anti-estrogenic substance.

The substance 17 α -(2'-methyl)-retro-testosterone is such a substance. Experiments specially designed to assess estrogenic action showed this substance to be active as an antagonist of estrogen. The substance was found to be able to antagonize both the body's own estrogens and those administered artificially, e.g. by injection. Persistent bleeding occurring after a non-ovulatory cycle is stopped by this substance within 48 hours.

To be assured of safe application of a retro-steroid as a pharmaceutical it is necessary to have a detailed knowledge of any changes it may undergo in the body. This is the subject of the next section.

Conversions of the retro-steroids in the body

In all the interactions of hormone regulation taking place in the body, it is obviously important that the steroid hormones should be inactivated and excreted

after performing their function. This inactivation takes place mainly in the liver: here the body has a series of enzyme systems at its command which attack the steroid molecules.

The first inactivating reactions consist of reductions of the 3-keto group and the double bond between C4 and C5. First, the double bond between C4 and C5 is reduced, which causes a new centre of asymmetry to appear at C5. There are individual enzymes which give rise either to the 5 α or the 5 β compounds (5 α and 5 β reductases). Reduction of the 3-keto group then follows. This is the normal way in which saturated 3 α -hydroxy-steroids are produced in the body. Apart from these initial reactions, various other conversions take place. If a keto group is present at C20, it is reduced; in a number of 17-OH compounds the whole side chain is split off, and sometimes hydroxyl groups are introduced at various places. The saturated compounds are coupled to glucuronic acid or sulphuric acid and this imparts a polar character to the whole compound which enables it to be rapidly excreted via the urine.

In accordance with the above, the urine of pregnant women is found to contain pregnane-3 α ,20 α -diol-glucuronide as the principal metabolite of progesterone (fig. 10).

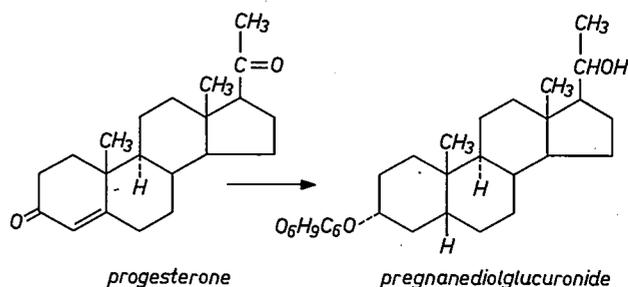


Fig. 10. Conversion of progesterone into a product which is excreted in the urine.

It appeared interesting to investigate the behaviour of the retro-steroids towards the enzymes that attack the natural steroids.

First of all, the behaviour of the 5 α and 5 β reductases, obtained from the liver of rats, was investigated. These occur at different places in the liver cells and can be obtained separately. The naturally-occurring steroids are quickly reduced by these enzymes. Synthetic derivatives from these natural steroids may sometimes be attacked with a little more difficulty; thus, for example a double bond between C6 and C7 is found to make

[8] H. F. L. Schöler and A. N. de Wachter, *Acta endocrinol.* 38, 128, 1961.

[9] H. F. L. Schöler, *Acta endocrinol.* 35, 188, 1960.

[10] P. M. F. Bishop, U. Borell, E. Diczfalusy and K. G. Tillinger, *Acta endocrinol.* 40, 203, 1962.

the reaction considerably slower. The behaviour of the above enzymes towards retro-steroids of widely varying kinds can be summed up quite simply: none of the retro-steroids is reduced [11]. This also applies in particular to the substances mentioned above, dydrogesterone, the 16 α -ethylether of dydrogesterone and 17 α -(2'-methallyl)-retro-testosterone. The chief initial attack which the normal steroids are subjected to in the body is apparently unable to take place with the retro-steroids.

As previously mentioned, the 20-keto group in progesterone is also reduced. The relevant enzyme, the 20-keto reductase, was now also investigated to obtain some knowledge of its behaviour towards retro-progesterone and dydrogesterone. In contrast to what was found with the 5-reductases, 20-keto reductase was in fact able to reduce the above retro-steroids. With the enzyme from the rat the reduction was found to proceed at the same rate as with progesterone; with an enzyme preparation of human origin the reduction of the retro-compounds proceeded rather more slowly than that of progesterone.

The enzyme that reduces the 20-keto group attacks a group located in the side chain of ring D of the steroid molecule, a location at quite a considerable distance from those locations that give the retro-configuration its specific character. Perhaps this explains why the 20-keto reductase does attack the retro-steroids while, as we have seen, the 5-reductase does not.

The above results, which were obtained with isolated enzymes *in vitro*, are in complete agreement with what is known up to now about the breakdown of retro-steroids in the body. If radioactively labelled dydrogesterone is taken by women, the principal metabolite appearing in the urine [12] is found to be the glucuronide of the compound which has been reduced at the C20 site (fig. 11). In contrast to what happens with progesterone, we find no reduction of the A-ring here, but only a reduction of the keto group at C20. There is thus quite a difference, compared with the metabolism of progesterone.

The reactions which dydrogesterone undergoes are thus more limited than those of, say, progesterone, but

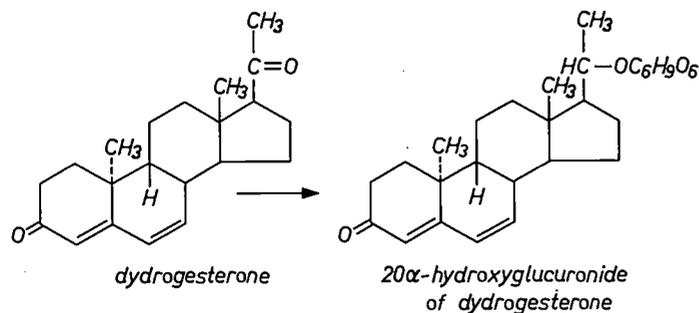


Fig. 11. Conversion of dydrogesterone into an excretion product.

in fact these reactions proceed so rapidly that if dydrogesterone is given to women orally, about 40% of the dose is excreted in the urine within eight hours.

Apart from the above reactions which lead to inactivation and rapid excretion, we are naturally also concerned with the many reactions in the body by which hormones are synthesized and converted one into the other. Let us consider for example the estrogenic steroids, which are characterized by an aromatic A-ring. Living tissues form these substances by aromatization reactions from testosterone and androsterone, for instance as shown by the diagram in fig. 12.

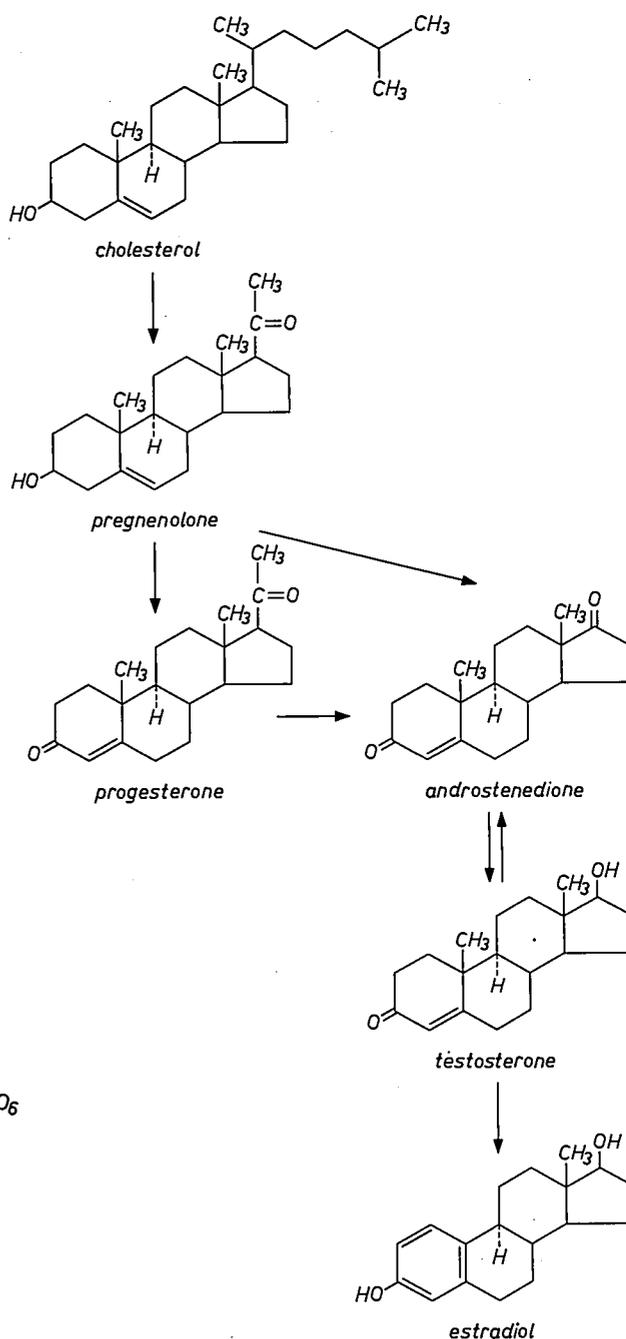


Fig. 12. Simplified diagram of the synthesis in the body of the female sex hormone estradiol.

These estrogens can also be made *in vitro* by using certain systems of enzymes occurring in the "microsome fractions" of the cells of various organs and allowing these enzymes to react with, for example, testosterone. An enzymatic system from human placenta, which has this property, was also investigated to see if it would aromatize steroids having the retro-structure. Here also it was found that the retro-steroids in question were not attacked by the particular enzyme system [13]. Aromatization of retro-testosterone, for example, could not be demonstrated, while on the other hand these steroids also exhibited no trace of estrogenic action in animal experiments after reaction with the enzyme. These experiments give an indication that when these steroids are administered to humans there will be no estrogenic side-reactions due to aromatization. This absence of estrogenic side-effects is again in agreement with what is known about the biological properties of dydrogesterone [14].

Owing to the dissimilar configuration of the retro-steroids it may of course happen that the molecules do not "fit" so well at the receptor sites of those organs through which the compounds are active. But once mol-

ecules are obtained with a sufficiently good "fit" to give biological activity, and dydrogesterone serves here as an example, we then have gained the advantage that the compound is less easily converted by the body into related steroids. As we have seen in the example of dydrogesterone the reactions which give rise to excretion products take place sufficiently quickly, but they are limited in number when compared with the natural steroids and their derivatives. With the retro-steroids there is therefore little fear of side-effects due to conversion into other hormonally-active steroids, so that in general an action exclusively in accordance with the purpose of the treatment is to be ensured.

Summary. The sex hormones which are produced by the ovaries and testes: estrogens, progesterone and testosterone etc., belong to the steroids. Retro-steroids are steroids with a $9\beta,10\alpha$ configuration, which can be obtained by the irradiation of provitamin D. The pharmacological properties and metabolism of three of these compounds, which differ from normal steroid hormones in their spatial structure, are examined here. Dydrogesterone ("Duphaston") can be taken orally as a substitute for progesterone, for example where there is a risk of miscarriage. It serves not only to maintain but also to promote pregnancy. The 16 α -ethylether of dydrogesterone is characterized by a specific action on the cervical mucus. The compound 17 α -(2'-methallyl)-retro-testosterone is characterized by an anti-estrogenic action which may explain why it can be used to stop persistent non-ovulatory menstrual bleeding. The conversions of the retro-steroids in the body take place differently from those of natural steroids. There is for example no aromatization of the A-ring, and therefore little fear of conversion into other hormonally-active steroids which could lead to side effects.

[11] J. N. Walop and N. de Lange, to be published.

[12] E. Diczfalusy, K. G. Tillinger, R. J. E. Esser and A. C. Houtman, *Nature* 200, 79, 1963.

[13] J. N. Walop and N. de Lange, *Biochim. Biophys. Acta* 130, 249, 1966.

[14] M. Marois, *Bull. Acad. Nat. Médecine* 146, 354, 1962.

In the Cyclotron and Isotope Laboratories



The Cyclotron and Isotope Laboratories have recently been set up by N.V. Philips-Duphar in the grounds of the Netherlands Reactor Centre at Petten. The chemical processing of the radioactive material produced in the laboratories is carried out in cabinets termed "hot cells". The maintenance gangway shown gives access to the back of these hot cells. When maintenance or repair work is required the lead shields of the hot cells can be moved aside or removed. This is of course only done after removing as much of the radioactive material as possible from the

hot cell. Access to the gangway is restricted to specially trained and equipped personnel, under the supervision of the laboratory medical department. The man working on the right is wearing an oxygen mask; the man on the left is wearing a pressure suit supplied with air by the hose lying on the floor.

The pressure in the gangway is kept below that in the hot cells to prevent radioactivity from being carried to the workroom in front of the cells by particles of dust, etc. The pipelines are for ventilating the cells.

Preparation of virus vaccines by means of tissue cultures

O. Bosgra and J. H. G. Roerink

Introduction

Within certain limits the body of man and animal is able to offer resistance to the invasion of pathogenic — i.e. disease-producing — organisms such as bacteria and viruses. Depending on the virulence and number of these organisms, however, a percentage of the infected individuals will become diseased, or may even die. In the individuals that survive the disease, antibodies are produced which circulate in the blood serum. These antibodies give the individuals immunity against the disease for a certain length of time, sometimes even for a lifetime. The immunity conferred is specific, that is to say, protection is provided only against a new infection by the same pathogenic organism.

The aim of vaccine inoculation is to give the individual a good start, as it were, by stimulating the body to produce antibodies before infection. Any ill effects which may result through such action should of course be very much less than those of the illness to be prevented.

Vaccine inoculation can be performed with "killed" or living organisms. In the latter case innocuous non-virulent strains of the pathogenic organisms are used.

The method of inoculation with vaccine is largely due to a discovery made by the English physician Jenner in 1796. Jenner found that people could be immunized against smallpox by infecting them with the cowpox virus, which is innocuous to human beings. The name vaccine, from the Latin "vacca" for cow, recalls this discovery.

Vaccine inoculation is a prophylactic measure, applied exclusively with the aim of preventing an illness. Other possibilities are also offered by the use of serum, which again is based on the effect of antibodies. A serum is prepared from the blood of animals that have recovered from a particular infectious disease or which have been inoculated several times with vaccine. Serum therefore, unlike vaccines, contains the appropriate antibodies in a high concentration, and it has an immediate effect. A serum may be administered as a preventive measure, to produce short-lasting immunity, or as a curative measure once the disease has appeared. Vaccines are solely preventive in their effect, but provided they are administered in good time they give more prolonged immunity against infection than serums.

The value of vaccine inoculation may be questioned

in view of the outstandingly successful results achieved with modern chemotherapeutics (sulphonamides for example), with antibiotics and with serums. One should bear in mind here that the diseases which have lost their grip on mankind as a result of the use of therapeutic agents are nearly all infectious *bacterial* diseases and not virus diseases.

A virus is a cell parasite. A virus that invades a cell forces the cell to put its metabolism to the service of the multiplication of the virus, which ultimately destroys the cell. The symptoms of disease observed after a virus infection are attributable *not* to the large amount of virus produced but to the resultant destruction of large groups of cells, and to the consequent malfunctioning of certain organs affected. But this means that we do not notice a virus disease — and cannot do anything to stop it — until the damage has already made a certain progress. Chemotherapeutics, antibiotics and serums, which have proved so successful in the control of infectious bacterial diseases, are also thrown into the fight against virus diseases but with little success. Viruses are as a rule not susceptible to the common antibiotics and chemotherapeutics and in many cases, because they multiply in the cells and not outside them as is usual with bacteria, viruses are not so readily accessible to these therapeutic agents. Taking this into account, we can understand how valuable vaccines are as a means of preventing virus diseases.

Cultivation of viruses in tissue cultures

In order to prepare virus vaccines, which can be either killed vaccines or living non-virulent vaccines, the viruses have to be cultivated in living matter. A particularly suitable medium for this was found to be incubated (embryonated) chickens' eggs [1]. Many vaccines in use today are made with the aid of chickens' eggs.

A method now rapidly gaining ground is the cultivation of viruses in tissue cultures. The tissues may be taken from any organ of mammals or birds, or of cold-blooded animals, or even insects. They may be tissues from kidneys, testicles, or muscles, etc.

There is a strong reason for making use of tissue cultures, and this is the much greater variety of viruses

Dr. O. Bosgra and Dr. J. H. G. Roerink are with N.V. Philips-Duphar, Research Laboratories, Weesp, Netherlands.

[1] See A. J. Klein and E. Hertzberger, *Philips tech. Rev.* 12, 273, 1950/51.

that can be cultivated in this way. We may mention the example of the polio virus, which cannot be cultivated in chickens' eggs but *can* be cultivated in tissue cultures from a monkey's kidney, a fact to which we owe the possibility of the polio vaccine. Other vaccines which have been prepared by means of tissue cultures are vaccines for measles, foot and mouth disease, canine distemper, infectious canine hepatitis, rinderpest and infectious rhinotracheitis in cattle.

The cell cultures in which viruses are to be cultivated require nutrition. The nutrient medium has to meet very rigorous requirements, particularly in regard to pH, the constituent salts, and amino-acid content. Widely employed nutrient media are those known as "Eagle's solution" and "Morgan's medium" [2].

Given appropriate ingredients of the nutrient medium the cells can be kept alive for a long time, and the metabolism of the cells is kept going in the normal way.

The most frequently used method of tissue culture consists in the formation of *primary cell cultures*. A particular organ, for example a kidney, preferably young and still growing, is cut into small pieces and treated with a proteolytic enzyme, e.g. trypsin. This treatment has the effect of attacking the structure of the tissue and of detaching cells which, although removed from their structure, remain alive. These cells are transferred to flasks, tubes or other vessels containing a nutrient medium. The conditions are arranged so as to promote the growth and multiplication of the cells, which soon results in the formation of a compact monolayer (one cell thick) at the bottom of the vessel (*fig. 1*). This is the primary cell culture. After

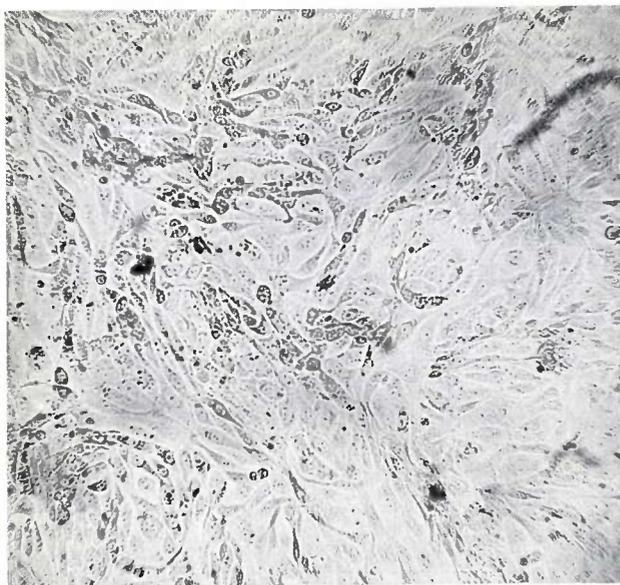


Fig. 1. Photomicrograph of a tissue culture of a dog's kidney. The culture is a monolayer, i.e. one cell thick. Magnification 100 \times .

further treatment with trypsin and transfer of the detached cells to a fresh nutrient medium, a secondary culture is obtained, and this procedure can be repeated. If the cells are required to be highly uniform this method of producing a *continuous line of cells* is the most suitable. Various such cell lines have been described in the literature, the best known being the Hela cell line which, cultivated from a cancerous tumour of a human cervix uteri has now been kept alive for more than ten years [3].

Tissue culture as an aid in diagnosis

Tissue cultures lend themselves particularly well to the demonstration of viruses. The changes produced in a monolayer by infection with a virus can easily be followed under a microscope. The abnormal picture observed some time later, called the cytopathogenic effect and abbreviated to CPE, has a pattern which is usually characteristic of the virus (*fig. 2*). Most viruses can be differentiated from each other by means of these patterns.

Tissue culture has thus become a valuable aid to diagnosis. The method has in addition led to the discovery of pathogenic viruses which could not have been found by other methods, one example being the virus responsible for infectious rhinotracheitis in cattle. Until about 1956 the true nature of this disease was unknown. The suspicion that it was due to a virus could neither be proved by experiments on animals, nor by cultivation in chickens' eggs. A tissue culture made from the kidney of an unborn calf finally brought the virus to light. Soon after this discovery it was possible to develop a vaccine against this disease and put it on the market.

Once the pathogenic virus is identified the vaccine can either be prepared as killed or as living vaccines. We shall not deal here with the preparation of killed vaccines, but simply mention that a suitable method of cultivation has to be found for the growth of the viruses. While it was formerly necessary to rely on chickens' eggs for this purpose, nowadays there is, as we have seen, a greater choice of tissues.

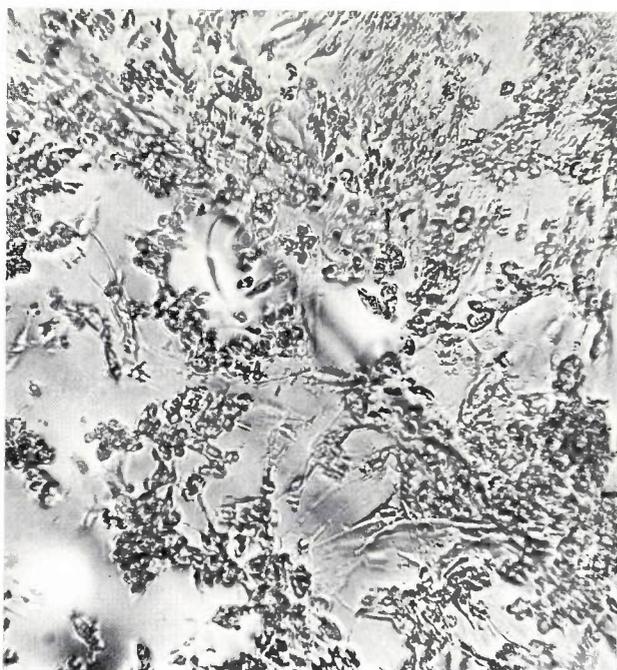
For the preparation of living non-virulent vaccines the possibility of tissue culture is of very much greater significance. The procedure for finding an innocuous variant of a particular virulent virus is to cultivate various generations of the virus in a host which is *foreign* to the type of virus concerned. The fact that this usually leads to the desired result may be understood in the following way. In addition to the virulent and rapidly multiplying virus, less virulent variants occur in a strain of viruses, but their existence is entirely overshadowed by the virulent species. This situation may be changed when the virus is transferred to another host:

one or more of the less virulent variants find it easier to adapt themselves to their new environment, and in their turn become virulent to the new host, and become dominant. By repeating this procedure a number of times — that is to say repeatedly transferring a small proportion of the changed virus population to the foreign environment — we are finally left with a population from which the virus which was virulent for the original host has been almost completely removed.

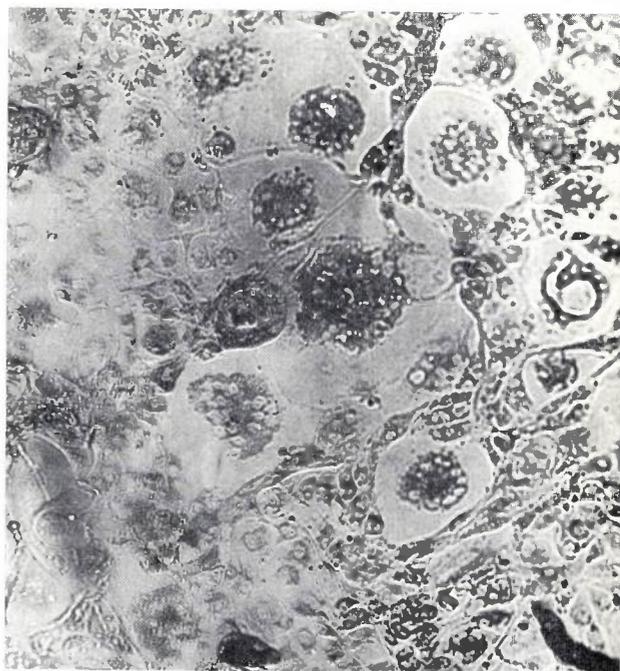
The method described will obviously not succeed in all hosts. When chickens' eggs were the only medium

properties. We shall not go into this subject here, but mention merely that this possibility is bound up with the fact that the nucleus of the virus governs the degree of virulence and other genetic characteristics of the virus, and that the protein coat of the virus is involved in the formation of the antibodies (likewise proteinous) in the host and thus determines the immunizing properties.

It will now have become clear that in developing a living non-virulent vaccine it is always necessary to establish what measure of immunization can be achieved.



a



b

Fig. 2. Monolayers of a dog's kidney showing cytopathogenic effect (CPE) caused by *a*) infectious canine hepatitis virus, *b*) canine distemper virus. This effect is generally so characteristic of the virus producing it that it can be used to identify the virus. Magnification 100 \times .

used for cultivation it was not always possible to develop living non-virulent vaccines. Tissue culture, with its wide variety of hosts, offers much more scope in this respect.

For the development of living non-virulent vaccines not only must the variant produced be non-virulent, it must also give the same or almost the same immunity to the original host as that which results when an individual survives a virulent infection.

It might seem surprising that it should be possible at all to find a virus possessing such a combination of

The immunity conferred by a particular vaccine can be determined by means of a serum neutralization test. In this test, blood is taken from an animal inoculated with a given vaccine and some test tubes are filled with various dilutions of serum prepared from the blood, for example, 1 : 10, 1 : 100 and so on. To each dilution a certain quantity of the appropriate virus is

[2] J. F. Morgan, H. J. Morton and R. C. Parker, *Proc. Soc. exper. Biol. and Med.* **73**, 1, 1950.

[3] G. O. Gey, W. D. Coffman and M. T. Kubicek, *Cancer Research* **12**, 264, 1952.

added, and a test is made by introducing a little of each dilution on a monolayer to determine the highest dilution at which the CPE is not found. The dilution factor of this dilution is then a measure of the immunity against the added virus.

Tissue culture versus culture in chickens' eggs

We now have a broad picture of the principal advantages of tissue culture as compared with cultivation in chickens' eggs.

- 1) A greater variety of viruses can be cultivated.
- 2) The measure of immunity obtainable with vaccines can easily be established from the CPE. The CPE is also a simple means of identifying viruses.
- 3) Vaccines prepared in tissue cultures contain virtually no protein. This is important in as much as precautionary measures are required if vaccines

cultivated in chickens' eggs are used on subjects allergic to egg protein.

- 4) The production of viruses presents relatively few problems. The multiplied viruses finally burst out of the cells of the tissue cultures and enter the nutrient medium. Purification is usually unnecessary, because of the absence of protein. It is only sometimes necessary to remove cell remains by filtering and centrifuging.

Summary. Prophylaxis by means of vaccines continues to be of very great importance today, particularly as a means of preventing virus diseases. This article discusses the method of preparing virus vaccines in tissue cultures, a method which has gained ground in recent years over the method of cultivating viruses in embryonated chickens' eggs. The new method has made it possible to develop various new vaccines against a variety of diseases, including polio and measles, and a number of diseases in dogs and cattle.



Vitamins in live-stock feeding

Th. J. de Man

Vitamins

If the figures given for the composition of feedstuffs in the tables used for live-stock feeds are examined, it can be seen that the routine method on which they are based (Henneberg and Stohmann's "Weender method", now a century old) amounts to a division into moisture, ash components, protein materials, fatty materials, crude fibre, and starchy materials.

This "Weender method" can be considered as a reflection of the "classical" nutrition theory. This maintained that, besides proteins, fats, and carbohydrates, only a limited number of minerals were necessary for the diet (potassium, sodium, calcium, magnesium, iron, chlorine, phosphorus, and sulphur). This classical nutrition theory, in general terms at least, held sway for

many years, and the peak of its ascendancy coincided with that of the energy approach to the problem. In those days this took the form of a comparison between the animal organism and a machine. The very limited value of such a comparison should however be recognized quite clearly: it is in fact very difficult to imagine a machine which for its fuel will not only manage with logs, peat, anthracite, coke, petrol, or crude oil, but which can also convert these kinds of fuel into structural material without stopping production, and which can if required use up this structural material later as a fuel. In "engineering" terms, this is what the animal organism can do.

Different approaches have led to the conviction that other quite different substances, as well as proteins, fats, carbohydrates and a few minerals, are indis-

Dr. Th. J. de Man is with N.V. Philips-Duphar, Amsterdam.

pensable in nutrition. The most direct approach makes use of animal experiments which have shown that if the feed consisted only of the above components in pure form, it was not sufficient to keep the test animals in optimum condition. The origin of the other approach cannot be given exactly but in any case it goes back a number of centuries. One could say that this approach is indicated by points at which an ever clearer understanding was reached of the fact that a number of diseases were the result of deficiencies in the diet. The diseases of greatest importance in this development were scurvy, beriberi, rickets, and pellagra. The great merit of Funk's work (and the chief reason why his famous publication [1] of 1912 attracted so much attention) was that he really recognized the above diseases as deficiency diseases, that is to say diseases resulting from deficiencies in antiscorbutic vitamin, anti-beriberi vitamin, antirickets vitamin, and antipellagra vitamin respectively. Even at that time he envisaged the possibility that these four vitamins might well be quite different compounds.

It is also of some interest to note that as early as 1901 Grijns had a clear understanding that beriberi was a deficiency disease. He was the first to suggest that beriberi occurred as a result of the deficiency of some substances (or one particular substance) in polished rice and that the necessary agents were present in the husks. In 1912 Funk thought (wrongly) that he had obtained the pure anti-beriberi vitamin. Success in this direction was not achieved until 1926 by Jansen and Donath, who in that year discovered how to isolate the pure substance from the husks of rice.

It was pure coincidence, though perhaps it is not without symbolic significance, that the isolation of the first vitamin (vitamin B₁) and the isolation of the first enzyme (urease) took place in the same year, 1926. Quite early on a connection had been intuitively established between the activity of vitamins and that of enzymes. The fact that vitamins exert their activities in very small amounts also indicated fairly clearly that it was worth investigating the possibility of a catalytic function. Extended investigations have now established that most vitamins are essentially components (co-enzymes) of the various fundamental enzyme systems which regulate the metabolism of the living cell.

The isolation of the anti-beriberi vitamin B₁ (aneurin, thiamine) in the pure state opened up a long road of successive isolations and purifications. Starting from classical nutrition theory some fifty compounds necessary in the diet were recognized, the nutrients. A good number of the nutrients are vitamins, and there are also essential amino acids and trace elements amongst them.

The more important vitamins, differentiated into

Table I. The most important vitamins.

Fat-soluble	Water-soluble
Vitamin A (axerophthol)	Vitamin B ₁ (thiamine, aneurin, anti-beriberi vitamin)
Vitamin D (antirickets vitamin)	Vitamin B ₂ (riboflavin, lactoflavin)
Vitamin E (tocopherols)	Nicotinic acid (P.-P. factor, anti-pellagra vitamin)
Vitamin K (antihaemorrhagic vitamin)	Pantothenic acid (vitamin B ₃)
	Vitamin B ₆ (pyridoxine, adermin)
	Biotin (vitamin H)
	Para-aminobenzoic acid (P.A.B.A.)
	Inositol
	Folic acid
	Choline
	Vitamin B ₁₂ (cyanocobalamine)
	Vitamin C (ascorbic acid, anti-scorbutic vitamin)

fat-soluble and water-soluble vitamins, are mentioned in *Table I*. Except for vitamin C all the above water-soluble vitamins belong to what is named the vitamin B complex.

In a relatively short time it also proved possible to elucidate the chemical structure of most of the vitamins. In a few cases, to the surprise of the research workers, the vitamin was found to be identical with a chemical compound already known. Such was the case for example with the antipellagra vitamin, which turned out to be identical with a well-known substance, nicotinic acid (nicotinamide).

Generally, after establishing the structure attempts to synthesize the vitamin artificially followed, preferably on an industrial scale in order to have a sufficiently cheap manufacturing method to make economically justifiable application possible. *Table II* sets out the most important vitamins, together with the year in which such synthesis became possible.

Table II. The vitamins and the year of their synthesis.

Vitamin D ₂	1931	Vitamin K ₁	1939
Vitamin C	1933	Vitamin K ₃	1939 [***]
Vitamin B ₂	1935	Pantothenic acid	1940
Vitamin B ₁	1936	Inositol	1940 [*]
Nicotinic acid	1937[*]	p-aminobenzoic acid	1941 [*]
Vitamin D ₃	1937	Choline	1941 [*]
Vitamin E	1938	Biotin	1944
Vitamin B ₆	1939	Folic acid	1946
		Vitamin A	1947

[*] Recognition as a vitamin.

[**] Confirmation of antihaemorrhagic activity.

Philips-Duphar in particular played a leading role in the industrial production of vitamin D [2] and of vitamin A [3]. This firm is now the biggest producer of vitamin D in the world.

Vitamins are nowadays added on a large scale to live-stock feeds. To carry out research in this field an experimental farm (*fig. 1*) has been set up in the neighbourhood of the Philips-Duphar laboratories. At this

literature are only rarely encountered in their pure form in animal husbandry as they only occur if the shortages are already relatively serious and particularly if they apply to just one particular vitamin. With very small deficiencies (in practice generally not of single vitamins) man finds himself in the border region between illness and radiant health, a region which is bigger than generally realized. In our farm animals, with minor shortages (hypovitaminoses) it is often not



Fig. 1. At the Philips-Duphar experimental farm: sow and litter in farrowing house.

farm feeding experiments are carried out with poultry, pigs, cattle and sheep, under conditions as which resemble the natural conditions as much as possible like those encountered in practice. These experiments are carried out with nutrients such as vitamins, antibiotics and enzymes, which are added in small amounts to the feeds to give optimum and economic results.

The recommended amounts of vitamins in live-stock feed

The deficiency symptoms clinically recognized as deficiencies of one or more vitamins and described in the

possible to say much more than that the animals "are not doing so well". Economically this makes itself felt in unsatisfactory breeding results, poor growth, unsatisfactory feed conversion, low production and

[1] J. Funk, *State Med.* 20, 341, 1912.

[2] There have been numerous publications on this since 1929 by E. H. Reerink, A. van Wijk, J. Boer, J. van der Vliet, and others, in *Rec. Trav. chim. Pays-Bas*, and also various dissertations written under the guidance of Prof. Dr. E. Havinga (Laboratory for Organic Chemistry, Leyden University, Netherlands).

[3] See the publications by H. O. Huisman and colleagues in *Rec. Trav. chim. Pays-Bas*.

Table IV. Vitamin content of live-stock feeds. The indication 0 may also mean that the amount is so small that in practice it can be neglected; the symbol + indicates a non-negligible amount of the vitamin concerned, but the information available does not justify quantitative indication.

	Vit. A	β -carotene	Vit. D	Vit. E	Vit. B ₁	Vit. B ₂	Nicotinic acid	Pantothenic acid	Vit. B ₆	Choline	Vit. B ₁₂	Biotin	Folic acid	Vit. C
	I.U.	mg	I.U.	I.U.	mg	mg	mg	mg	mg	mg	μ g	mg	mg	mg
Buckwheat	0	0	0	+	+	1.0	40	10	+	450	0			0
Barley	0	0	0	6	4	1.2	30	7	4	1000	0	0.1	0.6	0
Oats	0	0	0	5	6	1.1	12	10	3	900	0	0.2	0.2	0
Yellow maize	0	2	0	5	4	1.0	12	6	4	450	0	0.1	0.2	0
Maize gluten meal	0	6	0	+	0.2	1.5	45	9	7	300	0	0.1	0.2	0
Maize gluten feed	0	3	0	+	2	2.0	60	14	12	1300	0	0.2	0.2	0
Maize feed meal	0	1	0	+	8	2.0	40	8	10	900	0	0.2	0.3	0
Millet	0	0	0	+	1.5	1.0	25	10	3	450	0	0.1	0.15	0
Milicorn	0	0	0	6	4	1.0	30	10	4	500	0	0.2	0.2	0
Rye	0	0	0	10	3	1.2	12	7	3	450	0	0.05	0.6	0
Crude rice	0	0	0	6	3	1.0	30	8	4	900	0	0.1	0.4	0
Wheat	0	0	0	10	4	1.0	50	10	4	900	0	0.1	0.4	0
Wheat germ meal	0	0	0	100	18	5	25-50	12	12	3000	0		1.5	0
Wheat bran	0	0	0	15	8	3	50-100	20	12	1200	0	0.1	1.5	0
Ground-nut-oil meal	0	0	0	+	7	4	150	50	6	1750	0	0.3	0.6	0
Cottonseed-oil meal	0	0	0	+	4	4	30	10	5	2750	0	0.6	1	0
Coconut-oil meal	0	0	0	+	1	3	30	6	3	1100	0		1	0
Linseed-oil meal	0	0	0	+	7	3	35	7	6	1400	0		3	0
Sesame-oil meal	0	0	0	+	3	3	30	6	6	1500	0	0.3	1	0
Soya-oil meal	0	0	0	+	4	4	30	14	6	2750	0	0.2	0.6	0
Meat-and-bone meal	0	0	0	0	0.5	3	40	2	+	1500	10-50			0
Fish meal	+	0	+	0	0.5	6	60	6	3	1500	25-100	0.1	0.1	0
Fish solubles (condensed)	0	0	0	0	5	12	150	35	10	2500	50-200	0.1	0.2	0
Milk (whole)	200-1000	0.1-0.3	3-25	0.5	0.4	1.5	1	3	1	100	4	0.03	0.1	18
Skimmed milk powder	0	0	0	0	4	18	10	30	5	1000	30	0.2	0.6	+
Dried whey	0	0	0	0	4	25	10	40	5	1500	15	0.2	0.8	+
Potatoes, fresh	0	0	0	1	1	0.6	10	2	1		0	0.1	0.05	50-200
Brewers yeast (dried)	0	0	0	0	25-100	25-75	\pm 400	50-100	25-50	3000	0	1.5	9	0
Grass, fresh	0	15-75	0	\pm 20	1	2	7	2	2	200	0	0.05	1	500-1000
Grass meal	0	50-100	0	\pm 100	3	10	30	15	9	1250	0	0.2	5	+
Lucerne meal	0	50-100	0	\pm 100	3	10	30	20	9	1250	0	0.2	5	+
Molasses	0	0	0	0	0	1	40	2	2	600	0			0
Carrots, fresh	0	60-100	0	5	0.5	0.4	2	2	1		0		0.5	40-60

sorbed by the animal if this (fat-soluble) vitamin is in a "water-solubilized" form than if it is administered as the usual solution in oil.

This fact led us to the developing of what is known as the "vitamin AD₃ massive dose" ("Duphasol" [*] AD₃) in which both vitamins are in aqueous solution [4].

This question of resorption touches on a general problem which also merits the greatest attention in the development of pharmaceuticals and pesticides. The value of a pharmaceutical or pesticide, and the same applies to a nutrient, depends not only on the active ingredient in the product but also on the "formulation" of the preparation. Formulation means the bringing of this component into such a form that it works as effectively as possible. In two other articles elsewhere

in this issue various aspects of this problem are dealt with [5].

These resorption problems also play an important part in the assessment of "mineral stable" dry vitamin A (and AD₃) preparations suitable for the mash feed industry. These preparations are made in the form of grains with a coating to give protection against oxidation and mineral action. Coatings of this type can impair resorption by the animal. In the development of such vitamin preparations full attention should therefore be given to the utilization by the animal. A

[*] Registered trade mark, N.V. Philips-Duphar.

[4] See Th. J. de Man, J. R. Roborgh and E. J. ten Ham, T. Diergeneesk. 83, 380, 1958.

[5] A. Verloop, Availability of organic compounds in higher plants, p. 93, and W. Duyfjes, The atomization of concentrated dispersion of pesticides, p. 112.

coating is required which gives a maximum stability with a minimum impairment of the resorption, and which protects a vitamin preparation which can be fully utilized by the animal. These requirements are found to be difficult to combine.

To attack this problem we devised a biological assay for evaluation of vitamin by means of a chick growth test^[6]. This method lent itself quite well to mathematical treatment. This approach has led to successful preparation of vitamins A and D ("Dohyfral"^[*] Extra) which combine good stabilization with good utilization by the animal: in fact the utilization was so good that it was higher than that obtained with the normal oil-based solution.

Finally, we note that vitamin A is not just a single compound, but that various stereo-isomers exist, and these have quite different levels of biological activity. This shows how complicated the whole problem is, and it is clear that the number of units shown on the label (as determined by the chemist) will only give very limited information about the actual value to the animal of a particular preparation.

In Table IV, which shows a number of vitamins there is also a column reserved for a "provitamin". It had been noticed for some time that various vegetable materials exhibited vitamin A activity and that this activity must lay hidden among the members of a group of compounds known as the carotenoid pigments. These can be converted by the animal itself into vitamin A, so that they should be considered as a "provitamin".

The greatest vitamin A activity is exhibited by β -carotene, which is included in the table; 0.6 μ g of this substance can be considered equivalent in activity to 1 international unit of vitamin A. It should be remembered, however, that this activity relationship between vitamin A and β -carotene applies only to the special circumstances of an official rat-growth test, in which moreover both compounds were administered as a solution in oil. In ordinary live-stock feeding, however, the situation for the provitamin is much more unfavourable, one reason being that carotene is resorbed much less readily from vegetable material than from a solution in oil and another that most farm animals convert the provitamin into the vitamin much less readily than the rat. If we also bear in mind that this conversion takes place mainly in the intestinal wall it will be evident that in all cases where this intestinal wall is not in excellent condition the vitamin A value of these provitamins can best be neglected (i.e. for coccidiosis in poultry, heavy infestations with other internal parasites, e.g. worms, and for vitamin A deficiency).

[*] Registered trade mark, N.V. Philips-Duphar.

[6] Th. J. de Man, E. J. ten Ham, J. R. Roborgh and N. Zwiep, Neth. J. agric. Sci. 6, 237, 1958.

Vitaminization of live-stock feed

One of the most difficult tasks in live-stock feeding is to ensure that all the vitamins are available in sufficient amount, both qualitatively and quantitatively, during the whole life of the animals, including the prenatal period.

In practice it is necessary to investigate the extent to which the recommended amounts of vitamins (Table III) can be covered by the ordinary feed (the basic ration) and to what extent it is necessary to supply part of this by means of extra vitamin preparations. This still leaves the question of how best to supply the extra vitamins. The answer to this question is important not only to cattle or poultry farmers but equally to the manufacturers of mash feeds and vitamin preparations. We are firmly convinced that the need for extra vitamins can be met best by vitaminization of mash feeds. In countries with a highly developed mash feed industry vitamins are among the most important feed additives, with antibiotics and coccidiostats.

If a correct vitaminization of mash feeds appears to be the most attractive way of providing the animals with the requisite extra quantities of vitamins, the limitations of this method should still be kept in mind. First of all, the method is limited to an extent depending on the relationship between the mash feed and the total diet. This relationship may differ greatly from country to country and even from district to district, and there will also be annual fluctuations (in connection with prices and the harvest yields). Moreover the place occupied by mash feed in the total ration depends greatly on the animal species and a few other circumstances. Roughly speaking, mash feeds make up the smallest quantity of the total feed in cattle farming, and the biggest in poultry farming, while in pig farming the position is intermediate.

It will thus be self-evident that in all cases where the mash feed only amounts to a rather small part of the total ration, and particularly where this proportion is liable to change, it will be very difficult to comply with the vitamin requirements of the animals in an economically justifiable way with the help of mash feed vitaminization alone. But in point of fact the situation is that even if the animals are fed entirely on normal vitaminized mash feed it may still be desirable to make use of other ways of giving extra vitamins (in addition to this mash feed). This will always be the case if, for any reason, the requirements for one or more vitamins are abnormally high. One method of giving extra vitamins is by way of the "oral massive dose" mentioned above. This treatment consists of a single dose, and the intention is that this will meet the requirements for vitamin for a considerable period of time. Not only

does the success of this treatment depend on good re-sorption of the vitamins present in the preparation, it also depends on possession by the animal organism of a good storage facility (e.g. in the liver) which can be drawn on as required.

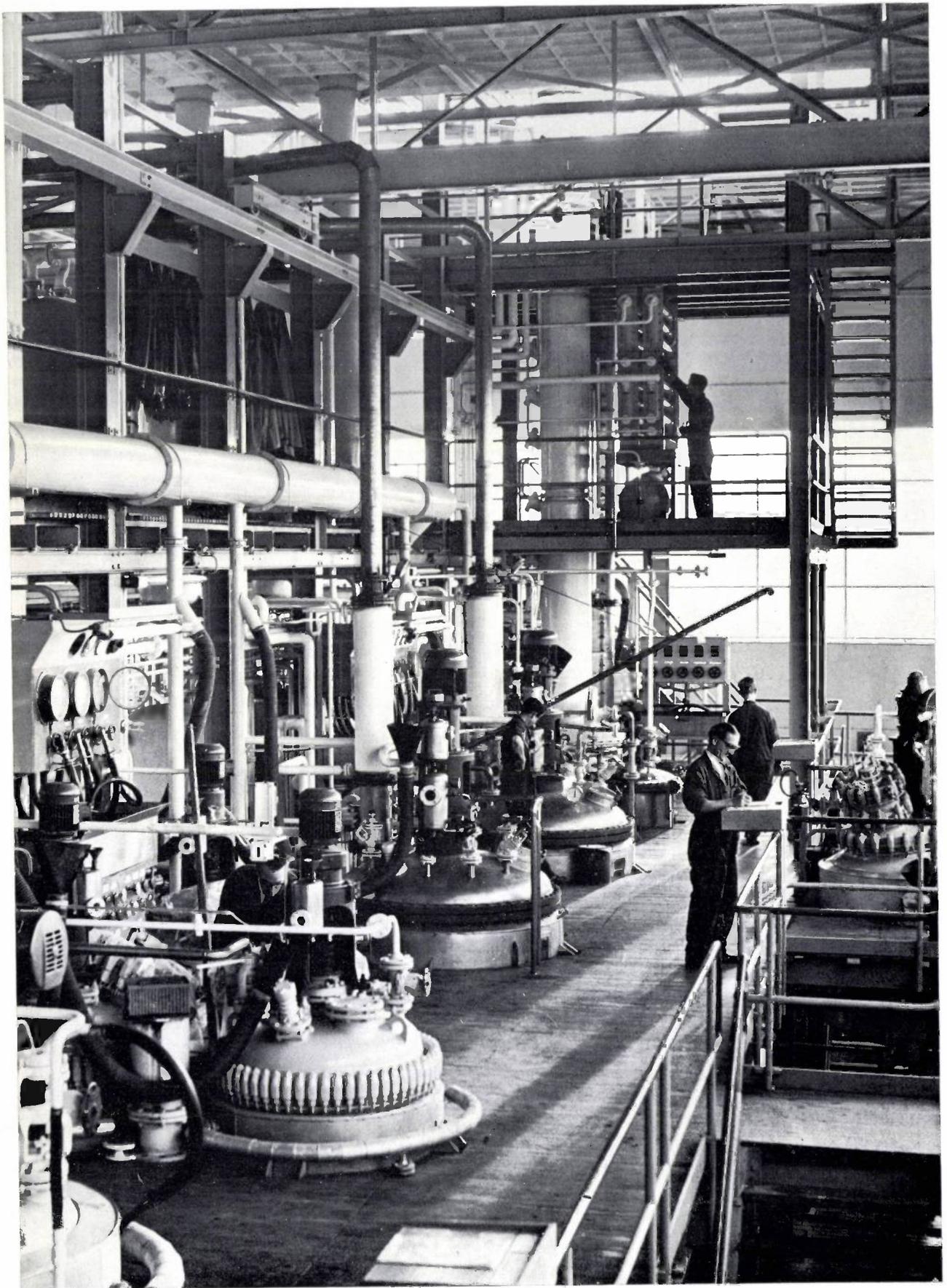
In practice this oral massive dose is frequently used for the fat-soluble vitamins A, D₃ and E, which for this purpose are given in a solubilized aqueous solution — a possibility already noted for vitamins A and D₃. In our opinion the components of the vitamin B complex are not very suitable for a true oral massive dose. The animal organism has only very limited storage possibilities for water-soluble vitamins, and it is therefore clear that this method of supplying vitamins of the B complex would not be economical and would give no certainty of protecting the animals against vitamin deficiency for a long period of time.

If circumstances compel the giving of extra B vitamins to the animals over and above the vitaminized mash

feed, one could choose to give a daily dose for a week or two. If desired, the preparation used for this (possibly completely soluble in the drinking water) might also contain the fat-soluble vitamins and, if necessary, trace elements and an antibiotic.

Summary. After a general introduction to vitamins, the article deals with the recommended amounts of vitamins in the feed for live-stock and of the vitamin contents of a number of usual live-stock feeds. A discussion follows about the problem of the extent to which the vitamin content as determined by chemical or physical methods is a reliable indication of what the animal is actually able to utilize. Thus it is found, for example, that vitamin A is resorbed to a greater extent if solubilized in water than if given as the usual solution in oil, and that the conversion of provitamin A to vitamin A depends greatly on the state of health of the animal. In this connection it is evident that a biological assay of a preparation yields better information about the actual value for the animal. These considerations have been taken into account in the development of new preparations, such as "Duphasol" AD₃ and "Dohyfral" Extra. In conclusion, attention is given to the various ways in which vitamin shortages can be made good in live-stock feeding.

Reaction vessels in one of the Philips-Duphar works



Availability of organic compounds in higher plants

A. Verloop

Introduction

Substances that bring about changes in living organisms are said to be biologically active. This property of chemical compounds depends as a rule upon a large number of factors, so that the relationship with the molecular structure is usually highly complex.

Attempts to analyse this relationship are based on a very simplified model (*fig. 1*). The biological system is represented by a membrane-enclosed space (compartment) which contains a receptor for the biologically

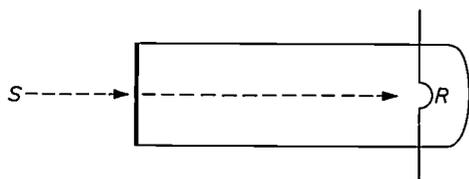


Fig. 1. Diagram showing the action of a drug in a biological system, for the simplest case possible. A receptor R is contained in a space enclosed by a membrane. The activity of the drug S is a function of the amount that can get through to the receptor (availability) and of the biological activity per unit concentration at the receptor site (specific biological activity).

active substance, the "drug". This receptor may be regarded as a small part of a biological macromolecule (an enzyme for example) possessing a very specific spatial structure on which the biochemical processes essential to the living organism take place. Interaction between the drug and the receptor can produce a biological effect by affecting or inhibiting the processes that normally take place [1].

In order to produce such an interaction the substance has to penetrate to the receptor; the concentration of the drug at the site of the receptor is called the *availability* of the drug. On the basis of the foregoing the biological activity of a substance could be defined as the product of availability and specific biological activity. By specific biological activity we mean the biological activity of unit concentration at the location of the receptor.

The availability at the site of the receptor is affected by two groups of processes. In the first place, the drug has to penetrate the cellular or sub-cellular barriers enclosing the compartment and then move to the recep-

tor (permeation and transport). Secondly, in the biological organism, it is subject to various breakdown processes, and part of the permeated substance may sometimes be prevented from reaching the receptor by other causes, such as adsorption and evaporation (from plants).

In actual research for biologically active compounds systems are encountered which are far more complex than the system of the model; a single cell alone is already more complicated than the model. Drugs administered to humans, animals and plants generally have to pass through a large number of "biological compartments" with their associated barriers, transport paths and metabolic processes, before reaching the receptor. In parasite control permeation in the parasite will often have to be taken into account as well as the availability of drugs in the host. In principle however the considerations applicable to the model remain valid.

Research on availability is important in the development of drugs. The object of this research is to learn something about the relations between the molecular structure or physical properties of chemical compounds and their availability in biological organisms. The application of these relations can be of value in the selection of classes of chemical compounds for drug research and can increase the efficiency of the search for optimum compounds in those classes in which biological activity has been found.

Research in this field has been carried out since 1900, particularly in connection with cell permeability [2], but the amount of research work has increased considerably in the last ten years. One of the motivations has undoubtedly been the need for safety in pharmaceutical application which makes it highly desirable to find out as much as possible about what happens to these substances in humans, animals and plants. Apart from this, it is now realized more clearly than in the past that a deeper insight into the availability of chemical compounds is of importance in the develop-

[1] The receptor concept was first put forward at the beginning of this century by Ehrlich and others. It has been elaborated upon in recent years by Ariens and associates; see E. J. Ariens, *Molecular pharmacology*, Vol. I, Academic Press, London 1964.

[2] See V. Wartiovaara and R. Collander, *Permeabilitätstheorien*, Springer, Vienna 1960.

ment of new drugs [3]. All this work is however still largely in an initial phase, in which attempts are being made to compile systematic data on the availability of these substances in various organisms [4].

Primarily with the object of illustrating the approach to the problems in this field, we shall discuss in this article two processes which affect the availability of biologically active compounds in *plants*.

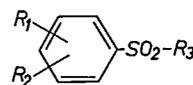
Plant processes which affect the availability of chemical compounds

The classical methods of controlling plant diseases were concerned solely with destroying parasites outside the plant, e.g. in the soil or on the leaves. In about 1945 the idea was conceived of trying to develop pesticides which, by analogy with the combating of disease in humans and animals, would be taken up and transported in the plant and thus provide it with internal protection against its invaders. These substances were referred to as *systemic* pesticides. The first indications that this protection method is in principle possible were obtained with organic phosphorus compounds developed in Germany during the second world war and which were found to possess systemic insecticidal activity [5]. Since then various other systemic pesticides have been developed. For the development of such systemic agents, and also of agents for controlling weeds (herbicides) it is essential to gain some understanding of their availability in plants [6]. In higher plants there are two important pathways by which the substances are translocated: the pathway via the parts above the soil (i.e. the leaves) and the pathway via the roots (and the soil). In the following we shall confine our attention to the second pathway. Through its roots the plant takes up water containing dissolved inorganic salts and other food materials. These are translocated through the transport ducts in the stems (the xylem) to the leaves. This "transpiration process" is maintained principally by the suction force exerted by the leaves as water evaporates from them.

The biologically active compounds which enter the

plant by this pathway, have to negotiate various "barriers". First of all, there are *adsorption* and *absorption* effects. Compounds of an apolar character are often strongly absorbed by the humus in the soil and by parts of the plants that contain a large amount of lipids. Ionized substances with a positive charge are often adsorbed on soil constituents and parts of plants (e.g. the walls of the xylem) which contain negative groups. Secondly the *permeation processes* play an important rôle: in order for a substance to enter the transport system in the plant it must be able to permeate through the roots, and for a good distribution (e.g. in the leaves) it must pass through a large number of cell membranes. Thirdly, many substances are subject to various *conversions* during the transport process. These include microbiological conversions in the soil and enzymatic processes in roots, stems and leaves. A final aspect, presumably unimportant in animals, but which can have a considerable effect on the availability of biologically active compounds in plants, is the *evaporation* from the leaves of substances taken up in the plant.

Each of these processes will generally have its own characteristic relationship with the molecular properties of the relevant compounds. There will therefore be a greater chance of learning more about these processes if the experiments are designed so as to permit separate study of the individual barriers in the various parts of the plant which a molecule has to traverse in penetrating from the outside to the receptor. The effects of permeation and transport on one hand and stability on the other should also be separately evaluated. By way of example we shall describe some experiments on root permeability and stability in the plant, making use of results obtained with the aromatic sulphones of the type



where R_1 and R_2 are arbitrary ortho-, meta- or para-substituents and R_3 is an alkyl group.

Root permeation

Method

Investigations on root permeation are carried out with an experimental arrangement making use of the exudation from tomato plant roots. Exudation is a physiological process in which the xylem in the roots continuously gives out water which can easily be collected (see *fig. 2*). We use roots from tomato plants which have been grown in identical conditions of illumination,

[3] See for example A. Albert, *Selective Toxicity*, Methuen, London 1965; and W. A. Sexton, *Chemical Constitution and Biological Activity*, Spon, London 1963.

[4] Special mention should be made of the important work done by the research teams led by Miller and McCallan, and by Brodie, on the availability of organic substances in fungi and animals, respectively. See for example S. E. A. McCallan and L. P. Miller, *Advances in Pest Control Research* 2, 107, 1958; B. B. Brodie, *ibid.*; T. B. Binns, *Absorption and Distribution of Drugs*, Livingstone, London 1964.

[5] G. Schrader, Monograph No. 62 in "Angewandte Chemie", Verlag Chemie, Weinheim 1952.

[6] With systemic pesticides the term "availability in plants" does not refer to the concentration at the receptor site (which is "in" the parasite), but to the concentration at the location where the parasite comes in contact with the pesticide.

temperature, etc. This standardization is necessary to obtain reproducible experimental results.

The roots are placed in a solution in water of the substance under investigation. The substance is taken up from this "external solution" and is metabolized in the root tissues, and the metabolites are then exuded

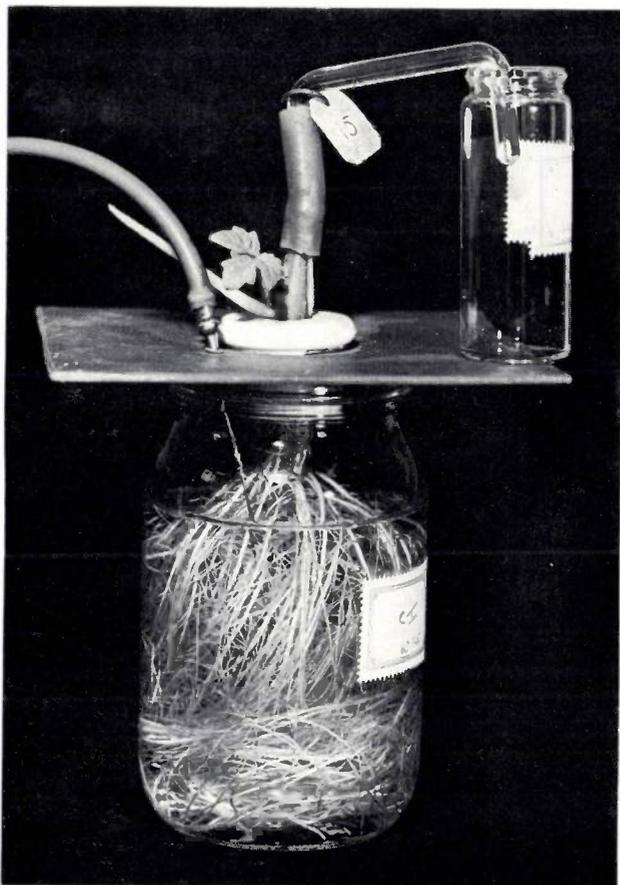


Fig. 2. Experimental arrangement for investigating root permeation (in tomato plants). The roots, cut off from the plants, continue to take up water and the substances dissolved in it, and exude it where the roots are cut off. The stability and the permeation of the dissolved substances are determined from the variation with time of the concentration of the external solution and of the exuded sap.

into the external solution. After some time the percentage conversion in the external solution will be approximately equal to that in the roots, unless the substance is particularly unstable in the roots.

The stability of the substance in the roots is determined by measuring the concentration in the external solution at the end of the experiment (after six days), and expressing this concentration as a percentage of the initial concentration after correcting for the amount that has permeated into the roots. The possibility that the conversions may have other causes such as chemical

instability in solution, bacterial conversions, etc., is checked by means of control solutions (with no roots present).

The permeation is determined by measuring the concentration of this substance in the exuded sap as a function of time. Some examples are shown in fig. 3.

Nature of the permeation process

According to the literature, the permeation of substances in biological systems may be *active* or *passive* in character. In active permeation the metabolism has a direct effect on the permeation mechanism. In passive permeation the passage of the substance through the biological membrane is determined entirely by physico-chemical factors.

In the literature various mechanisms have been described which are thought to play a rôle in active permeation. In many instances the *carrier* concept is used. A carrier is a natural substance occurring on the outside of the membrane, which can form specific bonds with an entering molecule. According to one theory the complex thus formed diffuses passively through the membrane along a concentration gradient, and then splits up again on the inside of the membrane. This

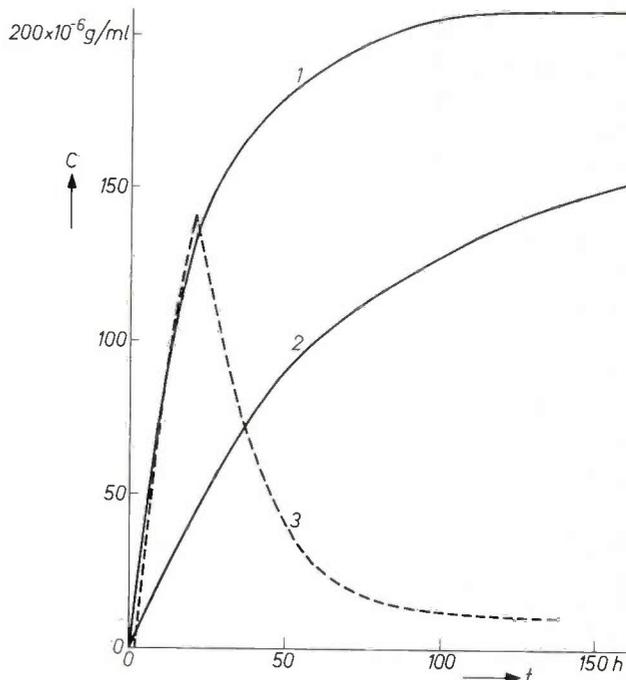


Fig. 3. Some results obtained with the experimental arrangement shown in fig. 2. The curves 1 and 2 show the variation with time of the concentration C of the compounds $\text{H}_2\text{N}-\text{SO}_2\text{NH}_2$ and $\text{CH}_3\text{CONH}-\text{SO}_2\text{NH}_2$ respectively in the sap exuded from tomato roots. The initial concentration of the external solution was 250×10^{-6} g/ml. The dashed curve 3 again shows the variation with time of the concentration of the first compound, but this time the roots were transferred to water 20 hours after starting the experiment.

splitting process, and possibly the biosynthesis of the carrier, is governed by the metabolism of the biological object.

Rosenberg and Wilbrandt have proposed various criteria for the recognition of active permeation [7]. These include:

- a) A constant rate of permeation above a critical saturation concentration, as all the available carrier molecules are then occupied. In root permeation experiments with a number of biochemically stable aromatic sulphones we found a practically linear relationship between external concentration and rate of permeation in the concentration range between 0.03 and 4 mmole/l. This shows that there are no saturation phenomena in this case.
- b) Inhibition of the permeation process by inhibitors, which stop the active process by disturbing the processes supplying energy. Using the compound p-aminophenylmethylsulphone an investigation was made into the effect of the inhibitors NaF, NaN₃, 2,4-dinitrophenol and KCN on permeation, and also into the effect of the supply of O₂ in the root system. Although the rate of exudation of water (which is also dependent on the metabolism for its energy supply) may indeed be affected by these factors, no direct effect at all on the permeation of the sulphone was found. On the other hand, no direct effect on the permeation was found after the addition of salts (e.g. KNO₃) which strongly stimulate the exudation.
- c) High structural specificity of the permeating substances, since they must "fit" the carrier molecules. Permeation through tomato roots was investigated for about 150 neutral organic compounds. Compounds were chosen which are stable enough in the relevant biological system to allow the permeability to be determined with reasonable accuracy. No marked structural specificity was found, although there were some indications of a gradual change in the permeation properties following structural variations which led to a change in the physico-chemical properties of the relevant compounds.
- d) High temperature coefficient of the permeation constant. Here also, there were no indications pointing to an active process in root permeation.

It may be concluded from the foregoing that root permeation of the organic compounds which we have investigated is a passive process. This contrasts with the root permeation of inorganic ions, whose active character has been fairly clearly established by other investigators.

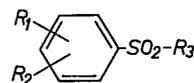
In view of the non-specific character of passive permeation it seems reasonable to assume that many other types of organic compounds can also permeate the

roots of plants in the same way. This is a vitally important point, because only for this type of permeation process does it seem possible to find rules which are generally applicable.

Relation between physico-chemical properties and root permeation

The hydrophilic/lipophilic balance of a molecule (HLB) was as early as 1900 described in literature as a physico-chemical quantity which has a very marked effect on the permeability in plant cells. This HLB may be defined as the ratio of the affinity of a substance for a more polar medium (e.g. water) to its affinity for a more apolar medium (e.g. ether or olive oil). A measure of this ratio is the distribution coefficient between, say, water and ether. In order to permeate into a cell the substances have to pass through the cell membrane. The observed effect of HLB on permeation led at the time to the assumption that the cell membranes must contain large amounts of lipophilic substances, in this case lipids. This hypothesis was later confirmed by analysis of the cell membranes.

Fig. 4 shows the permeation of neutral aromatic sulphones of the type



in the exudate, plotted against a measure of the HLB. No relationship is found whatsoever. However, when the results of those compounds which are not biochemically stable are omitted, a rough relationship can be seen (fig. 5). This illustrates how necessary it is to keep stability and permeation data separate when interpreting experimental results.

Just as with separate plant cells, the HLB of the molecules obviously plays an essential rôle in permeation in the much more complex systems of tomato roots. With other groups of substances and for the root systems of other plants similar relations have been found to exist between permeation and HLB parameters. These results lead to the conclusion that it is possible to formulate physico-chemical rules governing the permeability of (neutral) organic compounds.

Biochemical stability in plants

Method

We have already seen how the stability of a substance in the roots of a plant can be determined. To determine the stability of a substance in the plant as a whole a

[7] Th. Rosenberg and W. Wilbrandt, *Int. Rev. Cytology* **1**, 65, 1952.

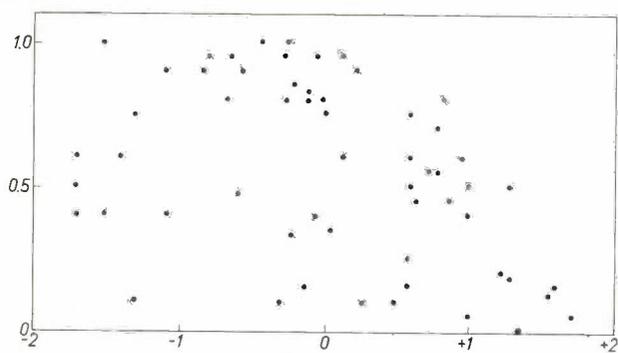


Fig. 4

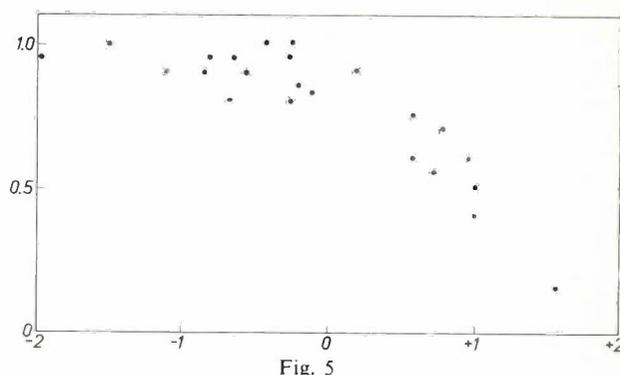


Fig. 5

Fig. 4. Permeation of aromatic sulphones through tomato plant roots, The logarithm of the distribution coefficient between a polar and an apolar solvent is shown on the horizontal axis; the vertical axis indicates the saturation concentration in the exuded sap divided by the initial concentration in the external solution.

Fig. 5. As fig. 4, but excluding the biochemically unstable compounds.

different method is employed, which we shall illustrate by describing experiments carried out with broad bean plants. The experimental arrangement is shown in *fig. 6*. The plants are placed with their roots in a solution in water of the substance under investigation; the uptake of the substance is governed by the transpiration mechanism previously described. To obtain comparable results all experiments were performed in identical conditions of temperature, relative humidity and

illumination, since these experimental conditions affect the evaporation of water from the leaves, and therefore the amount of water and substance taken up. The conditions are chosen in such a way that about half of the external solution is taken up during the course of the experiment, giving the best situation for obtaining quantitative results. By extraction of roots, stems and leaves, and analysis of the extracts, it is possible to determine the stability of the relevant substance in the

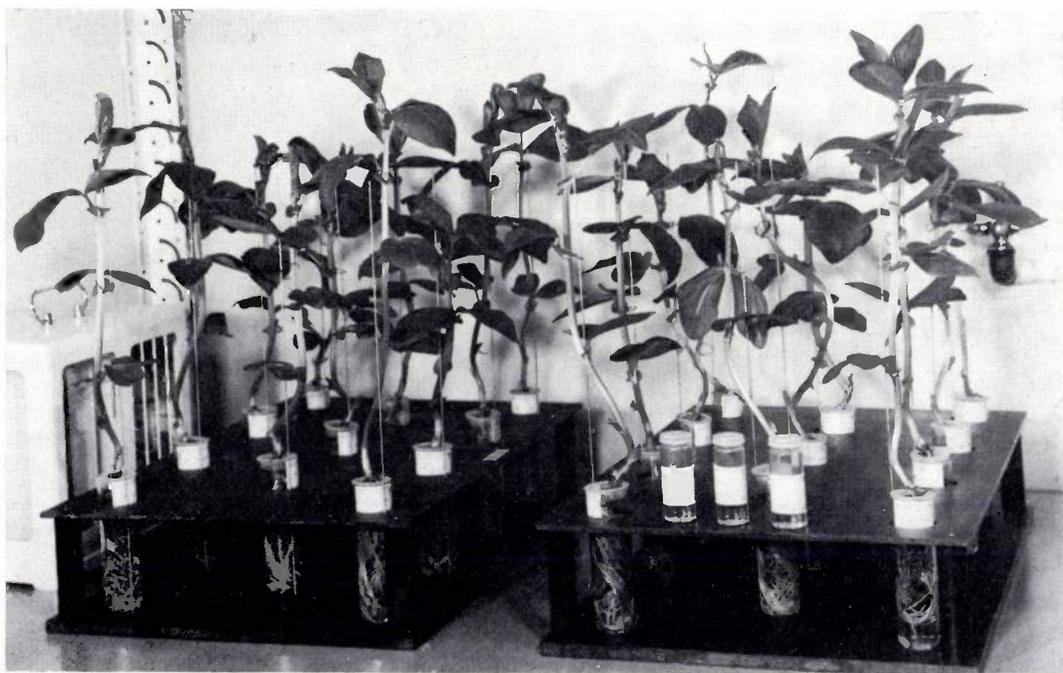


Fig. 6. Experimental arrangement for determining the stability of the chemical substances in plants (broad bean). When the plant has taken up about half of the external solution, roots, stems and leaves are extracted and analysed. The stability of the substance in the plant is expressed as the percentage ratio of the amount of substance found in the plant to the amount of substance taken up by the plant.

plant. This is generally defined as the ratio of the quantity of substance found in the plant to the quantity of substance taken up by it. The accuracy achieved is between 5 and 10%.

Stability of basic structure and substituents

In systematic research on a particular class of organic chemical compounds structural changes in the molecule are generally introduced only in one place at a time. In aromatic and heterocyclic compounds, for example, the ring system and all substituents except one are left unchanged; this part of the molecule might be called the basic structure. This basic structure may show considerable instability in some types of compounds. For instance, in all compounds investigated of the type $R-\text{C}_6\text{H}_4-\text{CSNH}_2$ (thiobenzamides) the stability in the external solution in the exuded sap experiments proved to be less than 50%, and it is even lower in the intact plants. Further investigation has shown that this is due to the marked instability of the thioamide group, which is found in even simpler molecules such as CH_3CSNH_2 (methyl thioamide).

In various other classes of organic compounds the basic structure is stable, so that experiments with substituted compounds can give information on the stability of the substituents. If conversion takes place the nature of the products formed must then be known, for it is of course also possible that the activating effect of substituents can make the basic structure itself unstable.

As an example we shall discuss the group of aromatic sulphones of the type $R-\text{C}_6\text{H}_4-\text{SO}_2\text{CH}_3$. Table I shows the stabilities of a number of substituted compounds.

As can be seen, a few substituents are converted to a considerable extent. These are:

the NO_2 group, which in the first instance is reduced to NH_2 ;

the $(\text{HC}=\text{O})$ -group, in which there is presumably conversion to COOH ;

the OH -group, where it is assumed, on the basis of similar results in the literature, that there is a conversion into a glycoside;

the NH -glucosyl group, where hydrolysis to the free NH_2 compound has been demonstrated.

On the other hand, with various other substituents such as CH_3 , OCH_3 , Cl and CN , the stability in the plant is very high.

Interesting behaviour is shown by the NH_2 substituted compound, in which it was found that the stability was strongly dependent on the concentration of the external solution. Upon substitution of the NH_2 group it is found in some cases (*di*-alkylation, acetylation, phosphorylation), that this dependence can be completely neutralized but in others this result is not

Table I. The stability of substituents R of aromatic sulphones of the type $R-\text{C}_6\text{H}_4-\text{SO}_2\text{CH}_3$ in experiments on broad bean plants (uptake during 48 hours).

Substituent R	Stability in plant (in %)	
	concentration of external soln. 2 mmole/l	concentration of external soln. 0.25 mmol/l
H	95	—
CH_3	100	—
OCH_3	90	90
Cl	80	80
CN	90	90
NH_2	70	20
$\text{NH}-\text{CH}_3$	60	30
	70	65
$\text{NH}-\text{CO}-\text{CH}_3$	90	90
$\text{NH}-\text{PO}-\{\text{N}(\text{CH}_3)_2\}_2$	100	90
NH -glucosyl	35	15
OH	20	10
$\text{HC}=\text{O}$	0	0
NO_2	0	0

found (*mono*-alkylation, glucosidation).

In order to find out more about the back-ground to these phenomena, an investigation was made into the behaviour of the compound $\text{H}_2\text{N}-\text{C}_6\text{H}_4-\text{SO}_2\text{CH}_3$, radioactively labelled with ^{35}S , in broad bean and tomato plants. It was found that about 10% of the substance taken up is converted into glucosyl and acetyl derivatives soluble in water. Another part, however, is bound to polymeric constituents of the plant (not soluble in water) and the extent to which this occurs depends strongly on the concentration used in the experiment. At a concentration of 2 mmole/l in the external solution about 10% is converted in this way, but at 0.25 mmole/l the percentage has risen to about 60%. The latter quantitatively much greater conversion is therefore responsible for the dependence on concentration shown by the stability of the free NH_2 compound. This effect of the concentration is presumably due to the presence of a limited number of active sites at the plant polymers. On the basis of results obtained and from model experiments it is assumed that the polymeric bonding also involves two types of reaction, one being the formation of a glycoside bond and the other being the formation of an acyl bond (see fig. 8).

When the plants are extracted with boiling water, the free NH_2 compound and the glucosyl and acetyl derivatives are quantitatively extracted. The same applies to the substance bound to the plant polymers by a glycoside bond, because in these conditions quantitative hydrolysis of this bond takes place. The *p*-aminophenylmethylsulphone which is bound to the polymer by a

bond of the type RNHCO-polymer, can only however be isolated after extraction using 0.5 mole/l HCl at 100 °C. This provides a neat method of determining the approximate location of this latter bond in the plant by making autoradiographs of the leaves before and after extraction with boiling water. It can be seen from the results presented in *fig. 7a* and *b* that this bond is located at or near the walls of the xylem vessels.

With the information obtained in this way it is therefore easy to explain the effect of the various N substi-

one might expect a similar effect on the reactivity of substituents towards enzymatic systems. Investigations using other stable basic structures have shown that this is in fact the case; the substrate-enzyme interaction or the reactivity of the substrate in the substrate-enzyme complex can be changed by altering electron density or the spatial situation of the substituent.

In addition there is another factor which can affect the stability of a substituent. An example of this is shown in *Table II*.

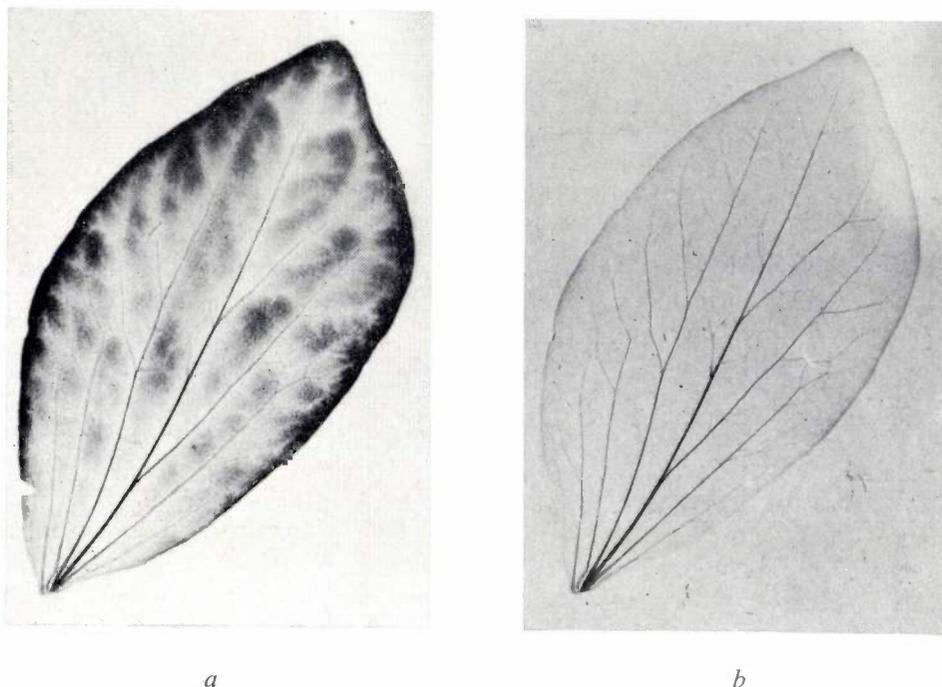


Fig. 7. Autoradiographs of a leaf of a broad bean plant which has taken up $\text{H}_2\text{N}-\square-\text{SO}_2\text{CH}_3$ radioactively labelled with ^{35}S , *a*) before and *b*) after extraction with water at 100 °C. This extraction removes all the radioactive compounds, except those which are bound by an acyl bond to polymeric constituents of the plant. This enabled the acyl bond to be localized: it appears to occur at or near the walls of the xylem vessels.

tutions. Thus, di-alkylation removes the possibility of the indicated reaction, whereas mono-alkylation does not; acetylation, phosphorylation and glucosidation also protect the NH_2 group, but the glucoside is not stable in the plants.

In all cases investigated the basic structure $\square-\text{SO}_2\text{CH}_3$ is found to be unaffected, so that in this case the stability of the *para*-substituents has in fact been determined. A summary of the results is presented in *fig. 8*.

The question arises as to whether such stability results found for substituents with the aid of a particular basic structure might possess more general validity.

In organic chemistry it is quite common for the reactivity of a particular substituent group to be affected by the rest of the molecule; on *a priori* grounds

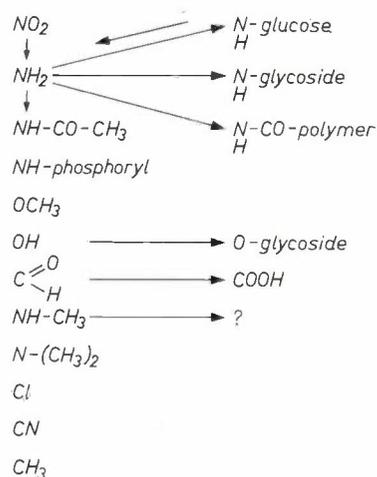


Fig. 8. Possible conversions in broad bean plants of the substituents R in the compounds $\text{R}-\square-\text{SO}_2\text{CH}_3$. If no indication is shown behind a substituent it is stable.

Table II. Effect of the hydrophilic/lipophilic balance (HLB) of the total molecule on the biochemical conversion of the CH_3O -group into a hydroxyl group.

Compound	Distribution coefficient water/ether	Stability in broad bean plants after 14 days (in %)
$\text{CH}_3\text{O}-\text{C}_6\text{H}_4-\text{SO}_2\text{CH}_3$	0.27	90
$\text{CH}_3\text{O}-\text{C}_6\text{H}_4-\text{SO}_2\text{C}_4\text{H}_9$	0.02	10

Whereas the electron density and the spatial situation of the CH_3O -group are found here to be virtually equal, the stability is seen to differ very considerably. There is however a great difference in the HLB of the two molecules, which is illustrated by the distribution coefficients given in the table. This effect on the stability of substituents is probably due to subcellular *localization* of the enzymes that cause the relevant conversions in a part of the cell with high lipid content.

Further research will have to show whether it is possible to obtain a sufficient understanding of the effect of factors we have mentioned to make it possible to predict the stability of substituents in potential pesticides.

Summary. The availability of chemical substances in a living organism is defined and discussed as one of the factors determining their biological action. The approach to the problems in this field is illustrated by considering certain processes that affect the availability of chemical substances in plants. These processes, namely root permeation and biochemical conversion, are discussed with reference to experimental results obtained with aromatic sulphones. The nature of the root permeation process and a relationship found with certain physico-chemical properties of the investigated compounds are dealt with in more detail. It is concluded that it should be possible to draw up general rules for this process. The effect of biochemical stability is discussed in terms of results obtained on the stability of various substituents in phenylmethylsulphone. It is shown that this stability of substituents does not necessarily have the same value in other classes of organic compounds since it is affected by the rest of the molecule.

Phenol synthesis and photomorphogenesis

G. Engelsma

Phenomenology of photomorphogenesis

The young seedling's struggle for life in the midst of other plants which may already have an advantage over it is essentially a struggle for a place in the sun. The sun provides it with the energy it needs to go on living after the reserves contained in the seed have been used up. Throughout the entire life of the plant the lighting conditions continue to be of the utmost importance, and it is not surprising that these are an important factor in the adaptation of the plant to its environment.

The mode of adaptation depends on the plant species. This appears for instance in the germination of seeds: seeds of some plants germinate only after they have been exposed to light for a short time; the germination of others, on the other hand, is inhibited by light. The same is true of flowering: there are plants that do not flower unless they have been exposed to light for longer than a certain continuous period per day (long-day plants); others will flower only when the daily light period is shorter than a critical length (short-day plants). The usefulness of these adaptation mechanisms to the plant would appear to be that they increase a plant's chances of survival and of propagating itself.

In order to illustrate some aspects of the development of the plant under the influence of light — photomorphogenesis — we have chosen the gherkin seedling, a plant which has been the subject of investigations at our Research Laboratories for a number of years [1].

Fig. 1 shows two seedlings of the same age, 5 days after being sown. The plant on the right was grown in the dark, and the plant on the left was exposed to light for 12 hours per day. The dark-grown plant makes a colourless impression, because the green pigment necessary for photosynthesis, chlorophyll, has not formed. The stem, the hypocotyl, is exceptionally elongated and is barely able to support the cotyledons, the original seed-lobes. The latter have increased in size only slightly since germination. Comparing this with the illuminated seedling, we see that the hypocotyl is much shorter and much sturdier. The initial bend in the part of the hypocotyl near the cotyledons, the plumular hook, has opened and the cotyledons, which have become green, are in the process of growing into organs with a leaf function. Apart from these differences visible to the naked eye, it is possible with special aids to



Fig. 1. Gherkin seedlings 5 days old. Right: grown in the dark. Left: exposed to light 12 hours per day.

ascertain internal changes which are due to the exposure to light, such as the accelerated breakdown of the storage products present in the cells of the cotyledons, an increase in the ascorbic acid content and an accelerated accumulation of phenolic compounds both in the cotyledons and in the hypocotyl.

If the first plant receives no light it will die after about nine days, even before it has used up all its reserves. The second plant, on the other hand, under the effects of light, has built up a photosynthetic apparatus capable of converting light energy into chemical energy [2]. One of the functions of the photosynthetic process is to produce from the carbon dioxide in the air and nitrogenous compounds in the soil the building elements it needs for further growth.

At this point it should be noted that a clear distinction should be made between photosynthesis and photomorphogenesis, that is to say between the energy-supplying and the formative effects of light. The formative effect is usually produced with much smaller

[1] G. Meijer, *Acta bot. neerl.* 6, 395, 1957; 7, 614, 1958; 7, 621, 1958; 8, 189, 1959.

G. Meijer, *The spectral dependence of flowering and elongation*, thesis, Utrecht 1959.

G. Meijer and G. Engelsma, *Photochem. Photobiol.* 4, 251, 1965.

[2] J. S. C. Wessels and M. van Koten-Hertogs, *Photosynthesis*, *Philips tech. Rev.* 27, 241-257, 1966 (No. 9/10).

amounts of light and, as we shall see presently, it also involves other pigments.

For every plant species the development from germinating seed to fruit-bearing plant follows a pattern which is characteristic of that species and is genetically fixed. An adaptation to the environment is obtained through the effects which the factors that constitute that environment, and light in particular, exercise on particular phases of the pattern of development.

The development of the plant may be divided into distinct phenomena such as germination, root formation, stem elongation, opening of plumular hook, leaf formation and flowering. All of these may to a certain extent be studied as individual processes. On the other hand, it is of course obvious that they cannot be conceived of as distinct from the developing plant as a whole. They are in fact so interrelated that the plant is able to evolve into a harmoniously functioning entity.

In this article we shall be concerned in particular with two phenomena that occur more or less simultaneously when gherkin seedlings grown in the dark are exposed to light. These phenomena are the inhibition of the elongation of the hypocotyl and the accelerated accumulation of phenols (hydroxycinnamic acids), phenomena which we shall consider both in their interrelationship and against the background of photomorphogenesis in general.

Let us now turn to the practical prospects of the investigation. From the above it follows that it is possible to guide the development of the plant in a certain direction by means of the factors which constitute the environment, in particular by means of light. This has been the point of departure of the investigations on light and plant growth at the Research Laboratories in Eindhoven [1]. Important applications exist in this field, which are however mainly restricted to hothouses. With the passage of time there has been an increasing interest in chemical methods of controlling plant development which could also be used for cultivation in the open. Work of this type is closely connected with the weed control work at Philips-Duphar. For both subjects it is highly important to have an insight into the regulating mechanisms underlying the development of the plant. One way of obtaining such an insight, is by investigating the way in which the metabolism of the plant responds to a particular stimulus from the outside, and light, which does not damage the plant, provides a very suitable stimulus for this purpose.

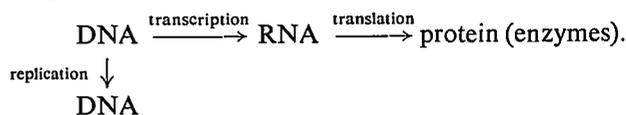
The molecular basis of photomorphogenesis

A plant, as is well known, is composed of a large number of cells which may be divided into different types depending on structure and function. Starting at

the outside, a cell consists in general of 1) a cell wall, which is formed by the rest of the cell, 2) protoplasm, in which nearly all processes essential to the maintenance and functioning of the cell take place and 3) one or more vacuoles, which are separated from the protoplasm by a membrane and in which organic and inorganic compounds are stored in an aqueous solution. The protoplasm contains various structured bodies, called organelles, each with specialized functions. Among these are the nucleus, which is the information centre of the cell, the mitochondria and the chloroplasts, which are responsible for the energy supply in the cell, the plastids, which are the storage compartments of the cell, and the ribosomes, where the enzyme synthesis takes place.

Enzymes play an important role in the building up of the cell and in its functioning. They control the chemical reactions that supply the elements for building up the structure of the cell and of its organelles, as well as the reactions that occur in and around these organelles.

Since, as will be shown, the *synthesis* of enzymes is an important aspect of photomorphogenesis, it is necessary to deal with it here at somewhat greater length. The mechanism underlying what is termed *de novo* enzyme synthesis, that is the synthesis of enzymes from their basic elements, the amino acids, may be represented in simplified form as follows:



This representation is based mainly on the study of bacteria and bacteriophages, but there is good reason to assume that the mechanism is roughly speaking analogous in the higher plants. The genetic information for the plant is contained in the nucleus of each cell, in the form of very long polymeric chains of DNA (deoxyribonucleic acid), whose four essential components form a code which determines the composition of the enzymes. The role of DNA is twofold: it can reproduce an exact replica of itself, which may become the information centre of a new cell; and it is able, by transcription, to transmit parts of its information in the form of a new code, exactly corresponding to the original, to the RNA (ribonucleic acid) a polymer of analogous composition. Part of this RNA can now serve as a matrix to which the amino acids become attached in a particular sequence to form polypeptide chains on the ribosomes.

It was mentioned at the outset that the cells which make up the plant differ in structure and function. These differences can be related to differences in the enzyme pattern of the specialized cell. It has been found from experiments on tissue cultures, for example,

that a whole plant can be grown from a single cell, and it may be concluded from this that each individual cell contains the genetic information for the entire plant, and hence for all enzymes encountered in the plant. Since the specialized cell possesses only part of these enzymes, we must assume that it makes only partial use of this information. Certain genes are blocked, or in other words only part of the DNA is capable of having its code converted into enzymes. By the action of certain compounds, known as inducers and repressors, particular parts of the DNA can be deblocked or blocked, thus causing shifts in the enzyme pattern. At present hardly anything is known about the mechanism responsible for this. In some cases it has been found that the addition of a particular compound leads to the induction or repression of one or more enzymes closely involved in the breakdown or synthesis of this compound, thus pointing to a regulating mechanism by means of which the living organism adapts itself to the changed supply of a particular metabolite. There are also indications that particular hormones, compounds which exercise a co-ordinating function between the various organs of the living organism, act via enzyme induction.

The mechanism referred to thus determines which enzymes are to be made and what quantity of a given enzyme is to be produced. A further regulating mechanism also exists which, by means of special compounds known as cofactors and inhibitors, increases or decreases the rate at which an enzyme functions.

It appears that both regulating mechanisms play a part in the formative effects of light on the plant. We shall try to explain this with reference to what is already known about phenol synthesis and the inhibition of elongation of the gherkin seedling.

Light-stimulated phenol synthesis

In the hypocotyl of the gherkin two hydroxycinnamic acids are synthesized, *p*-coumaric and ferulic acid [3], both compounds with a phenolic character, i.e. characterized by the presence of the phenol group: $\text{C}_6\text{H}_4\text{-OH}$. Since the two phenols are not converted into other compounds, or at least not quickly, the rate of phenol synthesis can be derived from the rate at which the phenols accumulate.

As long as the plant has not been exposed to light it contains only a small amount of these phenols. They are chiefly concentrated in the upper (apical) part of the hypocotyl. After the transition from dark to light it takes about two hours before the phenol synthesis becomes measurably faster (fig. 2). If the plant is contin-

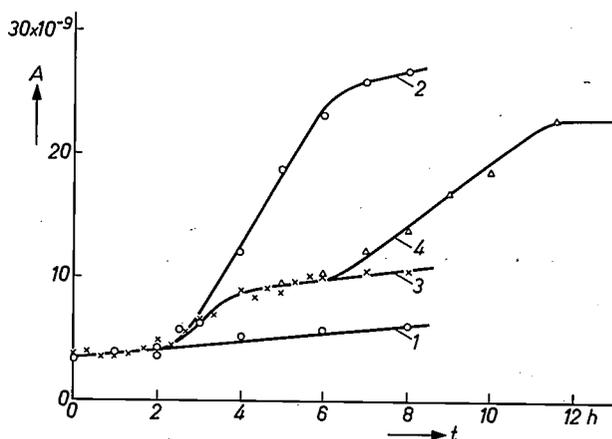


Fig. 2. The quantity *A* of hydroxycinnamic acids per hypocotyl in dark-grown gherkin seedlings three days old, as a function of time. The four curves refer to plants which after this period were: 1 left unexposed to light, 2 continuously exposed to light, 3 illuminated for one hour and then returned to the dark, 4 illuminated for one hour, then left three hours in the dark, and subsequently continuously illuminated. The quantity of hydroxycinnamic acids is given in moles. Illumination: blue light (400-500 nm), 600 $\mu\text{W}/\text{cm}^2$.

uously exposed to light, then after six to seven hours the quantity of hydroxycinnamic acids per hypocotyl reaches a level that changes very little during the next 40 hours — irrespective of whether the illumination is continued or stopped (curve 2). If the exposure is shorter, for example only one hour, the rate at which the hydroxycinnamic acids accumulate tends to decrease earlier, and the total quantity per hypocotyl remains at a lower level (curve 3). The plants thus illuminated have retained the capacity, however, to synthesize more hydroxycinnamic acids — up to about the level reached with continuous exposure — as is found when they are again exposed to light (curve 4).

The stimulation by light is always greatest in the apical part of the hypocotyl; it begins with an increased synthesis of *p*-coumaric acid, and later more ferulic acid is formed. It is consistently found that the ratio of ferulic acid to *p*-coumaric acid is greater in the basal (lower) part than in the apical part of the hypocotyl. The possible significance of the ratio of these phenols in connection with elongation will be dealt with presently.

Hydroxycinnamic acids occur in many higher plants. As a rule they are present in the form of sugar derivatives, as in gherkins. These compounds, or the corresponding alcohols, form the structural elements of lignin, a polymeric product that serves mainly to strengthen vascular bundles, but which in woody tissue also settles in and on the walls of other cells of the trunk or stem. Apart from their function in elongation, it has been postulated that phenols also play a regulating role in germination, budding, leaf formation and flowering.

Two chemically related groups of plant pigments, the flavonoids and anthocyanins, are compounds which may also strongly accumulate during a relatively short period following the mo-

[3] G. Engelsma and G. Meijer, Acta bot. neerl. 14, 54 and 73, 1965.

ment at which a seedling is first exposed to light^[4]. Exactly what purpose this serves for the plant is not yet clear. It is possible that these aromatic compounds, which absorb strongly in ultraviolet light, protect vital constituents of the cell, such as nucleic acids and enzymes, against damage from radiation.

It is generally assumed that the hydroxycinnamic acids in the plant are synthesized from sugars in the manner represented, in simplified form, in *fig. 3* (i.e. by the shikimic acid pathway). Of the intermediates postulated in this process, only L-phenylalanine and cinnamic acid are found to stimulate the synthesis of hydroxycinnamic acids in the gherkin seedling. In this case *d*-coumaric acid is mainly formed. On the basis of this information we can assume that the *p*-coumaric acid is

alanine-deaminase^[5], which is fairly stable, so that it is possible to determine enzyme concentration quantitatively in an *in vitro* test^[6]. If the dark-grown seedlings are continuously exposed to light, the enzyme concentration begins to increase after about 90 minutes (*fig. 4a*, curve 2). After three hours the concentration reaches a maximum, and then gradually decreases. If the seedling is illuminated for only one hour, the enzyme concentration reaches its maximum after a corresponding period of about three hours (curve 3). The maximum in this case is lower, and moreover the decrease is faster. After renewed illumination the enzyme concentration again begins to increase after 60 to 90 minutes (curve 4).

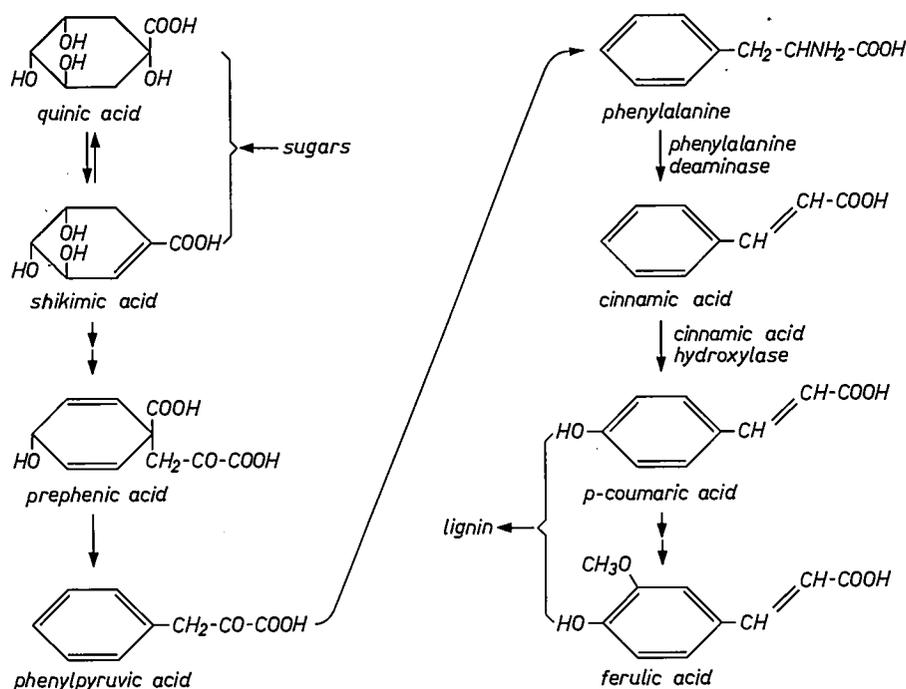
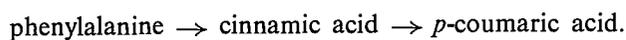


Fig. 3. Synthetic pathway for hydroxycinnamic acids.

synthesized from phenylalanine that was probably originally stored in the seed, and that formation of cinnamic acid is the intermediate step:



Theoretically, the stimulating effect of light in the synthesis of *p*-coumaric acid may be due to an increased supply of phenylalanine or to an effect upon the enzymes involved in the conversion of this compound. So far we have investigated the latter possibility only.

The first step, the deamination of phenylalanine, takes place under the influence of the enzyme phenyl-

Fig. 4b shows the rate of phenol synthesis (i.e. the first derivative of the curves in *fig. 2*) as a function of exposure time for the same illumination programs — continuous exposure; one hour exposure followed by darkness; one hour exposure, three hours of darkness, followed by uninterrupted exposure. Comparison of *figs 4a* and *b* shows a surprising correlation: increased synthesis of hydroxycinnamic acids is invariably preceded by an increase in enzyme concentration, and reaches a maximum rate at about the time when the enzyme concentration is at its maximum. The decrease in the quantity of enzyme is eventually followed by a decrease in the rate of phenol synthesis. The length

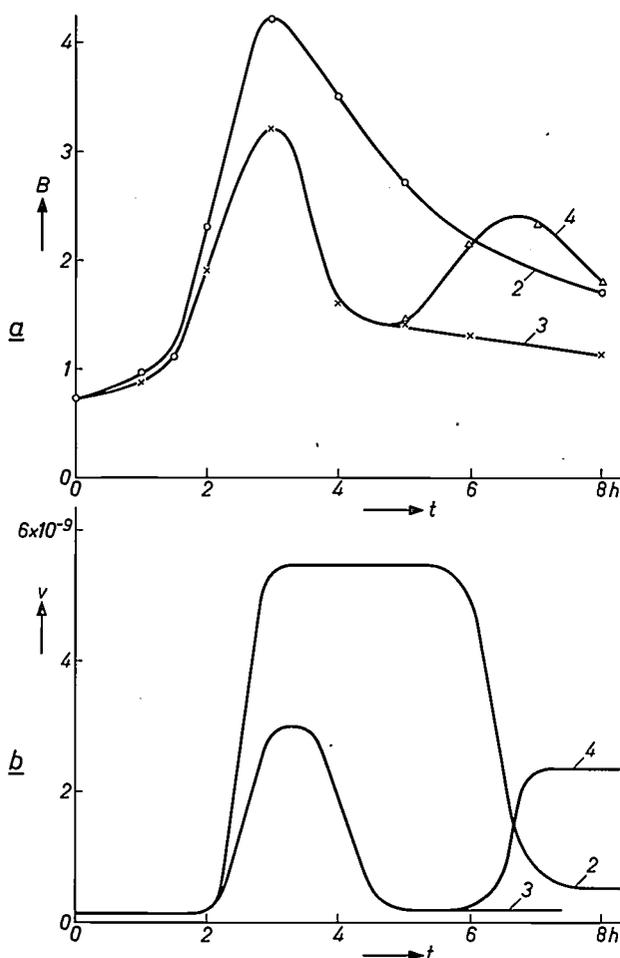


Fig. 4. *a*) The quantity B (in activity units, see [6]) of the enzyme phenylalanine deaminase per hypocotyl in dark-grown gherkin seedlings three days old, subjected to the same light treatments as in fig. 2.

b) The rate v (moles/hour) at which the hydroxycinnamic acids accumulate in the hypocotyl of these seedlings, as derived from the curves of fig. 2.

of the time interval between the two phenomena is presumably determined by the quantity of intermediates (cinnamic acid) that has temporarily accumulated. The rapid decrease in the quantity of active phenylalanine deaminase indicates the existence of a regulating mechanism which ensures that part of the induced enzyme quickly becomes inactivated in the plant. The nature of this mechanism is not yet clear.

As we can see from the figure, the underlying regulating mechanism is such that the plant, when stimulated by light to accelerate the phenol synthesis, does not continue this process. If this were the case, it would in fact be rather disastrous for the metabolism as a whole. Another notable aspect of the regulation of phenol synthesis is that the quantity of phenylalanine deaminase induced by light depends on the preceding illumination (see curve 4 in fig. 4*a*). If the preceding period of illumination has been sufficiently long — about 8 hours, with the light intensity used in this instance — then it is not possible to increase the enzyme

concentration again. The plant apparently has a mechanism that sets a definite limit to the quantity of enzyme that can be induced [7].

The next step in the synthesis of *p*-coumaric acid — the hydroxylation of cinnamic acid — is more difficult to study because the enzyme that catalyzes this reaction, cinnamic acid hydroxylase [8], is very unstable *in vitro* and therefore the enzyme cannot be isolated. Since we are therefore unable to ascertain the enzyme concentration *in vitro*, we have no alternative but to attempt to determine it *in vivo*. The activity of the enzyme is then derived from the quantity of cinnamic acid converted into *p*-coumaric acid in a segment of the hypocotyl under standard conditions [9]. It is found from these experiments that the cinnamic acid hydroxylase activity depends on the lighting conditions in much the same way as the phenylalanine deaminase activity.

It appears likely that, in addition to regulation of the enzyme level, regulation of the rate of enzyme activity by means of cofactors and inhibitors also comes into play. The activity of phenylalanine deaminase is greatly enhanced by glutathione; the action of cinnamic acid hydroxylase presumably depends entirely upon the presence of certain pteridine derivatives. In both cases reduced compounds are involved, and it is known that the concentration of some of these compounds in the plant likewise increases in the light [10].

The role of these compounds in the control of phenol synthesis has not yet been fully investigated. Further particulars of regulation at the enzyme activity level will be given when we come to discuss the inhibition of elongation.

With regard to regulation via enzyme synthesis, it should be noted that the enzyme synthesis in fact takes place in the manner already discussed, that is from the basic elements on an RNA matrix which in its turn is formed on a DNA matrix (*de novo* enzyme synthesis). That it is not the result of changes in previously existing protein molecules has been demonstrated with the aid of two compounds which block the mechanism for enzyme synthesis: actinomycin D, which prevents the transcription from DNA to RNA; and cycloheximide, which blocks the translation of RNA into protein. Both compounds inhibit the stimulating effect which

[4] H. W. Siegelman, in: *Biochemistry of phenolic compounds* (ed. J. B. Harborne), Academic Press, London 1964, page 437.

[5] J. Koukol and E. E. Conn, *J. biol. Chem.* **236**, 2692, 1961.

[6] G. Engelsma, in press.

[7] For a similar study on potato tuber slices, in which a correlation was also found between the induction of the enzyme phenylalanine deaminase and the synthesis of phenols, see M. Zucker, *Plant Physiol.* **40**, 779, 1965.

[8] P. M. Nair and L. C. Vining, *Phytochem.* **4**, 161, 1965.

[9] G. Engelsma, *Acta bot. neerl.* **15**, 394, 1966 (No. 2); *Nature* **208**, 1117, 1965.

[10] K. H. Erismann, *Int. Z. Vitaminf.* **32**, 36, 1962.

light exercises on the conversion of phenylalanine and cinnamic acid into *p*-coumaric acid in hypocotyl segments. Plants that have been sprayed with cycloheximide before being exposed to light produce much less phenylalanine deaminase and also synthesize a much smaller quantity of hydroxycinnamic acids.

We have touched here upon an important similarity to other photomorphogenetic phenomena. With the aid of the above-mentioned compounds and others that inhibit enzyme synthesis, it has also been established that *de novo* enzyme synthesis is an essential link in the mechanism underlying other processes governed by light, such as flower-induction, leaf development, chloroplast formation and anthocyanin synthesis. Which enzymes are involved in these phenomena is still an open question.

Returning now to the gherkin seedling, we know that when this plant is exposed to light the enzymes phenylalanine deaminase and cinnamic acid hydroxylase are synthesized *de novo* in the hypocotyl. Very little can yet be said, however, about the manner in which the synthesis is induced. There are indications that a factor is involved which, under the influence of light, is synthesized in the *cotyledons* and is then transported to the hypocotyl. If we remove the cotyledons before illumination, or cover them with aluminium foil so that only the hypocotyl is illuminated, the result is a substantial decrease in the stimulation of the formation of both enzymes and hence of the phenol synthesis. In blue light the decrease is about 50%, in red and far-red light it is 90% or more. If we illuminate intact plants for a short time (e.g. 10 minutes) and we wait some time before removing the cotyledons, we find optimum cinnamic acid hydroxylase activity when the time interval between the beginning of the exposure and the removal of the cotyledons is at least one hour. This would then be the time needed for the synthesis and transport of this factor.

Here again we come upon a similarity to other photomorphogenetic phenomena, such as flower-induction. In the latter case the leaves of the plant prove to be the light-perceiving organs, from which factors that induce or inhibit flowering are transported to the apices.

So far we have only considered the regulating mechanisms which relate to the enzymes involved in bringing about the response. In addition to this there is a regulatory system which is based on particular pigment transformations by light, as we shall see in the next section. A first indication that various types of light affect phenol synthesis in different ways has just been noted in connection with the effect of light when the cotyledons are removed or covered.

Pigment systems in phenol synthesis

Many photomorphogenetic processes in higher plants, including elongation and phenol synthesis in the gherkin, show an effect known as red-far-red antagonism. This effect can best be demonstrated for seed germination. With some varieties of lettuce, for example, only a small percentage of the seeds germinate in the dark. Brief illumination with red light is sufficient to bring the percentage of germinating seeds up to nearly 100. If, immediately after this irradiation, the seeds are exposed to far-red light, the percentage which germinate remains about as small as in the dark. Renewed irradiation with red light results in almost 100% germination once more, this is again antagonized by far-red illumination, and so on. As long as the time interval between the illumination periods is not too long, the sort of light that came last always determines the result. This effect has also been extensively studied on other light-dependent processes, such as anthocyanin synthesis, chlorophyll synthesis, opening of the plumular hook and flower-induction. Action spectra invariably show a maximum at 660 nm for the active red light and a maximum at 730 nm for the antagonizing far red.

From various plants that show this red-far-red antagonism a pigment has been isolated, called *phytochrome*, which occurs in two forms, one of which has an absorption maximum at 660 nm, the other at 730 nm (fig. 5) [11]. Illumination of a solution of phytochrome

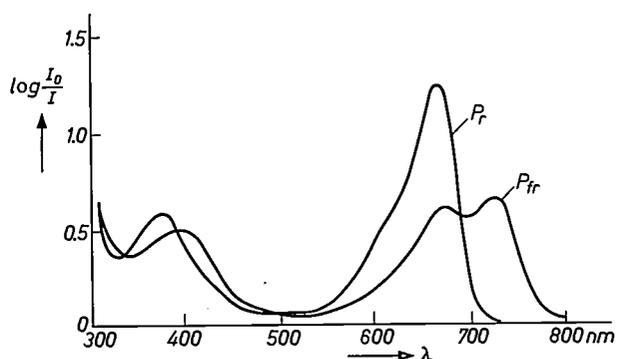


Fig. 5. Absorption spectra of the two forms of the phytochrome isolated from oat seedlings. Curve P_r gives the spectrum of the form which absorbs in red light, curve P_{fr} the spectrum of the one which absorbs in the far red. (Taken from Hendricks and Borthwick [11].)

at these wavelengths results in the conversion of the one form into the other. On the basis of absorption changes it is assumed that this photoreversibility also exists *in vivo*. Red light evidently converts a non-active form of the phytochrome into an active form, and far-red light does the reverse. As long as a plant has not been exposed to light the phytochrome is found entire-

ly in its non-active form with the absorption maximum at 660 nm. The active form, absorbing at 730 nm, slowly disappears in the dark *in vivo* in a manner as yet unknown. This agrees with the fact that red light is antagonized less the longer the time allowed to elapse before far-red illumination. Just how the active form of phytochrome (a protein of high molecular weight) functions, is as yet not known. Enzyme action is considered to be a likely possibility.

It is to be assumed that in addition to phytochrome another pigment, absorbing more especially in the blue part of the spectrum, is involved in phenol synthesis. This is seen if one compares an action spectrum for the induction of cinnamic acid hydroxylase activity (curve *act* in fig. 6) with the absorption spectra of both forms of phytochrome (fig. 5). Where the latter show a minimum, the action spectrum on the contrary shows a maximum (at 460 nm).

The fact that different pigment systems are involved in phenol synthesis is reflected in a difference in dependence on exposure time and light intensity for light of different wavelengths^[3]. An example is to be seen in fig. 7, where the phenol synthesis is shown as a function of exposure time for red (curve *r*) and blue light (curve *b*). In red light saturation is reached more quickly than in blue light, and the maximum level of the response remains much lower in red light than in blue. Unlike the effect of blue light, the red light effect is governed by the reciprocity law: effect = light intensity \times exposure time. We have just considered the particular antagonism between red and far-red light, and have also pointed to a difference between blue and red light in regard to the extent to which perception by the cotyledons is essential to the stimulation of phenol synthesis in the hypocotyl.

Some of these points indicate that red and blue light are perceived not only through different pigment systems, but also that these systems function differently. Red-far-red antagonism, a low saturation level and the validity of the reciprocity law can be brought into a certain relationship if we start from the premise that the action of red light depends on the conversion of an inactive form of phytochrome into an active one — without going into the question of what this activity is. The way in which the effect in blue light depends on exposure time and light intensity, on the other hand, lends support to the assumption that the effect of blue light is primarily based on the *excitation* of a pigment which exercises a particular effect from the excited state. By analogy with the primary process of photosynthesis this might for instance give rise to an electron transport. According to this assumption such an excitation could trigger a chain of reactions which would eventually lead to the speeding up of the synthe-

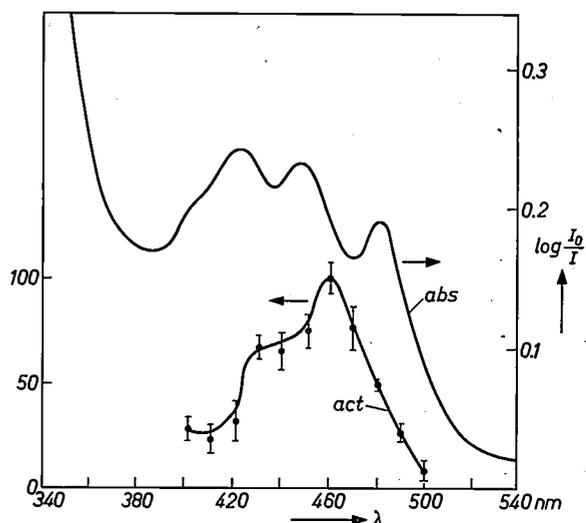


Fig. 6. Action spectrum for the induction of cinnamic acid hydroxylase activity in segments of gherkin hypocotyls (curve *act*) and absorption spectrum of the same segments (curve *abs*).

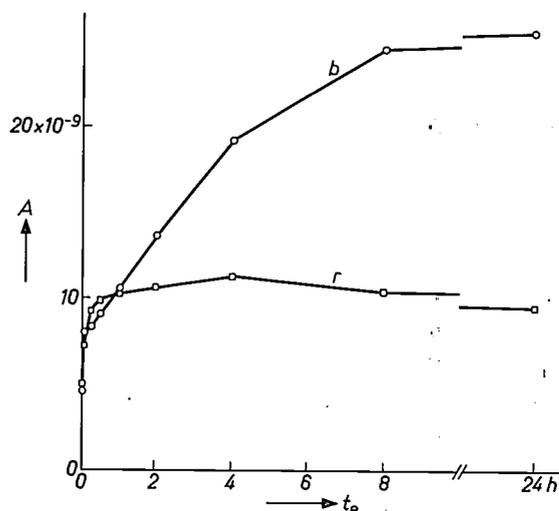


Fig. 7. The quantity *A* produced during 24 hours of hydroxycinnamic acids per hypocotyl as a function of the time of exposure to red light (curve *r*) and to blue light (curve *b*) in dark-grown gherkin seedlings three days old. The quantity of hydroxycinnamic acids is given in moles. Light intensity 800 $\mu\text{W}/\text{cm}^2$.

sis of phenylalanine deaminase and of cinnamic acid hydroxylase.

The effect of *far-red light* on phenol synthesis involves a complication which we have hitherto disregarded. On the one hand a small amount of far-red light, given after exposure to red light, has an antagonizing effect, which can be explained in terms of the conversion of

[11] For two recent survey articles dealing with the physiological significance of phytochrome and with its isolation, see: S. B. Hendricks and H. A. Borthwick, in: *Chemistry and biochemistry of plant pigments* (ed. T. W. Goodwin), Academic Press, London 1965, page 405; W. L. Butler, S. B. Hendricks and H. W. Siegelman, *id.* page 197. See also: H. Mohr, *Photochem. Photobiol.* 5, 469, 1966 (No. 6).

the active form of phytochrome into the inactive form. On the other hand, prolonged far-red irradiation stimulates phenol synthesis, and the effect then is even greater than that of red light. The dependence on exposure time and light intensity is similar to that for blue light, but as regards the importance of the cotyledons in effects due to light, the far red corresponds to red light. There are two alternative explanations for these effects: either the far red is perceived by a separate pigment which is distributed over hypocotyl and cotyledons in roughly the same way as phytochrome, or phytochrome is the percipient pigment but functions in much the same way when absorbing far-red light as the pigment responsible for the effect of blue light.

Summarizing, we can say that the induction of phenol synthesis by light involves at least one other pigment in addition to phytochrome. This other pigment, with a maximum around 460 nm according to the action spectrum in fig. 6 (curve *act*) must be sought among other pigments, as evidenced by the difference between this action spectrum and the absorption spectrum of the gherkin hypocotyl, given in the same figure (curve *abs*).

Phenol synthesis and elongation

The elongation of a plant is determined by processes of cell division and cell expansion, both of which are affected by light. At the time when our investigation into the elongation of the gherkin hypocotyl begins, 72 hours after sowing, a stage has been reached where growth is largely governed by cell expansion. We shall therefore confine our attention to this phenomenon.

The cell wall is built up from the protoplasm under the action of enzymatic processes. It consists of cellulose fibres which are held together by an amorphous mass (matrix) of polysaccharides of complicated composition. In the opinion of some investigators the cell wall also contains structural protein.

The cell wall has elastic and plastic properties. The wall is subjected from the inside to a pressure that depends on the difference in concentration between the molecules present in solution inside and outside the cell (cell osmosis). It is generally accepted that changes in the rate of cell expansion are the result of changes in the plasticity of the cell wall, due for example to the breaking or forming of certain bonds in the matrix, and not in the first instance to changes of pressure within the cell. Since the cellulose fibres are laid down by the protoplasm in a particular direction in the cell wall, the plastic and elastic properties of the cell wall are not the same in all directions (anisotropy). This explains why a cell subjected to internal pressure expands mainly in one particular direction^[12]. An increasing plasticity of the cell wall is usually accompanied by accelerated

incorporation of new matrix material, generally followed by an increased synthesis of cellulose.

One of the factors involved in the regulation of plant elongation is the growth hormone indoleacetic acid (IAA). It is assumed that the concentration of IAA is in turn determined by the activity of the enzyme that breaks down IAA, the IAA oxidase^[13]. Monophenols activate this enzyme, while o-diphenols have an inhibiting effect. The first group of compounds will thus tend to reduce the IAA concentration in the cell by accelerating the breakdown, while the second will show an action more sparing with IAA. In a number of plants it has been found, in complete agreement with this, that monophenols, including *p*-coumaric acid, inhibit elongation, whereas o-diphenols and other closely related compounds, such as ferulic acid, tend to promote growth^[14].

We have here an important link with the subject discussed in the previous section, the stimulation of the synthesis of hydroxycinnamic acids by light. Under the influence of light a shift is brought about in the balance between growth-inhibiting and growth-promoting phenols.

When a dark-grown seedling is brought into the light, it begins in the first place to produce a large amount of *p*-coumaric acid which, as we have seen, has the effect of inhibiting growth. We notice in fact that the rate of elongation declines. In a later stage more ferulic acid is made, which has the opposite effect. This corresponds to the finding that growth inhibition as a rule passes through a maximum. However, although other experimental data fit the hypothesis outlined here, for example the fact that the concentration of growth hormone is lowered in the hypocotyl when the plant is exposed to light, this mechanism *alone* does not provide a complete explanation for the inhibiting effect of light on elongation. The causal relationship assumed here between phenol synthesis and change of elongation implies that the latter should show at least as long a lag as does phenol synthesis. This, however, is not the case — at least not under illumination with blue light, where growth is inhibited within the very first minute (*fig. 8*). In red and far-red light the lag of growth inhibition varies for individual plants from about 30 minutes to longer than two and a half hours. This results in an average lag of the same magnitude as the lag in phenol synthesis. As gherkin seedlings appear to respond very rapidly to changes in IAA concentration there is good reason to assume that enhanced phenol synthesis is the primary cause of the growth inhibition in red and far-red light, although it is difficult to draw any definite conclusions^[15].

We might now consider whether there are other ways in which light could affect the growth of the plant.

Just as in phenol synthesis, *de novo* protein or enzyme synthesis also plays an important part in plant elongation. This appears from the growth-retarding effect of inhibitors of protein synthesis, such as actinomycin D and cycloheximide. In this connection one might think of the synthesis of structural protein for the cell wall or of the synthesis of enzymes responsible for producing other components, such as cellulose. The effects of light might thus arise from an effect upon the rate at which this *de novo* protein synthesis takes place. This, however, brings us up against the same difficulty as before. The lag to be expected with this mechanism might well explain the growth inhibition in red light and in the far red, but not the inhibition in blue light.

The very rapid effect of blue light on the growth rate might lead one to suppose that blue light affects a

an analogous process in yeast, namely budding due to local softening of the cell wall [17]. It has been demonstrated that this is caused by reduction of the disulphide bonds in a polysaccharide-protein complex by a specific enzyme, protein-disulphide-reductase, which depends for its action on a reduced cofactor. Light might be able to oxidize the cofactor, thereby regulating the enzyme activity and thus inhibiting elongation. A mechanism of this kind would explain the rapid effect of light on elongation, as found upon illumination with blue light.

Summarizing, we may conclude that the effects of red and blue light are brought about here as well via different pigment systems and presumably also via different mechanisms.

Concluding remarks

In a living organism everything is more or less bound up with everything else. High-molecular compounds, such as DNA, RNA and enzymes, are built up in close dependence upon one another. These compounds also determine which low-molecular compounds are to be made. Some of them influence in their turn the relation between the high-molecular compounds and determine, often again in co-operation with these compounds, the speed at which other metabolic processes take place. Outside intervention can have far-reaching effects on the system. Exposure to light is such an intervention.

A method that can be used to trace the regulating mechanisms of the plant, is to investigate the way in which the metabolism reacts to a light stimulus. For this reason we begin our investigations by looking for the start and finish of the photomorphogenetic process.

Starting points are formed by the pigments that perceive the active light. End points are found on one hand in the form of certain structures whose function already implies some degree of unchangeability, such as the cell wall, and on the other hand in the form of certain compounds which are not, or at least not very quickly, further metabolized, like some of the phenolic compounds, which probably accumulate in the vacuoles.

The differences in the way in which light of different

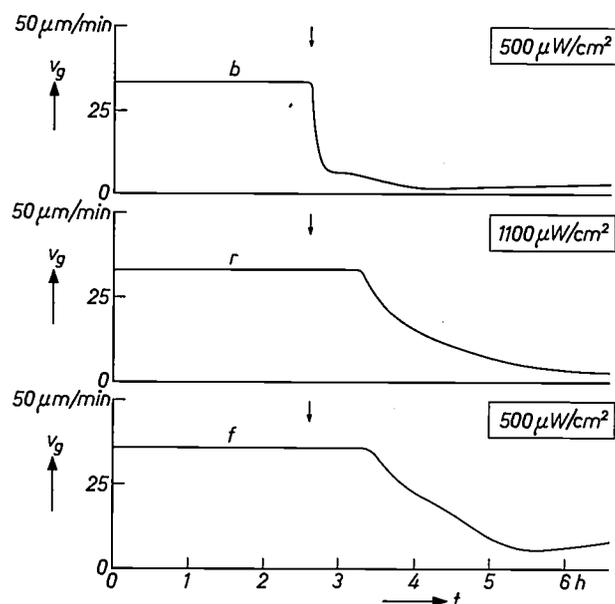


Fig. 8. Growth rate v_g of dark-grown gherkin seedlings three days old which were illuminated with blue (curve *b*), red (curve *r*) and far-red light (curve *f*) at the times indicated by an arrow. (According to investigations by Dr. G. Meijer of this Laboratory.)

process that is directly involved in the changes of plasticity of the cell wall matrix. An attractive hypothesis in this connection is that the changes of plasticity are caused by the opening and closing of disulphide bonds by controlled reduction and oxidation. This would tie up with what we mentioned as a possibility in the previous section, i.e. that the primary action of blue light consists in bringing about a particular transport of electrons, that is to say an oxidation-reduction reaction.

In this hypothesis the disulphide bonds to be opened and closed would form part of the structural protein of the cell wall [16]. This mechanism would correspond to

[12] P. A. Roelofsen, in: *Advances in botanical research* (ed. R. D. Preston), Academic Press, London 1965, page 69.

[13] W. S. Hillman and A. W. Galston, *Plant Physiol.* **32**, 129, 1957.

[14] J. P. Nitsch and C. Nitsch, *Ann. Physiol. vég.* **4** (3), 211, 1962.

[15] The relation between the effects of light on phenol synthesis and the development of the seedling has also been investigated for pea seedlings. See: M. Furuya and R. G. Thomas, *Plant Physiol.* **39**, 634, 1964, and W. Bottomley, H. Smith and A. W. Galston, *Phytochem.* **5**, 117, 1966.

[16] D. T. A. Lampert, in: *Advances in botanical research* (ed. R. D. Preston), Academic Press, London 1965, page 151.

[17] W. J. Nickerson, *Bacteriol. Rev.* **27**, 305, 1963.

wavelength regions stimulates phenol synthesis are reflected in various other photomorphogenetic phenomena. This makes it reasonable to conclude that a fairly high degree of similarity must exist between the pigment systems involved.

We can state that in the attainment of any photomorphogenetic response, some part is always played by enzymes. In light-induced phenol synthesis we found that the formation of the synthetic pathway enzymes was stimulated. There are indications that a number of other photomorphogenetic phenomena are also controlled by similar regulating mechanisms.

There are other links that can be established. For example, the rapid inhibition of growth in blue light has been brought into relation with a particular oxidation-reduction process. We are also able to state that both enzymes involved in the phenol synthesis depend for their action on certain reduced cofactors. We mentioned in the introduction that light stimulates the synthesis of ascorbic acid, a compound which is thought to play an important part in the oxidation-reduction equilibrium in the cell. We suspect certain relations here, but these are still only speculations.

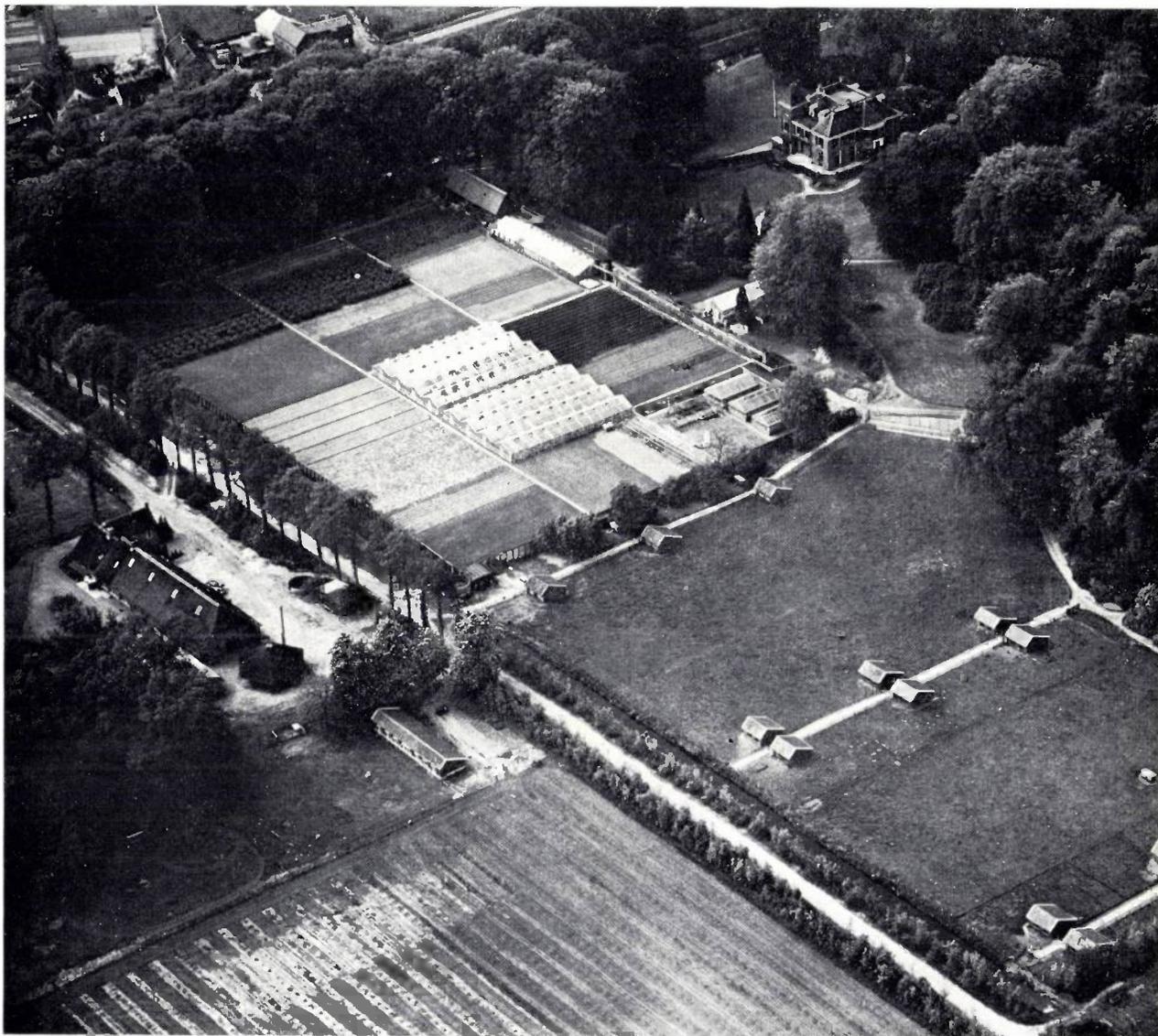
Finally, we would like to point out that phenol synthesis in the hypocotyl of the gherkin is determined to a large extent by a factor synthesized in the cotyledons as a result of exposure to light. This has brought us to a point which we had already reached in our phenomenological considerations at the beginning, namely that the photomorphogenetic phenomena can only to

a partial extent be studied as isolated effects because of their inherent interaction. To understand what takes place in the hypocotyl we are bound to take into account the changes induced by light in the cotyledons. During the same period in which light inhibits the elongation of the hypocotyl and causes hydroxycinnamic acids to accumulate in it, the cotyledons begin to unfold, the storage products in them to be broken down, chloroplasts are formed, while ascorbic acids and a variety of phenols are synthesized. And this is just a part of the jigsaw puzzle which, in our article, we have tried to begin to put together.

Summary. This article deals with the stimulating effect of light on the synthesis of the phenols *p*-coumaric acid and ferulic acid in the hypocotyl of gherkin seedlings. It is shown that this takes place as a result of the accelerated production, from the amino acids, of the enzymes phenylalanine deaminase and cinnamic acid hydroxylase which are involved in the synthesis of these phenols. It is also shown that at least two pigment systems — phytochrome, which is sensitive to red and far-red light, and a special pigment absorbing in the blue spectral region — are involved in the light-induced phenol synthesis. Special attention is given to a possible relationship between the stimulation of phenol synthesis and the inhibiting effect which light has on the elongation of gherkin seedlings, in connection with the fact that phenols are involved in regulating the concentration of the growth hormone indoleacetic acid.

The main purpose of the investigation is to get an insight into the regulating mechanisms of the plant. Conversion of phytochrome by light into a physiologically active or inactive form, induction and repression of the synthesis of the enzymes phenylalanine deaminase and cinnamic acid hydroxylase, and activation and inactivation of the enzyme that breaks down the growth hormone indolylacetic acid, form various examples of biochemical control mechanisms, which are brought out when considering the photomorphogenetic phenomena described here.

“Boekesteyn” Agrobiological Laboratory



Aerial photograph of “Boekesteyn” at ‘s-Graveland in the Netherlands, an estate dating from the 17th century, which is equipped as an agrobiological laboratory for N.V. Philips-Duphar. The grounds belonging to the estate are used for small-scale field tests, and there are hothouses in which plants of many kinds are grown for the investigations carried out at Boekesteyn [1].

In the Boekesteyn laboratory the chemical compounds synthesized in the Weesp Research Laboratories are studied to find out their action against mites, insects, fungi, weeds and nematoda. Investigations are also made of the biochemical and physico-chemical behaviour of these potential pesticides, both in and on

plants, in the soil and in isolated enzyme systems.

Differences in climatic and other conditions from one country to another make it necessary to subject a pesticide to comprehensive field tests in the actual country where it is to be used.

After compounds found to be effective have been tested in the fields and orchards at Boekesteyn, a whole programme of field tests on a larger scale is then carried out in the Netherlands and in various other parts of the world.

[1] See R. van der Veen, Philips tech. Rev. **16**, 353, 1954/55.

The atomization of concentrated dispersions of pesticides

W. Duyfjes

The discovery of the insecticidal action of DDT in 1939 triggered off large-scale investigations in a search for new chemical agents to control the many pests and diseases found in agriculture and horticulture. Among the hundreds of control agents resulting from these investigations, a few dozen have been of notable value in increasing the yield per acre of various crops.

Generally speaking, only small quantities of the chemical agents or pesticides (a term covering insecticides, acaricides, fungicides and herbicides) are needed to achieve the desired effects; say from 100 grams to a few kilograms per hectare. The costs of the actual active substances are therefore relatively low; the distribution of these substances over the crops presents so many problems however that the use of pesticides in agriculture and horticulture is nevertheless a cost factor to be reckoned with. The active substance cannot as a rule be applied straight to the plants, but must first be made up, by the addition of various adjuvants, in a form in which it can readily be distributed and in which the pesticidal action is most effective. This procedure is known as the *formulation* of the pesticide [1].

A pesticide can be spread over crops in various ways: by spraying or atomizing a liquid, by dusting with a dry powder, and in some situations by means of smoke. Spraying and atomizing are the most widely employed methods and differ mainly in the quantity of liquid distributed per acre. They are referred to specifically as high-volume, low-volume and ultra-low-volume spraying. The term high-volume spraying is generally used when the quantity distributed is more than, say, 500 l/ha (about 50 gallons imp./acre), atomizing, or ultra-low-volume spraying, relates to quantities of about 1 to 5 l/ha, and low-volume spraying refers to intermediate quantities. The liquid distributed is usually a disperse system (an emulsion or suspension), as most pesticides are not soluble in water.

A serious drawback of high-volume spraying is that with conventional spraying apparatus, the spraying liquid has to be very highly diluted to give a uniform distribution of the active substance over the crop. In many cases it is necessary to use at least 500 litres of diluting agent (usually water) per hectare. This means that if a large area is to be sprayed by this method, the tank of the spraying apparatus has to be continually re-

filled, and this obviously puts up the costs considerably, particularly when spraying from aircraft. It now proves possible to reduce the amount of liquid required per hectare very considerably, sometimes even to a tenth or less of the original quantity. The distribution of these small amounts of concentrated dispersions imposes strict requirements on the spraying and atomizing equipment and on the formulation.

In atomization from the air it is also important to ensure that the "deposition" is adequate, in other words that the droplets reach the ground and are not carried away by the wind. When herbicides are being spread, this "spraydrift" can have disastrous consequences for adjoining plots of land where crops are growing. Deposition can be increased by using special methods to ensure that no unduly small droplets are produced, and by adding substances which inhibit evaporation, so that large droplets are not reduced in size to such an extent that they can be blown away. With concentrated dispersions it is moreover particularly difficult to maintain stability, so that emulsions do not prematurely break down (i.e. separate into oil and water) and suspensions do not flocculate.

In spraying from aircraft there is a growing need to distribute several preparations at the same time. This may well give rise to difficulties: when mixing an emulsion and a suspension the two dispersions are often found to be "incompatible". The mixture may then flocculate and form a paste which blocks the spray system.

These problems will be discussed in this article. We shall first deal briefly with emulsions and suspensions. We shall then describe the "invert emulsion" (water-in-oil emulsion), a dispersion with useful properties which may well provide an answer to a number of problems.

Emulsions

Pesticides that can be spread as emulsions are generally marketed as "emulsifiable concentrates". They consist of an oil which emulsifies spontaneously on mixing with water, forming an oil-in-water emulsion. Sometimes the oil itself is the active substance, but usually the active substance is dissolved in an aromatic hydrocarbon oil. An emulsifiable concentrate also contains an emulsifier, i.e. a substance which enables the oil to emulsify in water and which at the same time stabilizes the emulsion. The emulsifiers originally used were "soaps"; and since a relatively high percentage of such soaps was needed to ensure good emulsifiability,

Ir. W. Duyfjes is with N.V. Philips-Duphar, Development Laboratories, Amsterdam.

emulsifiable concentrates used to be expensive. The emulsifiers available nowadays however are far more effective and better able to resist hard water. A much lower percentage of these emulsifiers now suffices.

These new emulsifiers are mixtures of two types: hydrophilic emulsifiers (i.e. with an affinity for water) and lipophilic emulsifiers (with an affinity for oils). The effectiveness of these mixtures is due to a complex process of selective adsorption at the oil-water interface: molecules of the two types of emulsifier becoming, so to speak, "anchored" to each other, forming a firm and stable interfacial film. The hydrophilic emulsifiers used are polyglycol ethers of alkylated phenols or partial esters of sorbitol, and the lipophilic substances are alkylbenzene sulphonates, e.g. calcium dodecylbenzene sulphonate. Given well-balanced mixtures of these substances a content of 3 to 5% in the emulsifiable concentrate is sufficient to emulsify the concentrate spontaneously into an oil-in-water emulsion.

These emulsifiers are more effective when the oil is very highly diluted in water (e.g. dilutions of 1/100 or 1/1000). They are not so suitable for the concentrated oil-in-water emulsions that are nowadays preferred^[2] with concentrations of 1/10, 1/5 or even higher, as in these emulsions there is no optimum orientation of the emulsifier molecules at the oil-water interface. Because of this, concentrated emulsions sometimes break down within a few minutes, with a total separation of the oil and water phases. A frequent occurrence at these high concentrations is the formation of a "multiple emulsion", i.e. one in which the dispersed oil phase itself contains a dispersed water phase (*fig. 1*). Such complex emulsions can be sprayed in the normal way. In certain cases however, a total phase reversal may occur, the oil-in-water emulsion being converted to a water-in-oil emulsion. This type of emulsion, which will be discussed in more detail at the end of this article, can be extremely viscous, particularly if the oil-to-water ratio is about 1 : 3 or 1 : 4. The viscosity increases further if the emulsion is agitated; and since the spraying of a concentrated fluid almost invariably involves vigorous agitation in the spraying apparatus, the emulsion thickens to such an extent that it can no longer be sprayed.

These effects are due to the lipophilic components in the emulsifier mixture, as these promote the formation of water-in-oil emulsions. In order to atomize highly concentrated oil-in-water emulsions a mixture is obviously needed that only contains hydrophilic emulsifiers, which facilitate the formation of oil-in-water emulsions. Such emulsifiers are only effective however if their content is high, and this makes them expensive to use. This kind of activity by lipophilic substances on concentrated emulsions leads to the cause of the diffi-

culties that arise when these emulsions are mixed with suspensions.

Suspensions

Pesticides sprayed in the form of a suspension are usually marketed as a powder which has to be mixed with water. These preparations, known as "wetable powders", consist of a solid carrier which is intimately mixed with the active substance and various adjuvants (this can be done by milling the ingredients together). Natural or synthetic silicates or colloidal silicic acid can be used as carriers. The adjuvants contained in wettable powders are a "wetting agent" and a "dispersing agent". A wetting agent is a substance

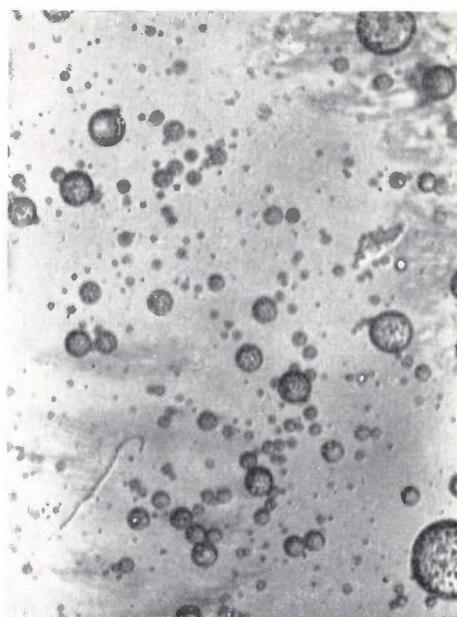


Fig. 1. Photomicrograph of a multiple emulsion. The emulsified oil droplets of the oil-in-water emulsion themselves contain small emulsified water droplets. Magnification 230 \times .

that lowers the surface tension of water sufficiently to enable the powder to be wetted by water and to form a suspension when mixed with water. The dispersing agent prevents flocculation.

The quality of wettable powders has reached a high standard in recent years. This is due to some extent to the considerable progress made in milling techniques, and also to the use of the carriers referred to above, which have been found to be very effective. The use of these carriers has also made it possible to employ a much greater proportion of active substance in the various preparations.

[1] W. Duyfjes, The formulation of pesticides, Philips tech. Rev. **19**, 165-176, 1957/58.

[2] W. Duyfjes, A breaking phenomenon of an oil/water emulsion, The formulation of pesticides, S.C.I. Monograph No. 21, 89-96, 1966.

It is important to follow the directions when mixing a wettable powder with water. The powder usually has to be mixed with a little water first to form a paste; the surfactants then dissolve quickly in the water and form a concentrated solution whose surface tension is low enough to ensure rapid wetting and in particular practically complete adsorption of the wetting and dispersing agents on the solid particles. After dilution of the paste with water a suspension is obtained that will remain stable for hours, even at dilutions of 1/100 or 1/1000. If however the powder is directly mixed with a

highly stable, at least when the powder is first mixed with a little water.

The disadvantage of a high surfactant content is the excessive frothing that occurs when the concentrated suspension is agitated in the spray apparatus. Because of the frothing, the tank can only be partly filled, otherwise it overflows. This frothing can in turn be effectively prevented by the addition of defoaming agents, which are preparations based on hydrocarbon oils or silicones. If defoaming agents are used, however, there is some danger of their becoming emulsified in the sus-

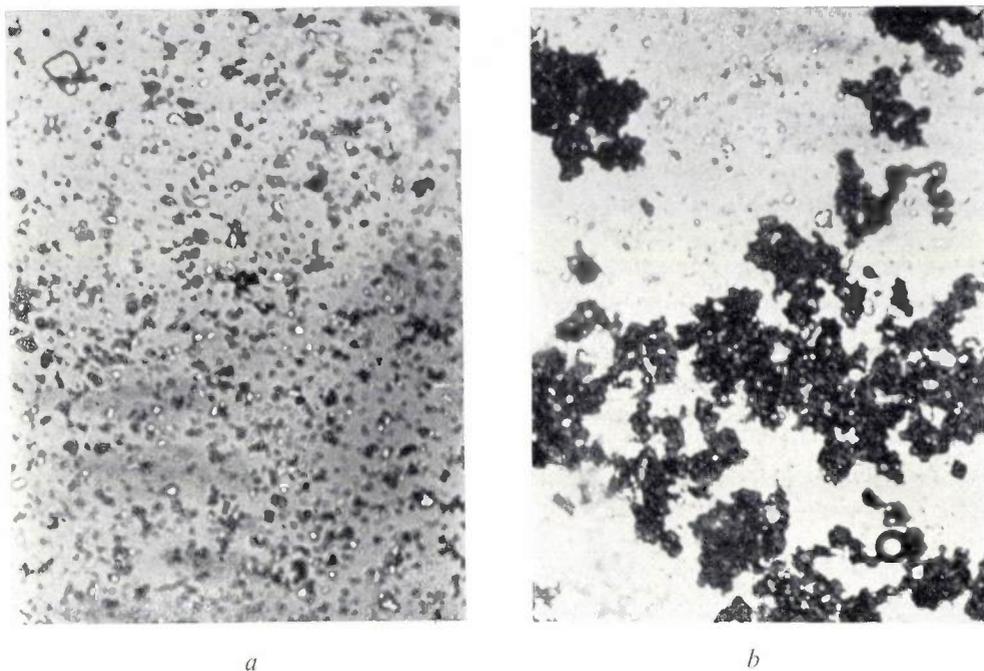


Fig. 2. Photomicrographs of a suspension of DDT. Magnification 110 \times .

- a) Without addition of a defoaming agent the suspension is very stable.
 b) After the addition of a defoaming agent the suspension is markedly flocculated. Owing to turbulence in the spray apparatus this agent has emulsified in the suspension and made it unstable.

large quantity of water, the adsorption of the surfactants on to the solid particles may not be sufficient to keep the suspension stable, and after a few minutes flocculation occurs.

It might be thought that a smaller quantity of surfactants would be needed in wettable powders formulated as *concentrated* suspensions for atomization. The carriers nowadays employed are however so finely divided (these powders sometimes have a total surface area of a few hundred square metres/gram) that they absorb a considerable portion of the adjuvants. This has the result that even at high concentrations, a fairly high content (a few per cent) of wetting and dispersing agents is required. Unlike the concentrated emulsions, the concentrated suspensions produced in this way are

pension as a result of vigorous agitation. Although originally highly stable, the suspensions may then suddenly start to flocculate (*fig. 2a* and *b*). Defoaming agents should for this reason only be used in very small amounts. It will be clear from all this that wettable powders for concentrated suspensions require very careful formulation.

The mixing of concentrated emulsions and suspensions

In crop-spraying practice it is more and more frequently required to apply two different pesticides at the same time, and it is then convenient to be able to mix the two preparations in the spray tank. If an emulsion and a suspension are to be used difficulties may arise due to breakdown and flocculation, especially if

the suspension contains strongly lipophilic substances. These problems are particularly acute if it is necessary to atomize mixtures of concentrated dispersions. We have already seen that concentrated emulsions are in general rather unstable, owing to the disorganized state of the emulsifier molecules adsorbed at the interface of the dispersed oil phase. The concentrated suspensions prepared from wettable powders are relatively stable, provided they are treated in the right way. If these dispersions are mixed, however, the state of equilibrium in both liquids is upset and their stability is destroyed,

with sufficient particles to give the maximum contact with the fungus spores.

In this situation, already complicated enough, a fatal complication may arise if the active ingredient of the suspension is soluble to some extent in the oil of the emulsion — whether or not in combination with an adjuvant. The solid particles of the suspension may then migrate to the emulsified oil, accumulate there and partly or completely dissolve in the oil. This can lead to a distortion of the interfacial film of the emulsion, resulting in greater instability. Such an effect is found

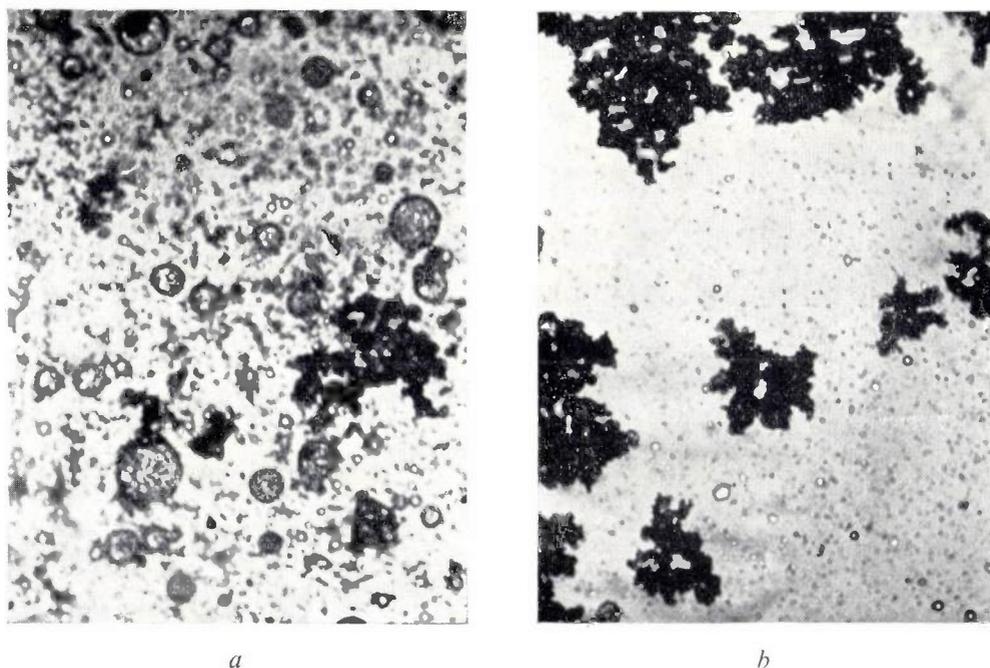


Fig. 3. Photomicrographs of a mixture of an emulsion of "Malathion" with a suspension of thiram. The stability of both dispersions is upset because the thiram tends to dissolve in the oil of the emulsion. Magnification 110 \times .

- a) The situation shortly after mixing. The emulsion starts to coalesce and the suspension starts to flocculate.
 b) In the final situation the suspension is totally flocculated into large pieces in which most of the oil has been taken up. Only a little of the oil is still emulsified.

so that after some time breakdown and flocculation occurs. If the suspension is incompletely stabilized because the powder has not been properly stirred into the emulsion, or, worse still, has been tipped straight into the emulsion, the mixture flocculates very quickly.

Mild flocculation does not always mean that the dispersion can no longer be sprayed; it can quite often still be sprayed provided the mixture is stirred while spraying. It does however as a rule tend to decrease the activity of the atomized droplets on the object. This can be a serious disadvantage with certain fungicides, which are usually sprayed or atomized in the form of suspensions. For these preparations to be effective, it is even more essential than with the other pesticides that the surface to be protected should be covered

when mixing an emulsion of "Malathion" [*] with a suspension of thiram (tetramethyl thiuram disulphide). Thiram is highly lipophilic and insoluble in water. This mixture would be useful in fruit growing for single-spray control of aphids and spider mites (with "Malathion") and spores of injurious fungi (with thiram). This particular mixture is a clear example of the kind of incompatibility described (*fig. 3a and b*): the suspension flocculates almost completely and the emulsion shows premature coalescence of the dispersed phase, leading to breakdown.

There are various ways in which mixtures normally incompatible because of certain properties of the active

[*] "Malathion" is a product of American Cyanamid Company, Wayne, N.J., U.S.A.

ingredients, or of the carriers, can be made compatible. This can usually be done by modifying the composition of one of the preparations, generally the emulsifier composition. If this is not an economic proposition, the two dispersions can sometimes be given sufficient stability by the addition of a colloid with a very specific action. *Fig. 4* gives an example of this action: a mixture of an emulsion of DDT and a suspension of zineb (a fungicide) is highly incompatible (*fig. 4a*); in the same mixture the emulsion and the suspension remains very finely dispersed if a colloid is added (*fig. 4b*). The use of a

principal characteristic of the spraying technique here is the pressure at which the liquid is atomized, especially if highly viscous solutions or dispersions are used.

Obviously, when preparations are sprayed from the air the droplet distribution can be affected to a very considerable extent by weather conditions before the droplets have settled. As we mentioned above, their slow rate of descent and the strong turbulence make the droplets very subject to evaporation, so that they gradually get smaller and are sometimes carried miles away by the wind. To reduce this spray drift two measures

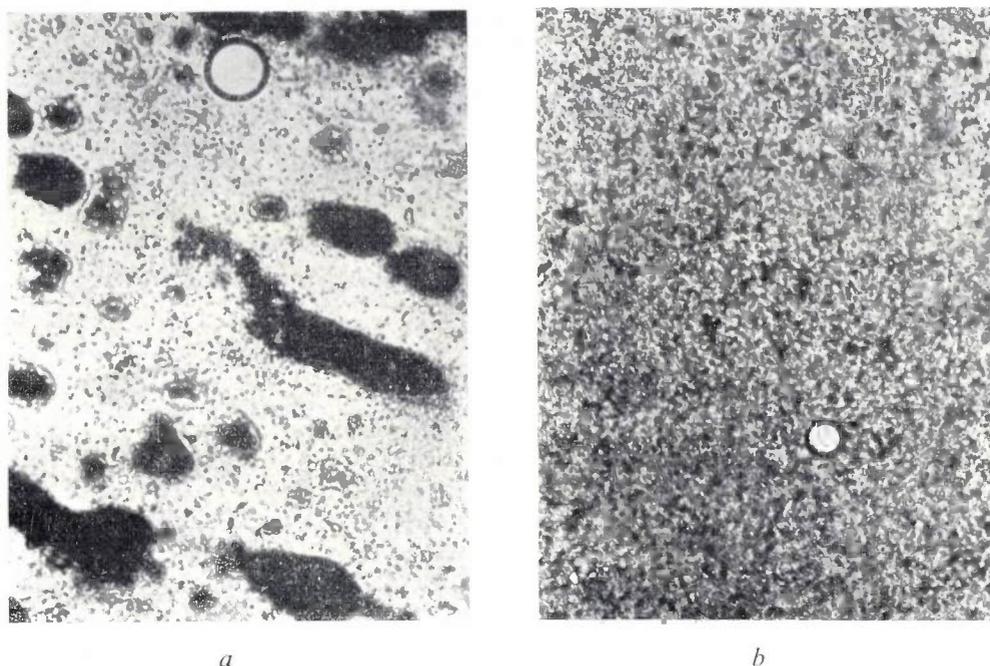


Fig. 4. Photomicrographs of a mixture of an emulsion of DDT with a suspension of zineb. Magnification 110 \times .

- a)* The dispersions are incompatible unless a colloid is added. There is a very marked coalescence of the emulsion and the suspension is almost totally flocculated. The clotted solid particles of the suspension are enclosed by the oil of the emulsion. (The circle at the top of the photograph is an air bubble.)
- b)* If a colloid is added the mixture is very stable; both oil and the solid particles remain very finely dispersed. (The circle is an air bubble.)

colloid can however reduce the effectiveness of one or even of both active ingredients of the mixture. In particular, strongly hydrophilic colloids may hinder the availability of the active substance at the object, owing to the formation of a film around the active substance once the droplets have dried.

Some physical aspects of the sprayed droplets

In every spraying technique the sprayed droplets have a characteristic size distribution. This distribution depends on various physical properties of the liquid, in particular on the viscosity and surface tension. The

are required. In the first place, spraying techniques have to be devised in which no unduly small droplets are formed. Droplets with diameters as large as 50 microns are too small; generally speaking, these do not reach the object. Secondly, something must be done to prevent the evaporation of the descending water droplets. If no measures are taken to reduce evaporation, then droplets as large as 100 to 150 microns can no longer be deposited when sprayed from the air at a temperature, say, of 30 °C and at low air humidity. On the other hand of course, the droplets should not be too large. The line is generally drawn at a diameter of

about 500 microns; if they are larger than this, the number distributed on the object per unit area may in some cases be too small, thus reducing the effectiveness of the control agent. This limit obviously depends to a great extent on the active substance to be sprayed. For example, the application of 40 l/ha (about 4 gallons/acre) of droplets with a diameter of 500 microns gives a deposited density of six droplets per cm^2 . This may be sufficient if a systemic pesticide is being used, i.e. one which is taken up in the sap stream of the plant, but for a fungicide (these are not usually systemic) this is quite insufficient.

Recently spray nozzles have been devised which give droplets with a very narrow size distribution. These have already been successfully employed for atomizing undiluted "Malathion" using a dosage of only a litre per hectare. We should not however conclude from this that all the problems of atomizing concentrated liquids have been solved by this method. In fact, "Malathion" is much more viscous and also much less volatile than water, and moreover good droplet deposition is possible only if there is little wind. In practice such conditions cannot be determined beforehand, and it will usually be necessary to spray larger droplets, and therefore to use solutions or dispersions of the active substance.

Bearing in mind the evaporation of the droplets, we see that water is not in fact a very suitable diluting agent for solutions and dispersions. This is because it has a fairly high vapour pressure; for example a water droplet of 50 microns diameter, at an ambient temperature of about 20°C and relative air humidity of about 40% will evaporate within four seconds. At a diameter of 100 microns it will evaporate in 16 seconds (the rate of evaporation is roughly inversely proportional to the square of the diameter). In spite of this undesirable property, water is very widely used as a diluting agent because it is cheap and because it is well-known to be harmless. In recent years however, efforts have been made to reduce the evaporation of water droplets by the addition of certain chemical agents.

Agents for reducing evaporation

The evaporation of water can be reduced by the addition of an agent that covers the surface of the water with a dense molecular layer which shuts off the water from the air. Good results have been obtained with substances like cetyl alcohol which consist of a polar group (in this case a hydroxyl group) and a long apolar hydrocarbon chain. The vigorous spreading property of this substance over the surface of the water gives rise to a thin layer of densely packed molecules. In each molecule the hydroxyl group is oriented towards the surface of the water, while the hydrocarbon group is

turned away from it. This principle is employed for the "conservation" of still water, such as that in drinking water reservoirs, to prevent excessive evaporation.

On the same principle one might try to minimize the evaporation of sprayed droplets. It is however no easy matter to compound the evaporation inhibitor in the preparation in such a way that the droplets after spraying are rapidly and adequately covered with a dense layer of the inhibitor. Although such inhibiting agents can be released into the suspension or emulsion in a very finely divided form, a rapid accumulation at the surface of the water droplet is difficult to bring about. Some other substances which show the same effect, e.g. saturated fatty acids like stearic acid, cannot be used at all because they are not sufficiently soluble in water, or because they show insufficient spread on water.

Some years ago, Hartley carried out experiments with certain salts of stearic acid (in particular with methyl dibutylamine stearate) which are in fact very soluble in water^[3]. These preparations were introduced as evaporation-inhibiting agents intended to be used solely with suspensions of finely divided pesticides. No wetting agents or dispersing agents could be used because the stearates themselves are surfactants, and a surplus of such substances would reduce the effectiveness of the stearate. This action is based on the dissociation of the stearate into stearic acid and a volatile base. The base evaporates from the droplet, and as the stearic acid is not readily soluble in water it should accumulate at the surface of the droplets.

In experiments with a suspension of copper oxychloride reduced evaporation was indeed found, but whether this could really be ascribed to the effect described is still an open question. Doubt arises because it is known that surface films of solids on water are susceptible to "strip effects" caused by the wind, that is to say more or less uncovered local patches where the water can evaporate. This effect might well be expected to take place in spray droplets. Moreover, the released stearic acid would probably be kept in dispersion (solubilized) by the still intact stearate in the droplet, so that it could not directly reach the surface. The reason for the reduced evaporation could also be related to the viscosity increasing effect of the stearates: this reduces the mobility of the water molecules in the liquid and this results in less evaporation.

Agents for increasing viscosity

Spraydrift can be considerably reduced with the aid of certain agents that increase the viscosity of the spray

^[3] G. S. Hartley and R. H. Howes, Special formulations for low volume spraying, Proc. 1st Brit. Insecticide and Fungicide Conf., Brighton 1961, 533-546.

fluid. We have already referred above to the lower evaporation that results from this, but what is even more important is that a more viscous liquid forms coarser droplets upon atomization and that the percentage of fine droplets is correspondingly reduced. Many years ago efforts were made in France to improve the deposition of spray liquids in this way; large-scale investigations have recently been undertaken in the U.S.A. to study the effects of viscosity in the atomization of herbicides. There is keen interest in this subject in the U.S.A. as in some States the spraying of herbicides from aircraft has been prohibited on account of the serious damage that has been caused by spraydrift.

The sprayed droplets should not however be too large. They might then easily be disintegrated by the strong turbulence that occurs during spraying, and this in turn could decrease the size of the droplets. Not only is viscosity important, it is also of interest to know whether the liquid possesses "visco-elastic" properties, in which the droplets have a tendency to contract, so that they do not break apart so quickly.

Good results have recently been achieved with the atomization of certain herbicides (solutions of trichloroacetic acid and emulsions of trichlorophenoxyacetic acid esters) when high-molecular cellulose derivatives were added in the form of a colloid. In atomization from aircraft a concentration of 0.75% hydroxyethyl cellulose in the spray liquid proved to be sufficient to eliminate spraydrift. A further advantage of the use of such a colloid in the atomization of dispersions is that these substances stabilize a dispersion. This is particularly true for mixtures of emulsions and suspensions, as we mentioned previously.

If we consider only the atomization and deposition of the liquid, the cellulose derivatives provide a very useful means of reducing spraydrift: by varying the concentration one can adapt the viscosity to any sprayer or atomizer in a wide range. However, looking at the total effect of the method, we see that its usefulness is again limited. In the first place, it is necessary to take into account the density of cover required at the crop for the droplets of the spray in use. With a systemic herbicide, the low density obtained when spraying with large droplets is no disadvantage; with a fungicide however, with its requirement for an extremely dense coverage of fine droplets, some spraydrift will have to be tolerated for this reason. Another drawback of the method appears after the droplets have settled on the crop. The active substance then has to be quickly released from the droplets and transferred to the plant. While a high viscosity is especially suitable in atomization it now proves to be a serious hindrance to the release of the active substance. This requires the droplets to spread out quickly, and this does not happen because of the

high viscosity. The droplet remains on the leaf, water evaporates from it, and as a result the viscosity of the droplet increases still further, and there is then a risk that the active substance will be encapsulated to an extent that can seriously impair its efficacy. Fortunately, not all active substances are affected to such an extent, but very lipoid pesticides which do not readily dissolve in water, or which only diffuse slowly through films of hydrophilic colloids (like these cellulose derivatives) are evidently not suitable for atomization in conjunction with these substances.

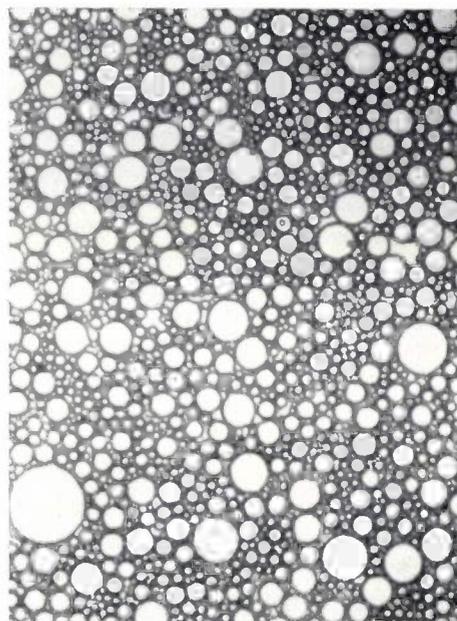


Fig. 5. Photomicrograph of a stable water-in-oil emulsion. The white discs are water droplets separated from each other by thin layers of oil. Magnification 330 \times .

It would be very useful to have a high viscosity during spraying and a low viscosity immediately after deposition. This cannot be achieved by the addition of a thickening agent. Systems can however be obtained which do exhibit this property; these are certain unstable dispersions whose viscosity is determined by a dispersed substance. If this substance coalesces after deposition, the desired change of viscosity can be obtained. A system of this description is found in "invert emulsions".

Invert emulsions

When discussing the conventional oil-in-water emulsions we referred to emulsions of the water-in-oil type (invert emulsions) only as an undesirable effect sometimes found in concentrated emulsions. Invert emulsions can be deliberately prepared however with

the aid of certain strongly lipophilic emulsifier mixtures^[4]. The viscosity of an invert emulsion shows an aspect that favours its use as a spray liquid; as a rule the viscosity of an invert emulsion is usually much higher than that of the constituent phases, so that after breakdown of the emulsion its viscosity is considerably reduced. For this reason the invert emulsions have attracted considerable attention in recent years for the atomization of herbicides and other agents from aeroplanes and helicopters.

In an invert emulsion water droplets are emulsified in a continuous oil phase. It proves possible to employ such an emulsion even when the quantity of water is greater than the quantity of oil. A system is then obtained in which large numbers of water droplets are separated by thin oil films which are nevertheless all joined. (see *fig. 5*). The viscosity of this emulsion is determined by the friction at the oil-water interfaces. If more water is taken up by the oil or if the water is more finely emulsified in it, the total surface area of the interfaces increases. This explains the surprising behaviour of these emulsions: the viscosity increases as more water is emulsified in the oil or as the water is more finely dispersed in it. (This is rather like the behaviour of whipped cream, which becomes stiffer the more air is beaten into it.) If one continues to add water, the emulsion usually breaks down (see *fig. 6*) and the oil appears as a film on top of the water.

Although the high viscosity may in itself be a good thing, too high a value sets limits to the usefulness of the emulsion. At a volume ratio of three or four parts of water to one part of oil the viscosity is already fairly high, and if further water is added the emulsion is almost impossible to spray. This means that for spraying purposes the emulsion cannot contain much water, so that a relatively small quantity of a fairly thick substance has to be distributed over a certain area. This sets difficult requirements for the spraying equipment. With many pesticides a quantity of 1 litre per hectare (1 pint per acre) is quite sufficient. With the herbicide 2,4,5-trichlorophenoxyacetic acid butylester, for example, 2 litres of a 50% solution in diesel oil is enough to destroy weeds over an area of one hectare. In this case an invert emulsion with about 6 to 8 litres of water with 2 litres of oil would only just be sprayable, which means that in the most favourable situation 10 litres of liquid would have to be distributed over one hectare.

Although the viscosity of an invert emulsion can be lowered and the quantity increased by diluting the emulsion with (say) diesel oil, there may be both economic and biological objections to this. Too much oil can "scorch" the plants, which impedes the uptake of systemic pesticides in the plant and thus reduces their efficacy.

Until techniques have been developed for spraying relatively viscous liquids, the application of invert emulsions will for the time being remain limited to pesticides which do not have to cover the plants completely. Good use can therefore be made of the method for spreading systemic agents.

In the U.S.A. the Stull Chemical Company has developed a special nozzle that does make it possible to spray invert emulsions with more water. In this system the oil and water are fed separately, under pressure, to the spray nozzles, where intensive turbu-

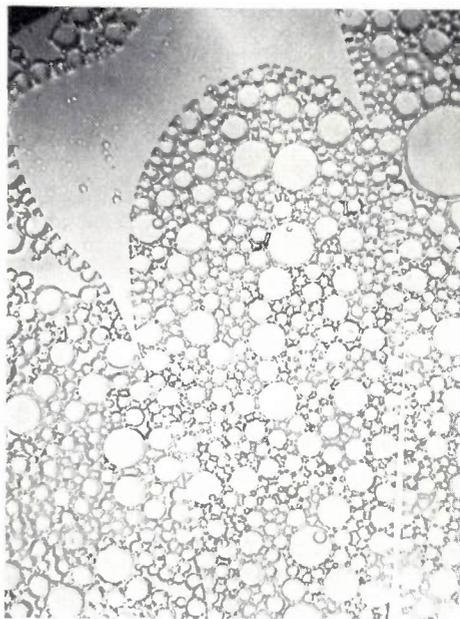


Fig. 6. Photomicrograph of a water-in-oil emulsion; so much water has been added that the emulsion has begun to break up. The water is shown white in the photograph. The oil phase can be seen to contract and to separate from the water phase. Magnification 400 \times .

lence produces a spontaneous water-in-oil emulsion. With this system water can be added in a volume ratio of 10 to 13 parts to one part of oil^[4]. As the emulsion is not formed until the liquids reach the nozzle, there is no high pressure or resistance in the system. It is clear however that this technique does require very active emulsifiers, which will also have a highly stabilizing effect on the viscous emulsion.

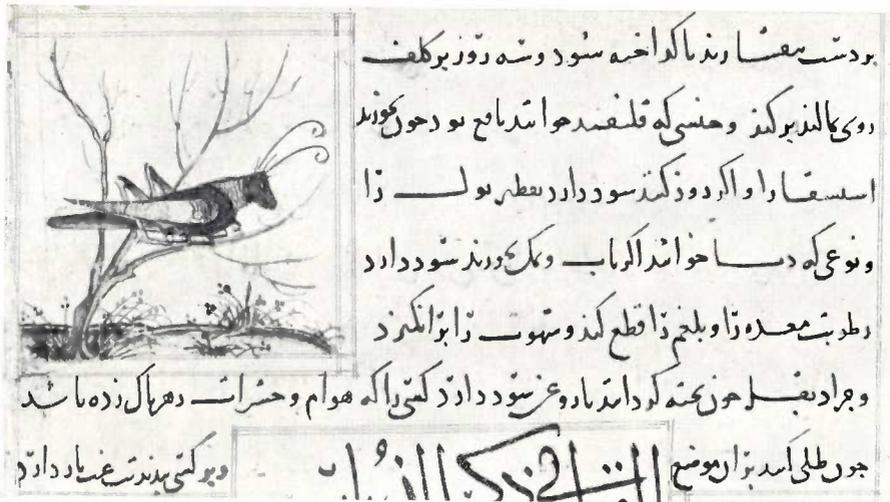
The reader will have gathered that there is as yet no method for atomizing concentrated dispersions which answers to the requirements of *all* pesticides. It is also very difficult at this stage to predict the direction

^[4] J. P. Colthurst, R. E. Ford, C. G. L. Furnidge and A. J. A. Pearson, Water-in-oil emulsions and the control of spraydrift, The formulation of pesticides, S.C.I. Monograph No. 21, 47-60, 1966.

which developments will take. Further refinement of spraying techniques will certainly be called for; we have mentioned the nozzles used for spraying undiluted "Malathion". It may well be that this will prove to be the best solution for pesticides where some spraydrift can be tolerated. For many pesticides however the investigations into methods of minimizing spraydrift will have to continue. In view of the wide variety of pesticides and the many different ways in which they act, it is unlikely that any generally applicable method will be found. Extensive research will still be necessary to find the best method for each substance.

Summary. The chemical control agents (pesticides) used in agriculture and horticulture are for the most part applied in the form of an emulsion or suspension in water. A difficulty in such methods is that with conventional spraying equipment the spray droplets can only be uniformly distributed over the crop if the liquid is considerably diluted. For spraying from aircraft, with their limited load capability, it is desirable to use concentrated liquids. The article discusses the various methods which can be used, and the physical and colloid-chemical problems involved. The main problem is the avoidance of "spraydrift", i.e. the dispersal of the droplets by the wind. This can be countered by ensuring that no unduly small droplets are formed in the spray nozzle, and by taking steps to prevent the droplets from growing

smaller because of evaporation during the descent. The article deals first with concentrated emulsions and suspensions and touches on the problems arising when emulsion-suspension mixtures have to be simultaneously atomized. Some adjuvants are then discussed which reduce the evaporation of the droplets, and also others which increase the viscosity of the droplets so as to increase their size when sprayed. All these methods of combating spraydrift have their disadvantages, such as reduction in pesticidal action because the active substance is not so readily released from the droplets after deposition. In conclusion, the "invert emulsions" are discussed (water-in-oil emulsions), which have various properties that may well make them useful for the atomization of certain concentrated pesticides.



Pierpont Morgan Library, New York

The control of agricultural pests and diseases through the ages

R. van der Veen

Plant diseases and pests have existed ever since man first began to till the soil.

Crops cultivated for food were a valuable possession that had to be defended against attackers, for a spoilt harvest often meant famine. It is scarcely surprising, therefore, that references to the control of diseases and pests in agriculture are not uncommon in historical writings, even of remote times.

Nowadays, when we talk of control we generally think of chemical agents, such as insecticides and fungicides, of repellents and of poisoned bait, etc. And if we want to use a term with a more contemporary ring, we talk about biological control, which is sometimes held to be the ultimate answer.

A study of agricultural pests of the past — on which several historical surveys have appeared in recent decades ^[1] — reveals that these modern methods of control are by no means so modern after all. Here, too, we find that there is nothing new under the sun. All we can say is that the old methods are now more refined and are put to use on a much wider scale. But then there are also many more people in the world.

The oldest agricultural regions known to us are Egypt, the Middle East and China. From all these regions we have records of ruined harvests, of the plagues that caused them and of the counter-measures adopted. The worst pests were apparently locusts, rats and mice, while in ancient Babylon almost two thousand years before the Christian era, the fungus causing smut in cereals wrought so much havoc that it had to be tackled with a magic charm. Locusts, rats and mice are still today among the major agricultural pests; and cereal smut was still the cause of one of the worst diseases in corn right up to the 19th century.

Insects affecting goods in storage must also have been quite a problem in ancient times. Remains of *Lasioderma serricorne* and *Tribolium castaneum*, which still do a great deal of damage today, have been found in a well-sealed Egyptian vase and in the tomb of Tutenkhamon. The excellent grain stores built in Egypt during the reign of Rameses II must not only have been designed to prevent rot but equally to protect the grain against infestation by insects.

Locusts seem to have been feared most, then as in

Dr. R. van der Veen is Professor of Botany at the University of Utrecht, Netherlands.

[1] Two of the most recent may be mentioned here:
H. Braun, *Geschichte der Phytomedizin*, publ. Parey, Berlin 1965.
K. Mayer, *4500 Jahre Pflanzenschutz*, publ. Ulmer, Stuttgart 1959.

The title photograph is reproduced from a Persian manuscript by Baktishu, written in c. 1295.



Fig. 2. Part of a fresco in Graz cathedral, painted in 1480 by Thomas von Villach. This part represents the three great pestilences of the age: on the left the locusts, on the right the Plague and in the middle the Turks. Graz was near the front line in the centuries-old struggle against Turkish invaders. The fresco of the locusts is reproduced here on a larger scale. (This picture is reproduced by courtesy of the Dompfarramt at Graz, from a restored copy; the original is in poor condition. The drawing is reproduced from E. Schimitschek, *Anzeiger für Schädlingskunde* 26, No. 1, 42, 1953.)

later times (see below). They are mentioned frequently in records of ancient Egypt; in about 700 B.C. the Assyrian Kings Sargon II, Sanherib and Assurbanipal had inscriptions made to commemorate the years in which locusts had been a severe scourge. In about 1075 B.C. the Chinese Emperor Chen Tsun issued a decree commanding landowners and headmen of villages, to gather and destroy the eggs of locusts and to catch and destroy the fully-grown insects, with a penalty of a hundred or more strokes of the birch for neglect of this duty.

Just how seriously the threat of insects to the harvest was regarded at that time appears from a proclamation by the minister Yao Tsung, who, in 714 B.C., called upon everyone to co-operate in the fight against insects in order to save mankind. One is reminded of the words spoken in recent times by one of the presidents of Brazil, who appealed for the most drastic measures to control the parasol ant (of the genus *Atta*), since this struggle would decide whether Brazil would henceforth be for the Brazilians or for the parasol ants.

Although one gains the impression that insect control was largely carried out by hand, a kind of chemical control had in fact long been practised in China at that time. In about 1200 B.C. lime and potash were used there against insects in stores and granaries, and before planting, seeds were protected from devouring insects by vegetable insecticides. China was also using arsenic as an insecticide before the Christian era.

In Greek literature, too, some examples are to be found of the use of chemical agents. Homer mentions the therapeutic action of sulphur. Aristotle recommends various fumigants, including sulphur, to combat disease and pests in agriculture. Democritus of Abdera reports the treatment of seeds with extract of aizoon (probably *Sedum acre*, or stone-crop). In this case it was evidently the tannins that provided the protection.

In the Roman Empire, somewhat later, adhesive bands were applied around the trunks of fruit trees, the adhesive used being a mixture of oil and bitumen.

In the first century of our era Pliny the Elder gives some directions on methods of chemical control, men-



tioning among other things arsenic as an insecticide. One of his remarks in this context, which is just as applicable today, is that one must beware of the remedy itself causing a disease, which can happen if the agent is applied too often or at the wrong time.

From all this it is clear that a method of control by no means merits the epithet "modern" simply because it is chemical. One might with equal justification use the term "classical"; it is only in formulation as sprays, dusts and aerosols that marked advances have been made. Although it is true that the number of agents used has increased enormously, several of the old agents, such as sulphur and arsenic, are still in use.

What, now, is to be said about the more modern sounding "biological" methods of control? Much less is to be found about these in ancient writings. Even so, biological control was not entirely unknown. To find a reference to it we must again go back to China, this time Cochin China. We find there a report from about the year 300 that ants were collected in straw sacks and sold on the market in order to be hung up in mandarin orange trees. The idea was that the ants would destroy the aphids and many other noxious insects infesting these trees.

Strangely enough, we come across almost identical advice in 1579 from Thomas Lupton^[2]. His recipe for getting rid of caterpillars in fruit trees consisted in hanging a bag full of ants in the tree after putting a ring of tar round the stem to prevent them from escaping (*fig. 1*).

Here, then, are two clear examples of biological control in earlier times.

Nor is there anything new about the use of repellents. K. Mayer found in an Egyptian papyrus scroll dating from about 1550 B.C. a description of the treatment of walls and floors of granaries with a substance that repelled mice.

Very many years later Glauber (1604-1670) describes a concoction for repelling game that is strongly reminiscent of the game repellent still available from Philips-Duphar. He writes: "From the hairs or horns of the animals a spirit is also made which stinks most evilly". If rags soaked in this be hung up around a piece of land, "no stag or boar will enter therein to do damage, for they are allrighted by the stench they do smell from afar and believe a hunter to be close by who would shoot them"^[3]. It makes one wonder about the odour that hunters carried around with them in those days.

In Europe, the practical art of controlling pests and diseases made no progress for many centuries. Bad harvests and plagues of insects were regarded as punishments from God and as such were put on a par with the Black Death (the plague) and with the invasions of the Turks (*fig. 2*). Plagues of insects and other scourges therefore had to be averted by prayer. The notion of divine retribution in a plague of locusts found iconographic expression in the St. Barbara legend: the punishment visited upon the treacherous shepherd who betrayed St. Barbara's hiding place to her enraged father was to see his flocks transformed into a swarm of gigantic locusts. This theme is to be found on several altar panels of the Middle Ages (*fig. 3*).

282

The tenth Booke

you haue done with it , foꝛ so it fastneth and byndeth the stronger: and in such soꝛt, that it fastneth peeces of glasse together.

- 51 If you woulde destrope Caterpyllers, do thus. Anoynt all the bottom of the tree round about with Tarre, then get a great soꝛt of Antes oꝛ Wyllemys, and put them in some bagge, and drawe the same by by a coꝛde into the tree, and so let it hang there, so that it touch the body of the tree, and the Antes litted to go downe from the tree by the meanes of the Tar, wyl (foꝛ want of foode) eate and destrope all the Caterpyllers there, without hurting any of the fruit. This was tolde me foꝛ a very truth.
- 52 Make a hoale in the ende of a Goose egge, and put all the whyte and polke out of it, then put into the shell, a Backe that flies about in the evening, and then glew oꝛ close it fast on the toppe, and you shal see the Backe flye away with the same Egge shell: to the great maruayle of them that knowes it not.
- 53 It is verry euill (foꝛ them that fallse syck?) when the Moone is applyng in conjunction to the Sun, and woosle, when she is within fyre degrees of the Sunne. This I knowe to be true by often prooffe and

British Museum, London

Fig. 1. Page from Lupton's book (1579^[2]), showing as item 51 the description of a biological method of control of caterpillars.

^[2] Thomas Lupton, *A thousand notable things of sundry sortes. Wherof some are wonderfull, some strange, some pleasant, diuers necessary, a great sort profitable and many very precious*, London 1579, p. 282.

^[3] See the book by H. Braun, page 25.



Fig. 3. One of the panels of the St. Barbara altar by Master Francke (Hamburg, c. 1420), from the church of Kalanti-Nykyrko, Finland. It represents a scene from the St. Barbara legend, in which the flocks of one of the two shepherds are transformed into gigantic locusts as a punishment for disclosing the hiding place of St. Barbara to her father^[4]. (Reproduced by courtesy of the National Museum of Helsinki, where the altar is now situated.)

It was not for nothing that locusts loomed so large in the popular imagination: their arrival signified wide-spread distress, and in broadsheets printed from woodcuts — the newspapers of the time — they were frequently represented as most fearsome creatures (fig. 4).

The control of pests by penance and prayer was again, of course, nothing new, for other and older religions than Christianity had their sacrificial rites

Later there were even ecclesiastical trials of harmful animals and insects, and later still they were prosecuted in the secular courts. In 1320 legal proceedings against the cockchafers were brought before the ecclesiastical court of Avignon. The cockchafers were summoned to appear, with the threat of excommunication from the Church if they failed to present themselves. They were provided with a defence counsel who pleaded the cockchafers' right to food. This right was recognized



Wick Collection, Zentralbibliothek Zürich

Fig. 4. Broadsheet dating from 1556 showing locusts. This illustration reveals how strong the feeling against the locusts was. They repeatedly destroyed the harvest in those days in many parts of Europe; in 1538, for example, they were responsible for famine in Rumania.

and supplications for a good harvest. Since social life in Europe in all its aspects was dominated so long and so completely by the church, there was little for it in the first 1600 years of our era but to accept famine as a punishment and to pray for good harvests. The exponents of this view were the patron saints, who were called upon to protect the farmer against injurious beasts and insects. The first of these was St. Magnus, Abbot of Füssen, who in the year 666 warded off an approaching swarm of locusts with the staff of St. Columbanus, a scene repeatedly found in church and monastery (fig. 5). St. Bernard pronounced an anathema upon the gnats in 921, whereupon they fell in heaps upon the earth. The Bishop of Lausanne, in full canonicals, solemnly cursed the cockchafers (fig. 6), the parish priest of Kalteren the locusts.

and the cockchafers were allotted a piece of land which was to contain sufficient food for their needs. The cockchafers were ordered to proceed to this defined territory within three days, and those that remained outside were outlawed.

In 1481 a similar trial was held in Basle, and notwithstanding the pleadings of the Freiburg Faculty on their behalf, locusts were declared anathema. In 1585 caterpillars came to trial in Valence and were sentenced to be banished the country.

The last ecclesiastical trial of this kind took place

[4] B. de Gaiffier, *Le triptyque du Maître de la légende de sainte Barbe*, Rev. belge d'Arch. et d'Hist. de l'Art 28, 3-23, 1959, which discusses the version of the legend as chronicled by Jan de Wackerzeel. (This version is accurately represented on a triptych, parts of which are preserved in Brussels and in Bruges.)



Fig. 5. The patron saint St. Magnus wards off a swarm of insects, presumably cockchafers and locusts, with the staff of St. Columbanus. Ceiling fresco in the Abbey Church of Schussenried (Württemberg, Western Germany).

in 1733 in Baumantant in France, and the last secular trial was in 1830 in Denmark.

Meanwhile the study of pests and diseases began to become more scientific. The great advances made in biology in the 16th and 17th centuries, partly as a result of the use of the microscope, and which resulted in the appearance of numerous works in which the principal insects (*fig. 7*) and many micro-organisms were dealt with, had prepared the ground for a more experimental approach. One of the pioneers in this field was Matthieu Tillet (1714-1791). He devoted a special study to smut in cereals, a disease that caused enormous damage in the cornfields. In 1755 he published his "Dissertation sur la cause qui corrompt et noircit les grains de bled dans les épis; et sur les moyens de prévenir ces accidens" (*fig. 8*).

After careful observations Tillet arrived at a number

wash was recommended for fruit trees. Bordeaux mixture came into use and powdered sulphur proved to be effective against mildew.

At the end of the last century institutions were set



Fig. 6. Pest control in the Middle Ages: the Bishop of Lausanne pronounces a curse on the cockchafers. (From H. Braun ^[1], page 14.)

of conclusions. He proposed experiments that well-nigh met the requirements of modern field tests. He predicted the results and, to many people's astonishment, all his predictions came true. One example was that healthy grains dusted with black spores before planting resulted in a diseased crop. If, however, the grains, after dusting with spores, were treated with nitric acid, quicklime or other agents, the result was a healthy crop.

The study of plant diseases now entered upon a period of swift development. One discovery followed another. Directives appeared on the disinfecting of seeds; lime-sulphur

up in many countries for the control of pests and diseases in agriculture. The causes became better known and better methods were found for applying the measures of control.

Natural disasters, caused by agricultural diseases, gradually became more infrequent. The last one produced by plant disease was the Irish famine. Ireland had specialized in the growing of potatoes. Between 1843 and 1845 the entire potato crop was attacked by a fungus called late blight, which ruined one harvest after another and resulted in widespread famine. Many of the population emigrated to America. The fungus also spread across Europe, and late blight is still a serious potato disease to this day. The growing of potatoes on a large scale is possible only if the crops are repeatedly sprayed with fungicides.

Many people believe that organic insecticides and fungicides did not come into use until after the last



Fig. 7. Frontispiece of the book "Schouburg der Rupsen, Wormen, Maden en Vliegende Dierkens" by S. Blankaert, Amsterdam 1688, one of the general works in the field of entomology in the 17th century.

DISSERTATION

SUR LA CAUSE
QUI CORROMPT ET NOIRCIT
LES GRAINS DE BLE
DANS LES ÉPIS;
ET SUR LES MOYENS
DE PREVENIR CES ACCIDENS.

QUI A REMPORTÉ LE PRIX AU JUGEMENT DE
l'Académie Royale des Belles-Lettres, Sciences & Arts de Bordeaux.

PAR Mr. TILLET, de Bordeaux, Directeur de la Monnoye de Troyes.



A BORDEAUX,

CHEZ la Veuve de PIERRE BRUN, Imprimeur Aggrégé de
l'Académie Royale, rue Saint James.

M. DCC. LV.

AVEC PRIVILEGE DU ROI.

Fig. 8. Title page of the treatise by Tillet, which was of great importance to the development of the experimental study of agricultural pests and diseases. (From H. Braun^[1], page 37.)

world war. This is not entirely true; probably the first synthetic organic agent was dinitro-orthocresol which was developed as long ago as 1892 by Elberfelder Farbenwerke. It was marketed for the control of the nun moth under the name of "Antinonin".

It was not until much later that the insecticidal action of HCH was discovered (1933) and later still that of DDT (1939).

In the meantime however, a number of other organic agents, which were derived from plants, became available for pest control. In 1690 La Quintinye recommended tobacco extract for use against aphids. Much later, in 1844, a decoction of tobacco leaf was often used in vineyards to keep down the insects. This agent, now concentrated and sold under the name of nicotine, is still a widely used insecticide, even though it is highly poisonous.

In 1840 Anna Rosauer discovered the insecticidal action of Pyrethrum, which soon afterwards appeared on the market as Dalmatian insect powder. It is probable that the inhabitants of Asia Minor and the Balkans were familiar with this action much earlier. This insecticide, which is quick-acting and harmless to human beings, is still in common use. Because of its rapid

action (often referred to as "knock down" effect) and low toxicity to humans, pyrethrin is a common constituent of aerosols. After its chemical structure had been elucidated, it was possible to produce synthetically various other related substances with the same action, one example being "allethrin".

Later the insecticidal action of the Derris root was discovered, and this was followed by an insecticide prepared from ground Derris roots or extracts containing rotenone.

After the second world war the group of synthetic pesticides made particularly great strides. This period cannot, however, be regarded as sufficiently "historical" to be dealt with here.

Nevertheless, one facet of the greatly increased application of chemical control should be mentioned. Since a number of the chemical agents at present employed are fairly toxic, both to humans and to domestic animals, the government authorities in nearly all countries have subjected their use to strict control. People had begun to feel rather anxious about the fact that much of their everyday fare had been sprayed or dusted with some poisonous agent or another.

Even this is nothing new. K. Mayer mentions a decree promulgated in Karlsruhe in 1808 making it an offence to trade in toxic substances such as arsenic and corrosive sublimate for insecticidal purposes, as there were other less dangerous preparations that could be used in their stead. The sale and use of these toxic substances were permitted only to merchants who

kept strict records of them and who complied with the safety regulations laid down. It seems highly probable that the use of arsenic and mercuric chloride in those days resulted in casualties now and then — through foul play or by accident, and the decree was no doubt meant to deal with the first contingency as much as the second.

In view of the very widespread use of pesticides at the present time, it is obvious that government measures to control their use had to become even more rigorous. Such control is necessary to assure the safety of the consumer. The reverse side of the coin is the inhibiting effect it has on the development of new and better pesticides, which cannot be introduced until they have been subjected to lengthy and very costly preliminary investigations into their toxicity and other possible side-effects.

Summary. A study of the history of agricultural pests and diseases shows not only that they were known in distant antiquity but also that chemical as well as mechanical means of control were already in use in very early times (certainly before 1000 B.C.). Even biological control was known many centuries ago. The religious climate in Europe up to the 16th century led people to take a long-suffering attitude to such plagues, and to pin their hope on the effectiveness of penance and prayer or on the intervention of their patron saints. A remarkable corollary, which seems quaint to us, was the prosecution of the offending insects in trials before the ecclesiastical and secular courts. In the 18th century, after great new developments in the science of entomology, agricultural pests and diseases and the measures of controlling them began to be studied experimentally. The article concludes with a description of some of the main stages in the progress towards present-day methods of control.

The whole of this issue is devoted to work carried out at the Mullard Research Laboratories, Redhill, England. The history and present main fields of activity of the Laboratories are described in an introductory article by P. E. Trier, the Director. A series of articles by members of the Laboratories then follows and provides a selected survey of work done in the recent past, with an emphasis on projects that have reached an element of completion and achievement. Current work of an in-

complete character, even where important, has only been briefly referred to.

The articles were selected and written under the guidance of an ad hoc Editorial Committee at the Laboratories under the chairmanship of Professor K. Hoselitz, the Deputy Director. It is a pleasure for the Editor to give credit to this Committee and to present to our readers this interesting cross-section of a varied programme of industrial research.

The Mullard Research Laboratories

An outline of their growth and function

P. E. Trier

Historical introduction

At the end of the Second World War, the Mullard Company was already one of the most important suppliers of valves and components to the British Electronics Industry. The late S. S. Eriks, then Managing Director, clearly foresaw the great coming expansion and diversification of this industry, and the new demands which in turn it would place on the component and building element industries. He also recognized the important role that the British Government Laboratories would play in the development of new electronic systems and applications for defence, telecommunications and nuclear technology. All these new activities would need a technical matching organism through which the Company could put itself abreast of the new requirements to help in forming its own policy for future products, and to lay the foundations for co-operative work with the Government on advanced projects.

To meet these aims, the Mullard Research Laboratories were started in November 1946, and initially occupied a small wing in an old radio factory at Sal-

ford, near Redhill, in Surrey. By 1952 the whole original building was occupied, and in 1954 a new building programme was started on the site which has continued ever since, with new buildings and land acquisitions from time to time to the south of the original site. The original single story building still forms part of the Laboratories and houses the Engineering Division.

The early work of the Laboratories was mainly directed towards new applications of the Company's products particularly in television and military systems. It was realized at the outset that, to effect the translation of experimental ideas into soundly engineered prototypes, strong engineering and model shop facilities must form an essential part of the Laboratories. In any case the research and development work on microwave structures alone demanded precision engineering of a very high order which could not have been met except by craftsmen intimately involved in the work. The first big step towards an enlargement of the scope was the transfer in 1948 of the Special Vacuum Physics Laboratory from the valve factory at Mitcham to the Research Laboratories, which allowed the start of new work on microwave tubes, photoconductor devices, and high-vacuum techniques.

By now the Laboratories were capable of undertaking major new development tasks and were also initiating research work in new fields. After the experimental demonstration by D. W. Fry of the linear electron accelerator, Mullard Research Laboratories were asked to design and make the first engineered machines at 4 MeV and 15 MeV [1]. The origin of the Laboratories' pioneer work in ultrasonics also dates from then; although many of the results are now standard products, there is still some continuity of activity in the work on ultrasonic signal processing [2].

The next step in widening the range of the work was taken in 1952, when a Solid State Physics Division was formed to study and exploit semiconductor phenomena, ferri-magnetism and para-magnetism, which were already beginning to transform the materials and devices available for electronic applications, thus marking the beginning of the new era of solid-state electronics.

The basic structure of the Laboratories was thus determined in the early 1950's and has maintained a recognizable pattern ever since. There are three main elements in the activities: fundamental investigations of the physical phenomena underlying the behaviour of electronic materials and devices, the creation and development of technologies for new devices, and the application of these devices in new equipments and systems. Although the Laboratories are organized on the basis of autonomous laboratory divisions with recognized boundaries of work, it is inherent in the nature of current trends that each division in varying degrees will comprise work under all the above headings, and that many projects demand an organization and activity that cuts across divisional boundaries.

In spite of the continuity in the discernible pattern of the work over the last 15 years, there have been many changes in subject matter and emphasis, as well as in personalities and external links. At the beginning of this period, much of the work was directed towards the development of devices and equipments which could be taken into production straight away in one of the factories of the Company or elsewhere. The accent was strongly on evolution rather than innovation. There was a rush to develop special products for which the factory development laboratories were not yet equipped and which were therefore often undertaken at Salfords, even including production of initial runs. It had however always been recognized that the engineering development of new production items is best undertaken in the production unit itself, to ensure design methods compatible with production processes and with the lowest costs. Many projects were therefore handed over as time went on to the growing factory laboratories, and this was often accomplished by transferring entire teams from the Research Labor-

atories with their work, either to strengthen existing development laboratories or to form the nucleus of new ones. Teams from the Research Laboratories have in this way helped to build up development laboratories in the factories dealing with semiconductors, microwave tubes, computer stores, electronic systems and also the Mullard Central Applications Laboratory at Mitcham; and there have been many individual transfers.

This shift allowed a growth of effort at the Research Laboratories on new principles, new device concepts and new systems embodiments, together with the growth of new technologies to allow their experimental realization. This growing emphasis on genuine innovation could only be accomplished by a gradual raising of the level of scientific understanding and the technical expertise of the staff, as well as their flexibility in tackling new tasks. It has been accompanied by a steadily increasing degree of contact and collaboration with the Philips Research Laboratories in Eindhoven and associated laboratories elsewhere in Western Europe. Simultaneously the co-operation with Government laboratories in Great Britain, which from the start was an important element, has been widened and deepened; and there has been an increasing volume of contact with Universities and Colleges of Technology.

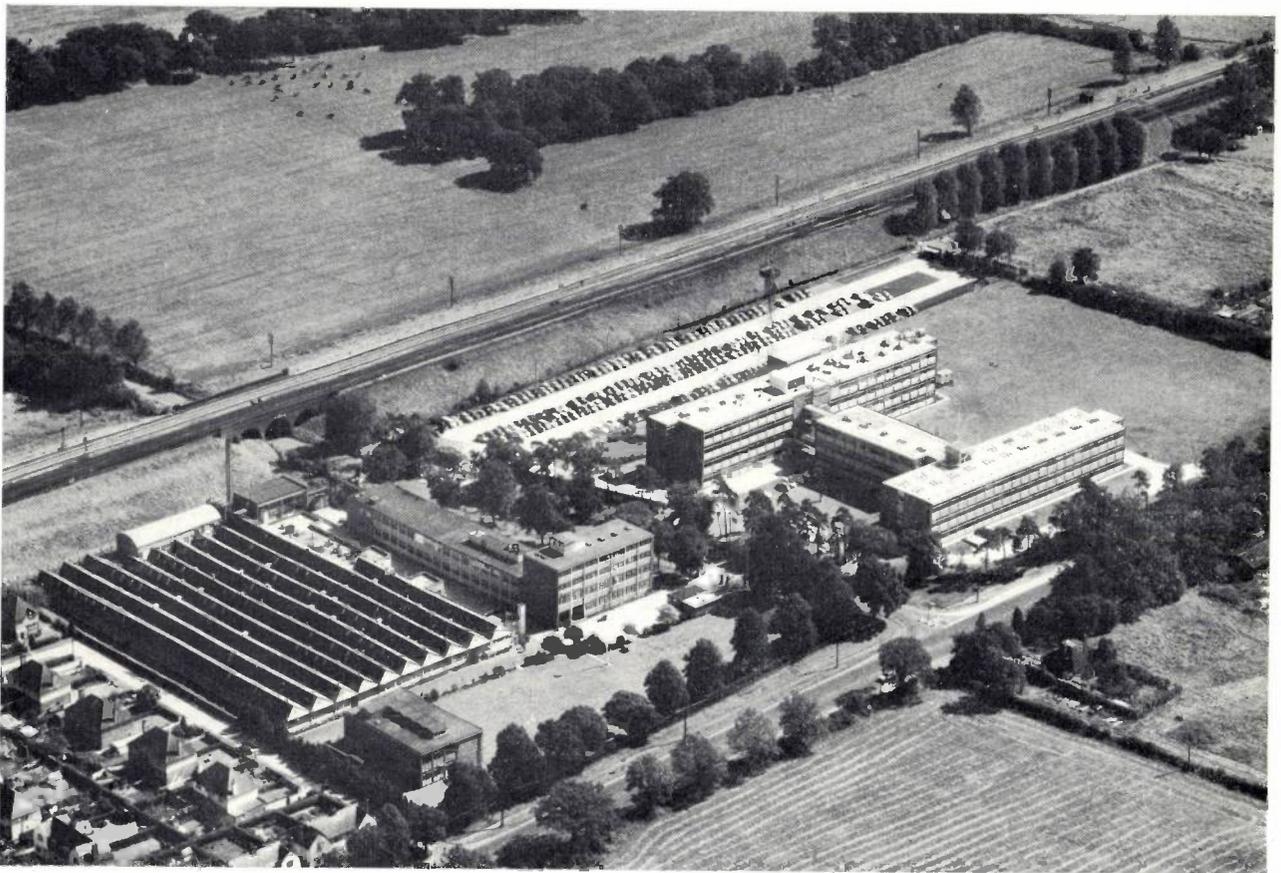
The size of the Laboratories grew rapidly in early years, and reached a level of nearly 700 in total numbers by early 1956. By then working conditions were exceedingly cramped. A policy of deliberate containment has therefore been pursued ever since; and the continued influx of new staff has been safeguarded by the continuous outward transfer of groups and individuals as already mentioned. Numbers are at present in the range 600-650. Of these, 200 are University graduates and a further 50-70 have equivalent qualifications. The gross space available grew from a beginning of 10 000 sq. ft. in 1946/47 to 50 000 sq. ft. in 1952, the full size of the original one storey factory building. Since 1954 a progressive building programme has added facilities step by step; the latest addition commissioned in 1966 brought the total gross space to 195 000 sq. ft. This is adequate for the scale and type of work on which we are presently engaged. The location of the Laboratories in a semi-rural setting but with good access to London and to the Company's main factories in the south of England has proved very useful.

-
- [1] D. W. Fry, The linear electron accelerator, Philips tech. Rev. 14, 1-12, 1952/53. C. F. Bareford and M. G. Kelliher, The 15 million electron-volt linear electron accelerator for Harwell. Philips tech. Rev. 15, 1-26, 1953/54.
- [2] J. S. Palfreeman, An opto-acoustic cross-correlator in radar signal detection, page 217 of this number.

Present scope

The Laboratory organization centres round four experimental Divisions, the Solid State Physics Division, the Vacuum Physics Division, the Circuit Physics Division and the Systems Division. These are supported by a range of central technical facilities working on behalf of all the Divisions, of which the largest is the Engineering Division comprising design and drawing offices, machine shops, construction shops, wiring shops, chemical and photographic technology. Additional

an emphasis throughout on the physics and materials for electronic devices, on the electrical characteristics of these devices and their assemblies, and on technologies for their realization. Even in the early days it was nevertheless realized that the evolution of successful new devices demands a deep penetration into their application in circuits and their use in complete systems; this alone allows the continuous mutual interaction between device supplier and equipment designer which is necessary in a field where all the time new



The Mullard Research Laboratories. The Laboratories adjoin the main London to Brighton railway line.

support facilities include computing, optical laboratories, glass technology, instrument maintenance, technical library and information services as well as the normal range of administrative and plant facilities needed for the running of a self-contained establishment.

To appreciate the scope of the work and the aims and boundaries of project choice, a number of factors must be considered. Primarily the Laboratories work within the framework of an electronic component and sub-assembly Company structure, and there is therefore

device possibilities trigger off new applications, but where conversely new user demands point the way to priorities for the invention and development of devices. The advent of the semiconductor integrated circuit and the further step towards large-scale integration make these observations almost commonplace, but they have in considerable measure been true long before this. Hence the Laboratories have always maintained a strong activity on circuit applications and systems studies. In the field of domestic consumer electronics, particularly

television, it has been possible to mount a substantial application effort leading to authoritative knowledge on every aspect of receiver techniques. This has helped set-maker customers to rationalize their designs, and at the same time helped the Company itself in the standardization of type ranges for display tubes, valves and semiconductors, deflector assemblies and other components. In the field of professional electronics comprising telecommunications, military systems, control systems, computers and many other areas, it is not possible to set up an equally comprehensive and authoritative research and application effort without the most intimate and selective user involvement. In this context the device and systems studies undertaken on behalf of Government laboratories as research or development contracts have been most valuable in providing user involvement, and in injecting technically advanced requirements as targets for new techniques and methods of approach. In recent years, the Government interest has been enlarged to include civil electronics, and this has been marked by the launching of new types of co-operative research projects involving several industrial firms in a common effort with the Government. The administration of such projects can lead to very complex problems, but the objective of bringing device and equipment designers into close partnership is thoroughly beneficial.

The Systems Division in the Laboratories has also over many years done work in the equipment field on behalf of the M.E.L. Equipment Company Ltd., an associated company, who in the course of time have taken over many advanced development projects from the Laboratories into engineering and production.

Another important feature in determining details of the range of work is the increasing co-operation and involvement with the work carried out at the Philips Research Laboratories in Eindhoven. There are of course a number of projects at the Mullard Research Laboratories on which confidential relations with outside interests based on commercial or other grounds preclude a free interchange of information with the Philips Laboratories. Such projects are however in the minority; in most cases where basic investigations of new physical principles or device techniques are concerned, a two-way contact and exchange of knowledge is entirely beneficial and is usually possible. At least two substantial advantages arise from this contact: first there is the possibility of each laboratory drawing on the experience of the other, and of saving much time particularly in the establishment of new materials technologies; secondly it allows the sharing out of some areas of research work, thus reducing duplication which is particularly wasteful in the case of speculative projects whose ultimate success and industrial

viability may still be in question. By way of example, work on masers is done only at these Laboratories, whereas work on new lasers is carried out almost entirely by the Philips Laboratories in Eindhoven. The staff of the Laboratories invariably welcome the possibility of contact with other scientists at an international level.

Some activity highlights

In the articles which follow, selected projects from the current range of work are described. Some connecting remarks will therefore be made here about the work of the Divisions and about some projects which are not reported on in the detailed articles.

The *Solid State Physics Division* works mainly on semiconductor physics, semiconductor devices, magnetic phenomena^[3] and quantum electronics. This work is supported by facilities for metallurgy, chemical analysis, crystal growth and other materials work, and there is also a group for theoretical physics and applied mathematics. One of the highlights of work in this Division has been the successful realization of the travelling wave maser^[4] practically simultaneously with similar work in U.S.A. A number of maser installations for low-noise receivers have since then been designed and executed in the Laboratories, mainly in the Systems Division. Two of these masers are in service for transatlantic telephony at the ground terminal for the Intelsat I (Early Bird) satellite in Cornwall. Potential new maser materials have also been examined^[5].

In the semiconductor field, a very big current effort drawing in work from all Divisions is devoted to a project in *large-scale integration, based on the Metal Oxide Semiconductor Transistor*, and aiming at a fast computer store using MOST structures. The ultimate aim is to make a substantial part (1000 bits) of a computer store on a slice of silicon. Work is currently being directed at establishing the techniques for making P-channel MOST bits at high yield and for defining an appropriate interconnection pattern by means of a computer controlled mask-making machine. Automatic probe testing with computer analysis of faults is

[3] F. W. Harrison, R. F. Pearson and K. Tweedale, Single crystal research on ferrites and garnets, page 135 of this number.

[4] J. C. Walling, Travelling wave masers, *Solid-State Electronics* 4, 225-234, 1962. J. C. Walling, Travelling wave masers, 5th AGARD Avionic Panel Conf. Oslo 1961, reprinted in *Low Noise Electronics*, 225-235, 1962. J. C. Walling and F. W. Smith, A travelling wave maser for satellite communications systems, *Quantum Electronics III-1*, 923-930, 1963. J. C. Walling and F. W. Smith, Solid state masers and their use in satellite communications systems, *Philips tech. Rev.* 25, 289-310, 1963/64.

[5] J. W. Orton, D. H. Paxman and J. C. Walling, The physics of maser materials, page 146 of this number.

now being used, and the results will be compared with optical and infra-red examinations of the slices to determine the causes of failure. In parallel with these and related technological studies [6] [7], selection circuits suitable for integration are being designed and an assessment is being made of the possibilities of making high speed stores using this approach.

A novel semiconductor device on which work is going on is the optically coupled transistor. The principle of this transistor was proposed independently at M.R.L. [8] and at I.B.M. [9]. The device has a similarity to the transistor but differs from it as its operation depends on the transmission of photons across the base rather than on the diffusion of minority carriers as in the transistor.

The essential components of the optically coupled transistor are an electro-luminescent diode — the emitter, and some form of photo detector — the collector. Because the device operates through photon transmission across the base it can be designed with a much thicker base and hence lower base resistance than a conventional transistor for operation at a similar frequency. The device thus appears to have very real possibilities as a microwave amplifier.

The problems of realizing this device are principally those of developing electro-luminescent diodes and collectors that are both efficient and fast. Our work has been based on the use of GaAs and its derivatives (GaInAs, GaAsP) for the formation of the emitter, base and collector. The present devices will not yet operate at microwave frequencies but their performance is sufficiently encouraging to justify further work.

The *Vacuum Physics Division* comprises groups working on electron emission, electron optics, image forming devices, high vacuum and gas discharges, microwave tubes and electron beam processes [10]. The work on microwave tubes has been progressively scaled down with the transfer of work to the factory laboratories; and the greatest single concentration of effort in the Division now is devoted to the field of *night vision*, and the associated image devices such as image converters [11] and channel electron multipliers [12] which feature in detail articles in this issue.

Over the past ten years a team in the Laboratories has made a number of important advances in night vision techniques [13]. Purely passive optical instruments such as telescopes or binoculars require additional active devices that can provide brightness amplification. These devices, image intensifiers, must have photoelectric detection surfaces that are at least as sensitive as the human eye and must be designed in such a way that every photoelectron emitted is registered as an event by the observer's eye.

The studies have led to a number of new image in-

tensifiers as well as important innovations in electron optics. An example of this has been a variable magnification electrostatic system with a "zoom" facility of 4 : 1. Tubes that incorporate fibre optic windows can be made where the ultimate resolution is determined by the window and screen characteristics alone. Further studies in screen technology give promise of higher resolution and maintained luminous efficiency.

The military significance of methods of improving visual acuity at low light levels has not been ignored and a great deal of the work has been devoted to this end. However, current research activities on channel electron multiplication give promise of a wider range of devices that might well be applied to civil use. Very low light level television may become a practical possibility.

The *Circuit Physics Division* has up to recently carried a substantial application effort on television systems and receiver techniques. Much of this work has now been absorbed by the Central Applications Laboratory at Mitcham, but the Division still carries out advanced work on receiver and display techniques, partly in relation to television, partly on other applications. Evaluation of potential new colour displays is still important; and the adaptation of integrated circuits to linear applications is a pressing topic of work. In the professional fields there is a group for applications of electronics in motor-cars [14], and a group for ferrite applications [15]; both these are represented by papers. The main effort in professional electronics is centred in a number of groups working on computer circuits.

In these, the most important current topic is the *Magnetic Thin Film Store* work [16]. Research work on

-
- [6] A. F. Beer, J. B. Coughlin and P. J. Daniel, Some applications of contour deposition, page 153 of this number.
 [7] K. H. Nicholas, Studies of anomalous diffusion of impurities in silicon, page 149 of this number.
 [8] J. R. A. Beale and P. C. Newman, British Patent Appl. 48360/62.
 [9] R. S. Rutz, Proc. IEEE 51, 470, 1963.
 [10] H. N. G. King, Electron beam processes, page 174 of this number. I. H. Lewin, Drilling of diamonds using electron beams, page 177 of this number.
 [11] P. Schagen and A. W. Woodhead, Image converter and intensifier research, page 161 of this number.
 [12] J. Adams and B. W. Manley, The channel electron multiplier, a new radiation detector, page 156 of this number.
 [13] P. Schagen, D. G. Taylor and A. W. Woodhead, An image intensifier system for direct observation at very low light levels, Adv. in Electronics and Electron Physics 16, 75-84, 1962. P. Schagen, D. G. Taylor and A. W. Woodhead, A two-stage electrostatic image intensifier with a large photocathode area, Adv. in Electronics and Electron Physics 16, 105-112, 1962. P. Schagen, Electronic aids to night vision, Television Soc. J. 10, 218-228, 1963. A. W. Woodhead, D. G. Taylor and P. Schagen, An experimental image-intensifier tube with electrostatic "zoom" optics, Philips tech. Rev. 25, 88-95, 1963/64.
 [14] R. W. Lindop, An experimental electronically controlled car transmission, page 179 of this number.
 [15] C. V. Newcomb and E. C. Snelling, Analysis of variability in ferrite cored inductors, page 184 of this number.
 [16] R. V. Peacock, Critical parameters of evaporated NiFe films for fast stores, page 188 of this number.

the magnetization processes started in 1957/58. Mechanisms of magnetization reversal were studied intensively. This resulted in important contributions to the understanding of the induced uniaxial anisotropy, the nucleation of reverse domains, the role of damping in rotational switching, the understanding of creep, direction and magnitude dispersion of anisotropies and its relation to magnetization ripple. The proposal of using these flat thin film elements for a fast computer store was phased with the material and switching study and this led to the development of a demonstration store of 2560 bits (64 words of 40 bits) with a cycle time of about 200 ns. The feasibility of using flat thin magnetic film techniques for fast and very fast storage was thus demonstrated and the circuitry problems associated with these techniques emerged and could be tackled. A prototype main store of about 10^6 bits is under construction, to establish the manufacturing technology and speed and size limitations for such equipment. Results suggest that flat magnetic thin film stores will be used as main computer stores, especially as cycle times faster than those of core stores come into demand.

The *Systems Division* works in very diverse fields, including several microwave activities devoted to special components and devices such as masers, microwave semiconductor devices and broadband passive components, as well as complete microwave systems. In addition there are groups on signal processing (including optical character recognition)^[17], industrial electronic systems and instrumentation, a new group on gas chromatography, and a continuing group on problems of particle accelerators. Some of the characteristic aims of the Systems Division have already been mentioned above, and several topics are covered in the following articles, but two fields will be summarized here, microwave solid state devices and broadband microwave receivers. Substantial research and development is proceeding on microwave solid-state devices. The main effort is in the field of parametric amplifiers, and there has been a substantial effort, now diminishing, on solid-state masers which has already been mentioned. After successful designs of masers (3 GHz and 9 GHz bands) considerable progress has been made with masers in the 30 GHz band and it has been demonstrated that a maser with reasonable gain at 20 MHz bandwidth can be designed.

Several novel microwave sources have been developed, including a relatively cheap and simple microwave source with an output of 10 mW which uses a step recovery diode driven by a 100 mW UHF transistor. A solid-state 50 mW 30 GHz source, suitable for a local oscillator or for pumping a parametric amplifier, consists of a transistor oscillator at 500 MHz

(2 watts) driving three cascaded quadruplers.

A tunable (10.7-11.7 GHz) tunnel diode amplifier with 100 MHz instantaneous bandwidth at 15 dB gain and only 6 dB noise figure is useful for microwave links.

One of the main areas of development and research has been in parametric amplifiers using varactor diodes. These amplifiers are described in some detail in the present issue^[18].

The importance of this field lies in the fact that these new microwave devices will tend to replace low and medium power microwave tubes in many applications.

A field in which a team at Mullard Research Laboratories has made quite unique contributions to the state of the art is that of *Broadband Microwave Receivers*^[19]. For many years, microwave receiving techniques have been studied under Government and Company sponsorship. The main theme of this work has been the creation of novel receiver systems for the detection and measurement of signals in frequency ranges exceeding one octave. With techniques such as the microwave phase discriminator and the homodyne, highly sensitive receivers have been designed to operate without tuning and to provide automatic measurement of signal characteristics such as frequency and angle of arrival. Special microwave components^[20] are needed and the close association of component design, data handling technique and system study has greatly benefited the work. For example, by applying digital methods to microwave measurement, a world lead has been established in certain bearing and frequency measuring equipments^[21], and some of the earlier developments have already resulted in considerable quantity production of these equipments.

Conclusion

The aim of an industrial research laboratory must be innovation, to allow the Company progressively to enter new product fields and to adopt new techniques successfully. In this process of innovation, understanding of fundamentals is as essential as competent experimental work; and the laboratory team while always welcoming work in new and unproven fields must also be prepared to carry promising work to the proving stage, where development laboratories can take it

[17] P. Saraga, J. A. Weaver and D. J. Woollons, Optical character recognition, page 197 of this number.

[18] C. S. Aitchison, Low noise parametric amplifiers, page 204 of this number.

[19] A. J. Lambell, A tapped delay line compression network for linear F.M. signals, Conference on delay devices for pulse compression radar, IEE Conference publication No. 20, February 1966.

[20] S. J. Robinson and P. T. Saaler, A survey of coaxial and strip-line microwave components, page 211 of this number.

[21] R. N. Alcock, A digital direction finder, page 226 of this number.

over with confidence. Equally important is selection of topics and communication of progress and results; this demands continuous contact with the Laboratories' counterparts in the Company's industrial divisions, with laboratories in the Universities, in the Government establishments, in associated and other companies. It has been the aim in Mullard Research Laboratories to foster these contacts, and to use them all the time for monitoring the purpose and usefulness of the laboratory programme. Ultimately however, no external check can take the place of the determination

from within the Laboratories to improve continuously in discrimination, perception and quality of work. These are the aspirations towards which we strive.

Summary. Introductory paper for issue devoted entirely to Mullard Research Laboratories. It describes the historical growth of the Laboratories and the range of activities. It continues with a review of the present functions of the Laboratories, relationships with the industrial activities of the Mullard Company and also with outside scientific and technical activities, particularly those sponsored by the British Government. It concludes with a mention of some special topics, including some not covered elsewhere in the range of papers.

Single crystal research on ferrites and garnets

F. W. Harrison, R. F. Pearson and K. Tweedale

Introduction

It was decided in 1954 that the newly formed Solid State Physics Division at Mullard Research Laboratories should add a programme of work on magnetic materials to the semiconductor studies which it had already begun.

Ferrites, oxide materials combining strong magnetic properties with high resistivity, which play an important role in the electronics industry, had been invented and developed mainly by J. L. Snoek at the Philips Research Laboratories in Eindhoven during the late 1930's and early 1940's. By 1954 ferrite components had been in production for some years and new applications far beyond those originally imagined had arisen.

The work begun in Eindhoven had been followed by extensive research on the properties of the new materials but there was still a number of fundamental phenomena which required to be understood and which could best be studied if the materials were available in single crystal form, rather than having the polycrystalline nature of the commercial ceramic product. It was in this field that it was felt a contribution could be made at Mullard Research Laboratories and accordingly a team was assembled.

Single crystal preparation

The preparation of single crystals, an essential foundation for the programme, was undertaken initially by means of the Verneuil or flame fusion method (*fig. 1*) in which an oxy-coal gas or oxy-hydrogen flame from a vertical burner provides the temperature higher than 1500 °C necessary to melt the raw material powders which are arranged to fall through the flame and crystallize on a refractory pedestal at the tip of the flame. Progressive crystallization occurs as the refractory pedestal is lowered uniformly to maintain the growing top of the crystal at a constant position in the flame, and a column of single crystal is built up. This method was used to produce manganese ferrite (*fig. 2*) and a range of compositions based on it with additional manganese, iron or cobalt-substitution.

The Bridgman method, in which a conically-tipped crucible containing the melt is lowered slowly through a temperature gradient in a vertical electric furnace so that crystallization begins at the pointed tip, was used in the growth of magnetite, ferrous ferrite, and magnetite incorporating gallium or aluminium (*fig. 3*). All these compositions required an atmosphere of carbon dioxide at the melting point for chemical equilibrium in view of the large ferrous iron content. The crucibles were made of a platinum-rhodium alloy.

Rise in interest in magnetic materials with garnet structure, e.g. $R_3^3+Fe_5^3+O_{12}$ where R is a rare earth

F. W. Harrison, Ph. D., R. F. Pearson, Ph. D., and K. Tweedale, B.A., are with Mullard Research Laboratories, Redhill, Surrey, England.

over with confidence. Equally important is selection of topics and communication of progress and results; this demands continuous contact with the Laboratories' counterparts in the Company's industrial divisions, with laboratories in the Universities, in the Government establishments, in associated and other companies. It has been the aim in Mullard Research Laboratories to foster these contacts, and to use them all the time for monitoring the purpose and usefulness of the laboratory programme. Ultimately however, no external check can take the place of the determination

from within the Laboratories to improve continuously in discrimination, perception and quality of work. These are the aspirations towards which we strive.

Summary. Introductory paper for issue devoted entirely to Mullard Research Laboratories. It describes the historical growth of the Laboratories and the range of activities. It continues with a review of the present functions of the Laboratories, relationships with the industrial activities of the Mullard Company and also with outside scientific and technical activities, particularly those sponsored by the British Government. It concludes with a mention of some special topics, including some not covered elsewhere in the range of papers.

Single crystal research on ferrites and garnets

F. W. Harrison, R. F. Pearson and K. Tweedale

Introduction

It was decided in 1954 that the newly formed Solid State Physics Division at Mullard Research Laboratories should add a programme of work on magnetic materials to the semiconductor studies which it had already begun.

Ferrites, oxide materials combining strong magnetic properties with high resistivity, which play an important role in the electronics industry, had been invented and developed mainly by J. L. Snoek at the Philips Research Laboratories in Eindhoven during the late 1930's and early 1940's. By 1954 ferrite components had been in production for some years and new applications far beyond those originally imagined had arisen.

The work begun in Eindhoven had been followed by extensive research on the properties of the new materials but there was still a number of fundamental phenomena which required to be understood and which could best be studied if the materials were available in single crystal form, rather than having the polycrystalline nature of the commercial ceramic product. It was in this field that it was felt a contribution could be made at Mullard Research Laboratories and accordingly a team was assembled.

Single crystal preparation

The preparation of single crystals, an essential foundation for the programme, was undertaken initially by means of the Verneuil or flame fusion method (*fig. 1*) in which an oxy-coal gas or oxy-hydrogen flame from a vertical burner provides the temperature higher than 1500 °C necessary to melt the raw material powders which are arranged to fall through the flame and crystallize on a refractory pedestal at the tip of the flame. Progressive crystallization occurs as the refractory pedestal is lowered uniformly to maintain the growing top of the crystal at a constant position in the flame, and a column of single crystal is built up. This method was used to produce manganese ferrite (*fig. 2*) and a range of compositions based on it with additional manganese, iron or cobalt-substitution.

The Bridgman method, in which a conically-tipped crucible containing the melt is lowered slowly through a temperature gradient in a vertical electric furnace so that crystallization begins at the pointed tip, was used in the growth of magnetite, ferrous ferrite, and magnetite incorporating gallium or aluminium (*fig. 3*). All these compositions required an atmosphere of carbon dioxide at the melting point for chemical equilibrium in view of the large ferrous iron content. The crucibles were made of a platinum-rhodium alloy.

Rise in interest in magnetic materials with garnet structure, e.g. $R_3^3+Fe_5^3+O_{12}$ where R is a rare earth

F. W. Harrison, Ph. D., R. F. Pearson, Ph. D., and K. Tweedale, B.A., are with Mullard Research Laboratories, Redhill, Surrey, England.

ion or yttrium, resulted in the use of the method of growth from solution, particularly suitable for such compounds, based on work originating at the Bell Laboratories. The slow cooling of solutions containing lead oxide or lead oxide-lead fluoride as solvents was used to produce a wide range of rare earth iron garnet single crystals and yttrium iron garnet with various rare earth dopings (fig. 4).

Further details of the application of the above methods in these Laboratories have been given elsewhere [1] [*].

Investigations

It was the aim of the work at the outset to undertake studies which would lead to a more complete understanding and control of some of the fundamental factors affecting technical parameters such as permea-

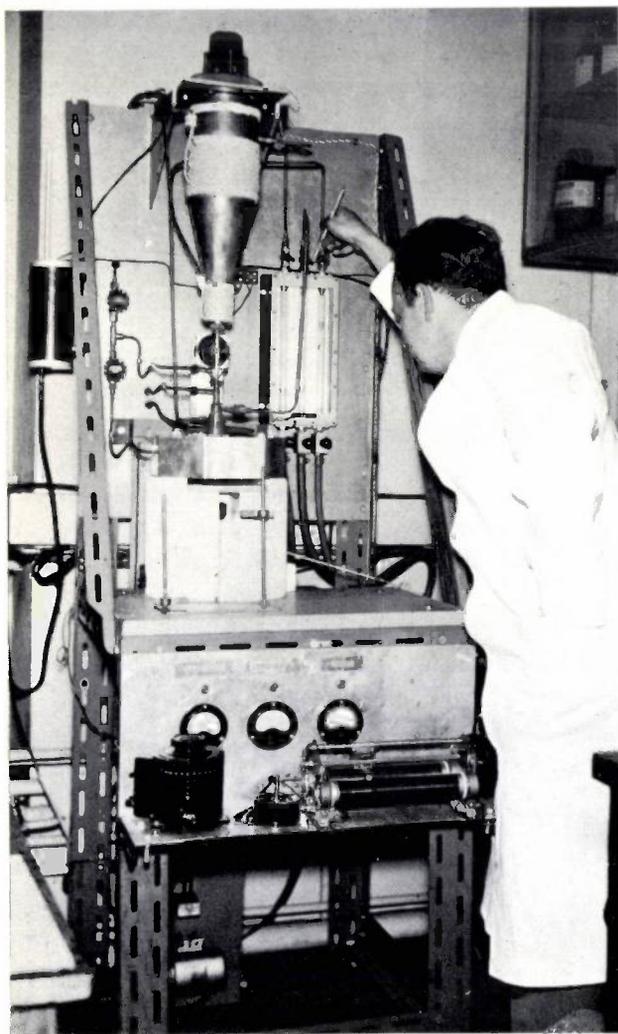


Fig. 1. The flame fusion apparatus used in the production of single crystals of manganese ferrite and associated compositions. The conical hopper contains a vibrating canister which supplies controllable amounts of powder to the vertical burner. The nozzle of the burner is directed into the top of the firebrick furnace, within which the crystal grows.

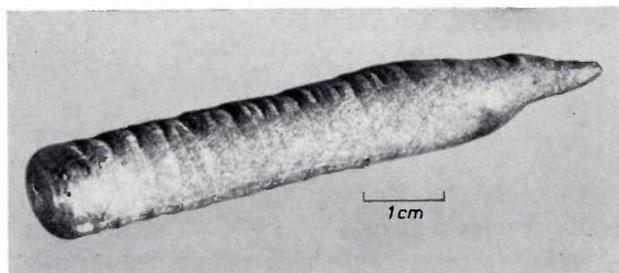


Fig. 2. A manganese ferrite boule grown by the flame fusion method. Growth began with a narrow neck to reduce the probability of more than one nucleus increasing. Surface irregularities result from variations in powder flow, flame conditions and lowering rate during growth.

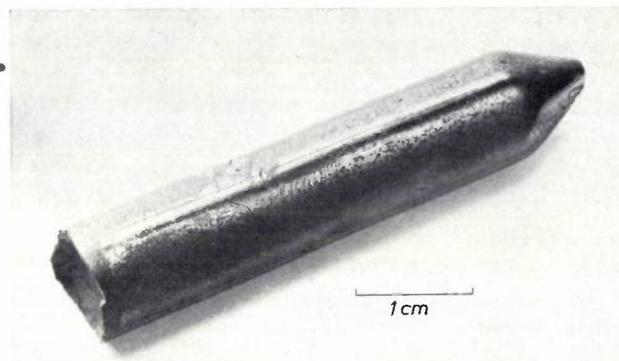


Fig. 3. A crystal of magnetite grown by the Bridgman method. The platinum-rhodium crucible, in which growth from the melt took place, has been stripped off. Growth was initiated at the pointed tip. At the other end the surface of the crystal exposed to the furnace atmosphere reoxidized slightly on cooling from the melting point.

bility, coercive force, hysteresis loop shape and microwave absorption. In a polycrystalline material such parameters will in general be structure-sensitive, affected by the grain size, grain boundaries, porosity and homogeneity as well as by the structure-insensitive properties intrinsic to the crystalline material itself which are dominated by the chemical composition and the type of crystal lattice. Measurements on single crystals enable the structure-insensitive properties to be separated from the structure-sensitive and indeed present the only precise method for studying those properties which depend upon the orientation of the magnetization with respect to the crystal lattice.

A brief survey of some of the more important investigations now follows.

Magneto-crystalline anisotropy

In general the magnetic energy of a ferrimagnetic crystal contains a contribution which depends on the direction of the magnetization with respect to the crystal structure as follows:

$$E = K_1(a_1^2 a_2^2 + a_2^2 a_3^2 + a_3^2 a_1^2) + K_2 a_1^2 a_2^2 a_3^2,$$

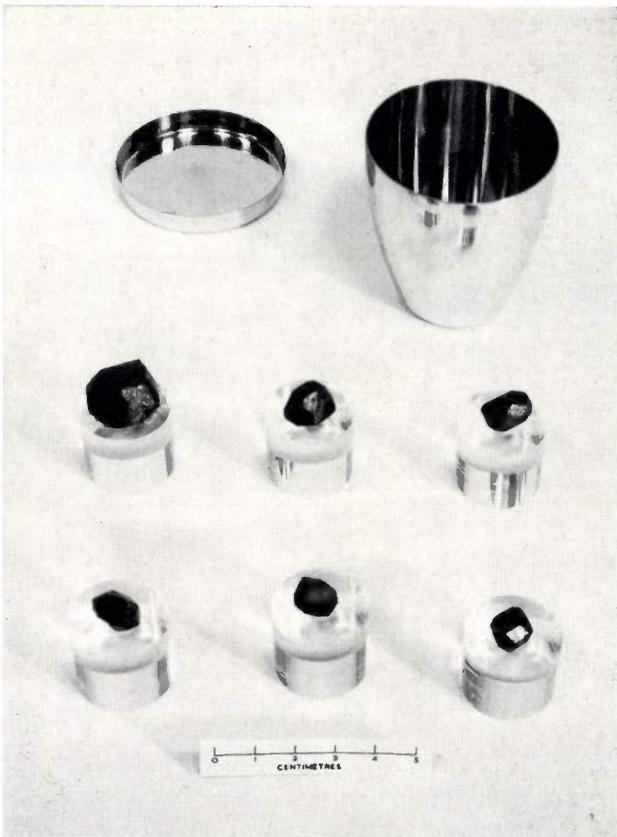


Fig. 4. Garnet crystals grown from lead oxide-lead fluoride solution. The crystals display a dodecahedral habit with (110) and (211) faces. The platinum crucibles used have a tight-fitting re-entrant lid to minimize the loss by evaporation of solvent during growth.

where K_1 and K_2 are known as anisotropy constants and a_1, a_2, a_3 are the direction cosines of the magnetization relative to the structure. Placed in a magnetic field the crystal will experience a torque tending to orient it in such a way that E is a minimum, the field and hence the magnetization lying in a so-called easy direction. In our work K_1 and K_2 were derived from torque measurements on single crystals using an automatic recording torque magnetometer. This instrument plots the torque as a function of orientation with respect to field of small spherical samples (up to 1.5 mm diameter) at temperatures between 1 and 500 °K in magnetic fields up to 15 kOe.

Results on cobalt-substituted manganese ferrite crystals^[2] showed that the temperature at which K_1 went to zero in these compositions could be correlated with the temperature at which the equivalent polycrystalline composition showed an initial permeability peak^[3] (see fig. 5). The cobalt contribution to K_1 varied linearly with cobalt composition, as shown in fig. 6, indicating that a single ion theory of anisotropy

applied. Such a theory, proposed by Slonczewski to explain the case of Co^{2+} in Fe_3O_4 , was found to need modification in detail for the case of Co^{2+} in MnFe_2O_4 . The effect of Co^{2+} substituted in manganese ferrous ferrite was also investigated, as well as cobalt-free compositions with varying ferrous ion content. This work confirmed that Fe^{2+} can give a positive contribution to the anisotropy^[4] and favourably affects the permeability of certain compositions.

An extensive series of measurements was made on the rare earth iron garnets^[5]. These compounds have very large anisotropy at low temperatures and the anisotropy exhibits large values of K_2 and higher order terms. This is illustrated in the case of samarium-iron garnet (SmIG) in fig. 7 where the anisotropy energy surfaces at 77 °K and a somewhat lower temperature are plotted. In such diagrams the length of each radius vector from the origin to the surface represents the anisotropy energy when the magnetization lies along that direction. At 77 °K, where $K_2/K_1 = -1.25$, $\langle 111 \rangle$ is the easy direction along which the magnetization prefers to lie, since the energy is then a

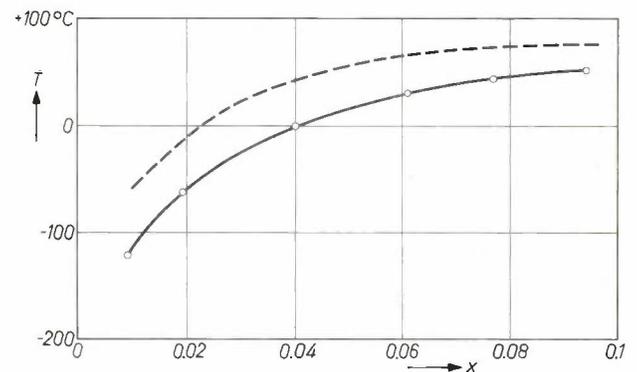


Fig. 5. Temperature at which $K_1 = 0$ in cobalt-substituted manganese ferrite crystals plotted as a function of cobalt concentration x (full line). The results are compared with the temperature at which the equivalent polycrystalline compositions show an initial permeability peak (dotted line)^[3] and are seen to follow a similar composition dependence.

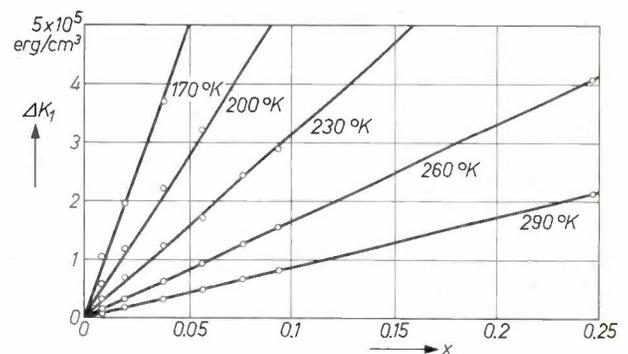


Fig. 6. Cobalt contribution to K_1 in cobalt-substituted manganese ferrite crystals plotted as a function of cobalt concentration x at different temperatures. The linear dependence indicates that the single ion theory of anisotropy is applicable.

[*] The references are found on page 144.

minimum. At the lower temperature K_2/K_1 has changed to -4 and $\langle 110 \rangle$ becomes easy.

As a result of the large anisotropies present in the rare earth iron garnets at low temperatures and the dependence of the spontaneous magnetization on applied

way in which the rapidly changing anisotropy with temperature affects the magnetization in different directions is shown in *fig. 8* where the behaviour of SmIG can be related to the changes in anisotropy constant, demonstrated in *fig. 7*.

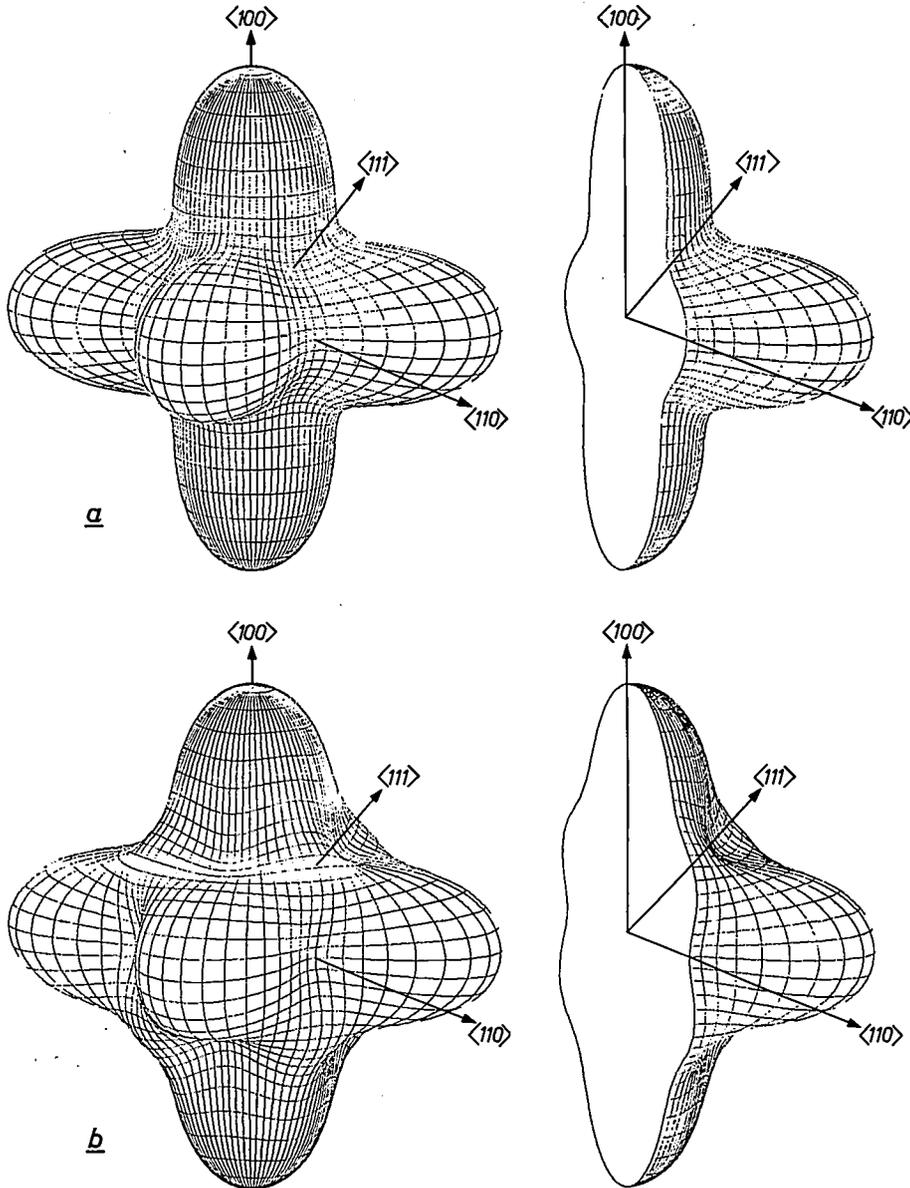


Fig. 7. *a*) Anisotropy energy surface at 77 °K for samarium-iron garnet showing $\langle 111 \rangle$ as the easy direction and $\langle 100 \rangle$ and $\langle 110 \rangle$ as hard. Both the complete surface and a section along a $\langle 110 \rangle$ plane are shown.
b) The same at a lower temperature where $\langle 110 \rangle$ is now easy, and $\langle 111 \rangle$ and $\langle 100 \rangle$ are hard. It must be emphasized that the easy direction is a true minimum and not just a saddle point appearing as a minimum in a particular plane.

field it is impossible to make precise magnetization measurements on polycrystalline samples and in order to know this important parameter in such regions it was necessary to carry out measurements with the applied field in the easy direction [6]. An example of the

Measurements of the field required for ferromagnetic resonance in rare earth doped yttrium iron garnet at helium temperatures showed large peaks along certain crystallographic directions [7]. Torque curve measurements [8] indicated that these peaks were an intrinsic

property of the crystals and could be explained in terms of discontinuities in the energy surface resulting from the crossing over of energy levels in the rare earth ion (fig. 9).

These anisotropy results have proved invaluable in

checking the theory of anisotropy in garnets [9] and also have been used in conjunction with data on dynamic anisotropies derived from high frequency measurements to investigate the relaxation mechanisms in such magnetic systems [10].

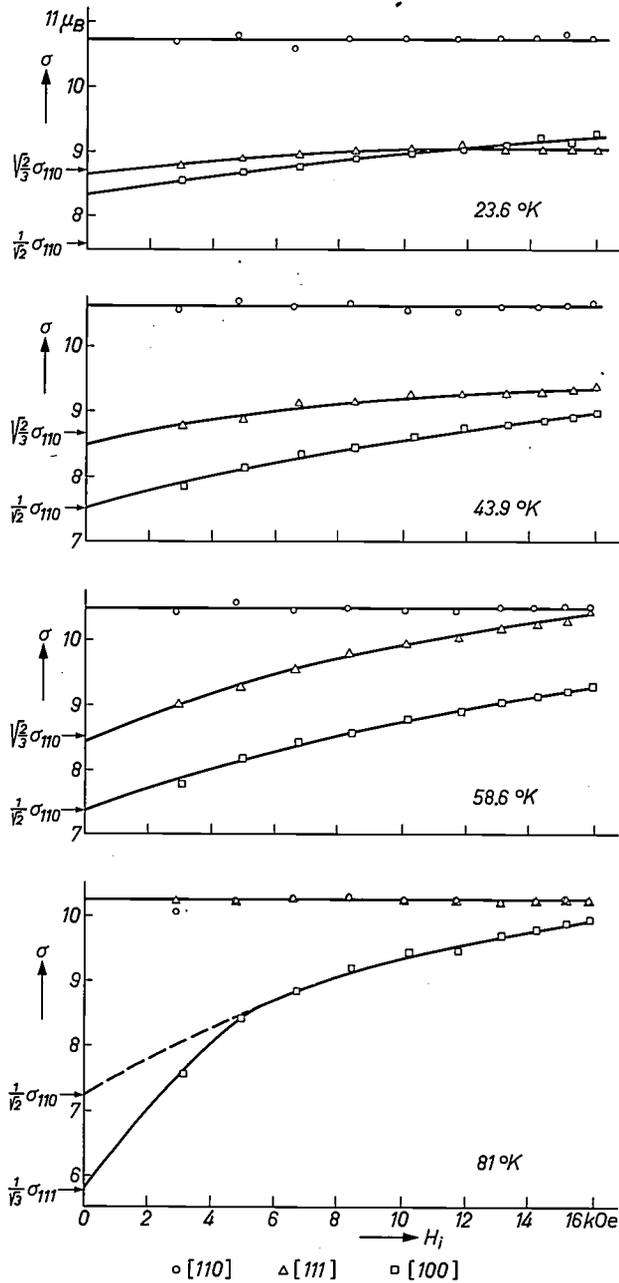


Fig. 8. Magnetization in Bohr magnetons μ_B per formula unit ($5\text{Fe}_2\text{O}_3 \cdot 3\text{Sm}_2\text{O}_3$) versus internal field H_i (applied field minus demagnetizing field) in the three principal crystallographic directions for samarium-iron garnet. [110] is the easy direction at the lower temperatures, the [111] and [100] being unsaturated, the magnetizations extrapolating at $H_i = 0$ to the appropriate values relative to the easy direction. At 23.6 °K [111] is slightly harder than [100] but becomes less hard with rise in temperature and takes over as easy direction between 58.6 and 81 °K. Thus at 81 °K with decreasing internal field in [100] the magnetization vector initially rotates towards [110] but veers toward [111] in the lowest fields. These measurements accord with the anisotropy surfaces shown in fig. 7.

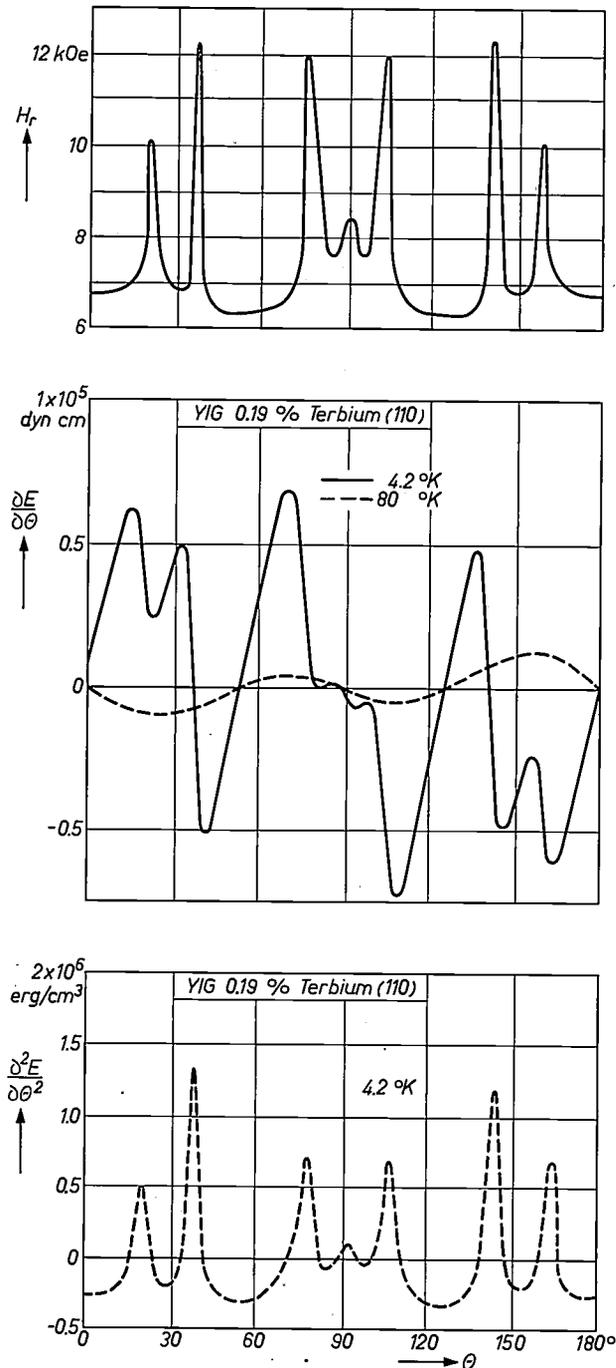


Fig. 9. A comparison of the derivative of the (110) torque curve $\frac{\partial^2 E}{\partial \theta^2}$ for terbium-doped yttrium iron garnet with the ferromagnetic resonance results of Dillon [7] at 4.2 °K. θ is the angle in the (110) plane between the magnetization and a $\langle 100 \rangle$ direction, H_r the field for resonance, and E the magnetic anisotropy energy. The derivative (lower curve) obtained from the torque/cm³ (middle curve), $\frac{\partial E}{\partial \theta}$, reproduces very faithfully the resonance results (upper curve) of Dillon, indicating that the anomalous peaks are independent of frequency.

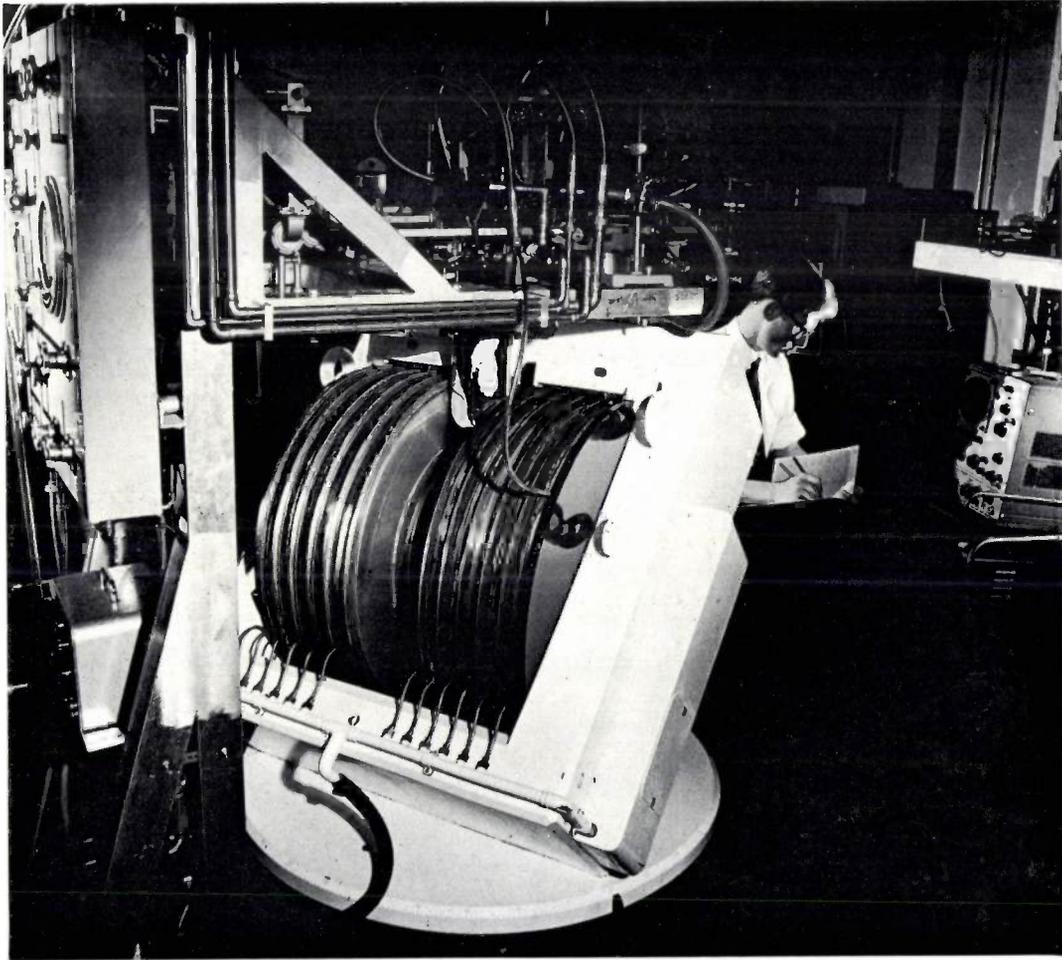
Ferromagnetic resonance

Studies of ferromagnetic resonance were initially aimed at understanding the contribution which an anisotropic ion such as cobalt made to the linewidth. The apparatus used is illustrated in *fig. 10*. The anisotropy studies on the garnets further stimulated investigations of the ferromagnetic resonance behaviour of the rare

Magnetization processes and low field losses

Work on magnetization processes and low field losses originated with a study of the loop shape and switching characteristics of square loop ferrites.

Later, investigations were made to study the origin of the troublesome effect in high quality inductor ferrites known as disaccommodation, in which the permea-



Photograph Walter Nürnberg

Fig. 10. Apparatus for carrying out ferromagnetic resonance studies at microwave frequencies and low temperatures on single crystals of ferrites and garnets.

earth ions, which showed that materials containing these or other anisotropic magnetic ions could give rise to large losses at microwave frequencies in a hitherto unsuspected way. The origin of such losses also manifests itself in the frequency dependence of the observed anisotropy and is further relevant to losses associated with magnetic domain wall motion at lower frequencies. It will be described more fully in a subsequent section.

bility changes with time, or as a result of mechanical shock. Sensitive measurements were made of the change in rotational permeability on a single crystal of magnetite when the orientation of a saturating field is altered [11]. From the angular dependence of this change it was possible to show that the effect arises from the diffusion of cation vacancies located on the octahedral sites of the ferrite structure. By firing the ferrite in such a way

that all the vacancies are filled, the disaccommodation can be minimized.

An unexpected development of this work was the discovery that, due to disaccommodation, domain walls can be made to dig "wells" for themselves of a shape and distribution controlled by an alternating applied field [12]. This presents the possibility of tailor-making loop shapes to order for novel computer storage systems. Thus in *fig. 11* we can see the effect with time of an a.c. field on a minor hysteresis loop of a magnetite toroid initially in the demagnetized state. A practical material would display such behaviour at elevated temperatures so that, for example, a square loop characteristic so formed could be preserved by quenching to room temperature.

above 200 °K, having a maximum value around 20 °K. More detailed measurements were done by Clarke [13] on 5% Yb in YIG and *fig. 12* illustrates this.

The resonance experiments also exhibited any loss mechanism acting in the system through the linewidth of the observed resonance line. *Fig. 13* illustrates the results in the case of Yb-doped YIG for the linewidth. (The same sample was used for the resonance field measurements.) There is a maximum between 100 °K and 150 °K, depending on the frequency of measurement, and a well-defined shoulder to the curve at 20 °K.

The origin of the discrepancy is to be found in the assumption that the measured anisotropy is independent of the frequency of measurement. This was clearly

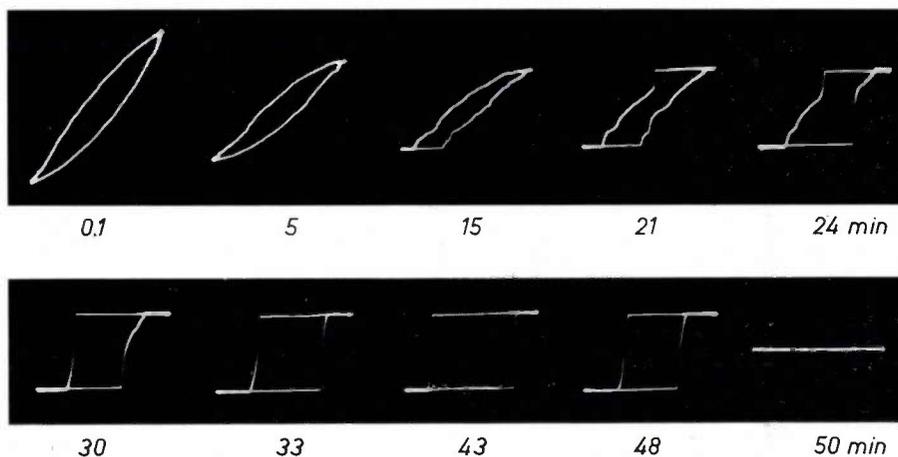


Fig. 11. Effect with time of an a.c. field on a minor hysteresis loop of a magnetite toroid initially in the demagnetized state. As time proceeds potential wells begin to form at the limits of excursion of a domain wall, increasing the length of stay there of the wall. The loop becomes increasingly square until the well is sufficiently deep for the wall to be trapped permanently, when the loop disappears. For each loop the time elapsed after demagnetization is indicated.

Frequency-dependent anisotropy and microwave relaxation

As we have pointed out above, the magneto-crystalline anisotropy was measured by static torque experiments and the ferromagnetic resonance work was also providing information about the anisotropy, not the static anisotropy now, but the anisotropy measured at microwave frequencies, 9 GHz and 16 GHz.

For some materials, and in particular for Yb-doped yttrium iron garnet (YIG), there was a notable discrepancy in the values obtained from the torque measurements and the resonance measurements performed at these Laboratories [10]. The main feature of the discrepancy was that the measured resonance fields were depressed from the ones calculated on the basis of the torque measurements. The so-called dynamic shift was zero at low temperatures (4.2 °K) and at temperatures

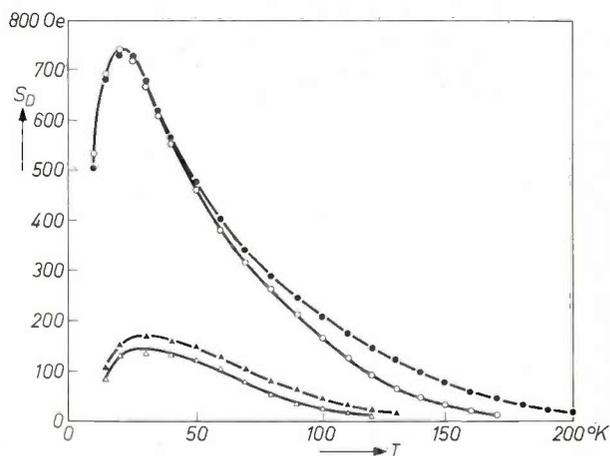


Fig. 12. Dynamic shift S_D of the field value required for resonance for 5% Yb in YIG, versus temperature T , measured in the [111] direction (circles) and [100] direction (triangles) at 9.3 GHz (open symbols) and 16.8 GHz (closed symbols).

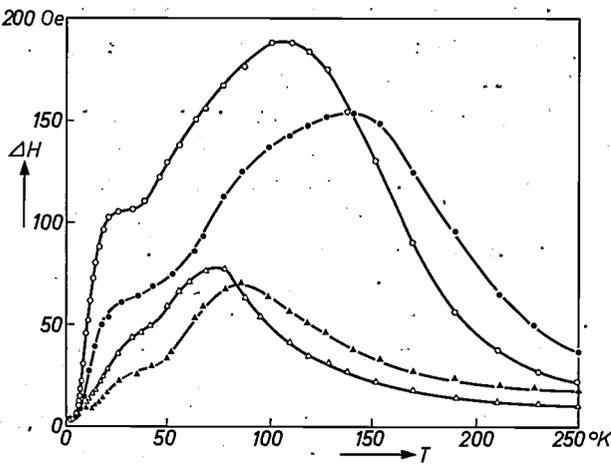


Fig. 13. Linewidth ΔH versus temperature T for 5% Yb in YIG in the [111] direction (circles) and the [100] direction (triangles) at 9.3 GHz (open symbols) and 16.8 GHz (closed symbols).

not so and the theory of the origin of the anisotropy had to be rethought.

Pure YIG has a relatively low anisotropy which can be increased enormously by substituting rare earth ions for some of the yttrium ions. The greatly increased anisotropy is therefore almost wholly due to the substituted ions and the presently accepted theory of anisotropy is the single-ion model in which one calculates the anisotropy from each ion and simply adds the resulting contributions to get the total anisotropy.

Associated with each ion are a number of possible states that it can be in, each one having a different energy: these different energy levels are created by the interaction of the ion with the electric fields produced by its crystalline environment and are modified by the exchange interactions with the neighbouring ferric ions through the intermediate oxygen ions. The separations of these energy levels depend on orientation of the magnetic moment of the ferric ions with respect to the crystalline lattice: these levels are in general anisotropic. This is illustrated in *fig. 14*: for simplicity we shall assume only two levels. θ denotes the orientation of the net ferric ion moment, and hence the magnetization, relative to a specified crystalline direction in some convenient plane, and the energy is shown by E plotted on the vertical axis.

At any given time each ion can only be in one of the two states and the distribution of the ions over the states is determined by the temperature. At low temperatures (low means such that $kT \ll \Delta E$) most of the ions are in the ground state, ε_1 , whilst at high temperatures ($kT \gg \Delta E$) there would be almost equal numbers in states ε_1 and ε_2 . The distribution of ions between the two states is given by the Boltzmann distribution law.

The magneto-crystalline free energy is then given by

adding the free energies from each ion, and because the levels are anisotropic the total free energy can be anisotropic; the temperature dependence of the anisotropy has its origin in the changing distributions of n_1 and n_2 between the two levels at varying temperatures.

There is one other parameter that is relevant to the present discussion: the relaxation time of the system. If we assume that it is somehow possible to disturb the populations, n_1 and n_2 , from their thermodynamic equilibrium values, the system will take a finite time τ to re-establish equilibrium. This time is a measure of the strength of the interaction of the ions with the surrounding crystal. This interaction is usually by means of phonons, but in a ferrimagnetic crystal spin waves, or magnons, may provide the necessary interaction. τ is typically of the order 10^{-10} s.

Let us now consider the two types of measurement mentioned above, the torque method and the experiment involving ferromagnetic resonance.

In a torque experiment the magnetization is moved relatively slowly, i.e. $\tau d\theta/dt \ll 1$, and so the system has time to re-establish equilibrium at each angle. Thus the resulting anisotropy is called the static anisotropy.

However, in a microwave resonance experiment the magnetic moment is oscillating about some mean position at a frequency ω , and it may be that $\omega\tau \geq 1$. In this case the system does not have time to establish thermodynamic equilibrium at any angle and the populations n_1 and n_2 differ from the Boltzmann distribution.

If the resonance experiment is being done at a point such as θ_2 in *fig. 14*, then to first order in $\Delta\theta$ (the precession angle of the moment) the energy separation does not change and so the populations of the two

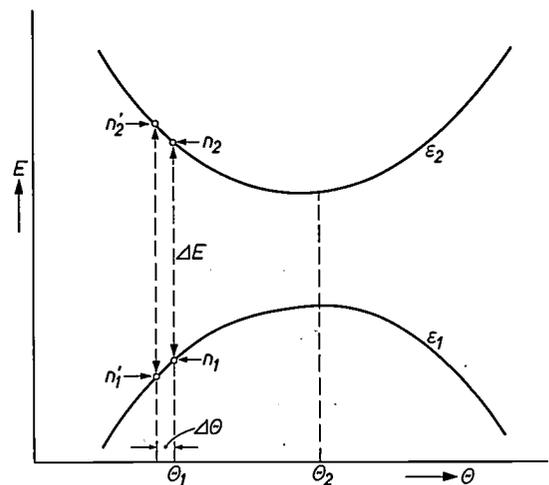


Fig. 14. Dependence of the rare earth ion energy levels E on the crystallographic orientation θ of the magnetization.

levels, n_1 and n_2 , do not need to change as a function of Θ to maintain equilibrium. At points such as these, the static and dynamic experiments will agree. However, if we go to a point such as Θ_1 , where there is a definite gradient of the energy levels, then n_1 and n_2 will need to change (to n_1' and n_2') as Θ is varied about the equilibrium position and, as usual, because of this phase lag, there will be a resultant energy loss.

The resulting expression for the linewidth ΔH of the resonance field because of this frequency dependent anisotropy is [14]:

$$\Delta H \propto \frac{\omega\tau}{1 + (\omega\tau)^2} \frac{1}{kT \cosh^2 \frac{\Delta E}{2kT}} \left[\left(\frac{\partial \Delta E}{\partial \Theta} \right)^2 + \left(\frac{\partial \Delta E}{\partial \varphi} \right)^2 \right]. \quad (1)$$

The first factor on the right of equation (1) shows the dependence of the linewidth on the frequency and relaxation time. This gives a maximum at $\omega\tau = 1$ and because τ is a function of temperature there is a temperature dependence implicit in this factor. This accounts for the main peak in fig. 13.

The second factor is the explicit temperature dependence which arises from the thermodynamic distribution between levels ϵ_1 and ϵ_2 separated by the energy gap ΔE . The shape of the function is shown in fig. 15. At low temperatures ($kT \ll \Delta E$) the function is zero because all ions are in the ground state and a slight change in ΔE does not affect the distribution significantly. At high temperatures ($kT \gg \Delta E$) the function again approaches zero (as $1/T$). This is because the two levels are nearly equally populated and redistribution does not mean very much. The shoulder on the linewidth in fig. 13 is due to this factor.

The last factor shows the dependence on the shape of the energy levels. (Because the moment precesses in a resonance experiment and so does not stay in one plane, we have to invoke two angular variations in planes perpendicular to one another.)

The above treatment (an adaptation of an earlier suggestion by Clogston [15]) explains the apparent discrepancy between the static and resonance results for Yb in YIG. (In this system the energy levels were known from optical measurements and so a direct verification could be made.)

The theory, which has general applicability, has also been used to provide information about energy levels of Nd in YIG [16] and to explain the results of microwave experiments in $Mn_xFe_yO_4$ [17]. It has been applied in detail to Eu in YIG [18], to Er in YIG [19], and to Co^{2+} [20].

In the description that we have given above, the dynamic case ($\omega\tau > 1$) was illustrated by a microwave

resonance experimental situation. However, if we consider what happens when a domain wall moves through a ferromagnetic crystal we come across the same loss mechanism. In this case, instead of the moment precessing about some fixed position, as the wall moves past a given point in the crystal the moment at the point changes direction through a large angle rotation. If the domain wall velocity is v and the thickness of the wall is d the effective angular velocity is given, for a 180° wall, as $\omega = \pi v/d$. For typical velocities of 10^3 cm/s and a thickness of 10^{-6} cm, we see that $\omega\tau \approx 1$, for $\tau \approx 10^{-10}$ s. The relaxation mechanism now causes energy loss in the wall and this acts as a damping force on the wall. In the temperature region where the second factor of equation (1) attains its maximum value we can expect that this loss might be the dominant one limiting wall velocity. (In pure YIG the domain wall velocity is limited by a process not entirely understood at present and any effect of anisotropic ions will act together with these unknown effects.)

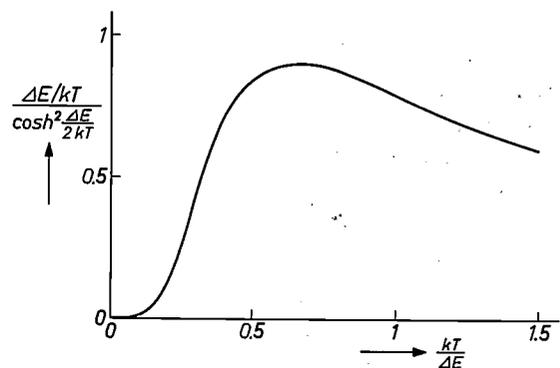


Fig. 15. Plot of the explicit temperature dependent factor in equation (1).

Experiments to measure initial susceptibility also invoke domain wall motion and a detailed application of a similar theory to that for rotational resonance yields frequency-dependent damping constants and also predicts a shift in the resonant frequency of the walls in an exactly parallel way to the microwave resonance measurements.

Another aspect of this model that may be referred to is related to the high power behaviour of ferrites. This behaviour depends on the linewidths associated with spin wave modes, and for spin waves whose wavelength is very much larger than ionic dimensions the same loss mechanism will apply. So anisotropic ions will also influence high power microwave devices using ferrite materials.

Thus, in conclusion, we can see that the loss mechanism we have described can play an important part not only in simple rotational processes at microwave frequencies, but also in spin wave losses and in domain wall processes, both by limiting the switching speed and by producing frequency-dependent losses in the initial susceptibility.

The ions that have been mentioned above, viz, the rare earths, Co^{2+} , Mn^{3+} , and Fe^{2+} are those ions which in general produce a marked effect upon crystalline anisotropy. The theory is believed to have significance for all anisotropic ions in the appropriate frequency range but, of course, in polycrystalline materials other processes may dominate. Fe^{2+} appears in many technologically valuable polycrystalline spinels and although the detailed structures of the energy levels of this ion are not known, the effect of this particular form of energy loss may be important. Another aspect of the Fe^{2+} ion is that the different energy levels may arise from a process of electron migration between different octahedral sites, rather than from the intra-ionic energy level schemes. Co^{2+} affects the anisotropy very strongly and if the relaxation time is of the order ω^{-1} then the model predicts very large linewidths and dynamic shifts: however, preliminary measurements [21] indicate that $\tau \approx 10^{-12}$ s at all temperatures and so only at very high frequencies does the model indicate significant effects. Mn^{3+} has a similar relaxation time at room temperature, but at lower temperatures $\tau = 10^{-10}$ s and large effects now occur in the microwave region.

Single crystal applications

The introduction to this article indicated the aim of studying ferrites in single crystal form more completely to understand and control their technical parameters.

It would, however, be misleading to leave the impression that ferrite and garnet single crystals are only useful as an aid to the production of more precisely designed polycrystalline materials. Indeed single crystals have had for some years, and continue to have, their place in sophisticated applications at microwave frequencies. It is perhaps appropriate in closing to mention that garnet single crystals, which are transparent in the infra-red, may have uses in novel devices for manipulating light and a part of current effort on single crystals at these Laboratories is devoted to this end [22].

Summary. Studies at Mullard Research Laboratories, aimed at achieving a more complete understanding and control of some of the fundamental factors affecting the technical parameters of ferrites and garnets, are described. They include the effect on magneto-crystalline anisotropy of Co^{2+} and Fe^{2+} in spinel ferrites and of rare earth ions in ferrimagnetic garnets. It is shown that the frequency dependence of this anisotropy is the outcome of a relaxation process taking place within the anisotropic ions which

References

- [1] F. W. Harrison, The growth of oxide single crystals containing transition metal ions, *Research* **12**, 395, 1959.
- F. W. Harrison, The growth of ferrite and other magnetic oxide single crystals, *Proc. Brit. Ceramic Soc.* **2**, 91, 1964.
- [2] R. F. Pearson, The magneto-crystalline anisotropy of cobalt-substituted manganese ferrite, *Proc. Phys. Soc.* **74**, 505, 1959.
- [3] C. M. van der Burgt, Controlled crystal anisotropy and controlled temperature dependence of the permeability and elasticity of various cobalt-substituted ferrites, *Philips Res. Repts.* **12**, 97, 1957.
- [4] R. F. Pearson, Magnetic anisotropy in ferrimagnetic crystals, *J. appl. Phys.* **31**, Suppt., 160S, 1960.
- [5] R. F. Pearson, Magnetocrystalline anisotropy of rare-earth iron garnets, *J. appl. Phys.* **33**, Suppt., 1236, 1962.
- [6] F. W. Harrison, J. F. A. Thompson and K. Tweedale, Single crystal magnetization data for rare-earth iron garnets below room temperature, *Proc. Int. Conf. on Magnetism, Nottingham 1964*, p. 664.
- F. W. Harrison, J. F. A. Thompson and G. K. Lang, Single crystal magnetization data for anisotropic rare-earth iron garnets at low temperatures, *J. appl. Phys.* **36**, 1014, 1965.
- [7] J. F. Dillon Jr. and J. W. Nielson, Ferrimagnetic resonance in impurity doped yttrium iron garnet, *J. appl. Phys.* **31**, Suppt., 43S, 1960.
- [8] R. F. Pearson and R. W. Cooper, Torque measurements on rare-earth doped yttrium iron garnet, *J. appl. Phys.* **32**, Suppt., 265S, 1961.
- [9] R. F. Pearson and K. Tweedale, Field dependence of anisotropy in ytterbium-doped yttrium iron garnet, *J. appl. Phys.* **35**, 1061, 1964.
- R. F. Pearson, Magnetocrystalline anisotropy of ytterbium iron garnet $\text{Yb}_3\text{Fe}_5\text{O}_{12}$, *Proc. Phys. Soc.* **86**, 1055, 1965.
- [10] R. W. Teale, R. F. Pearson and M. J. Hight, Ferrimagnetic resonance and torque measurements on ytterbium-substituted yttrium iron garnet, *J. appl. Phys.* **32**, Suppt., 150S, 1961.
- [11] J. E. Knowles, Induced uniaxial anisotropy in magnetite at room temperatures, *Proc. Int. Conf. on Magnetism, Nottingham 1964*, p. 619.
- [12] J. E. Knowles and A. Broese van Groenou, A new manifestation of magnetic after-effect, *Phys. Stat. sol.* **14**, 91, 1966.
- [13] B. H. Clarke, K. Tweedale and R. W. Teale, Rare-earth ion relaxation time and G tensor in rare-earth-doped yttrium iron garnet, I. Ytterbium, *Phys. Rev.* **139**, A 1933, 1965.
- [14] R. W. Teale and K. Tweedale, Ytterbium ion relaxation and ferrimagnetic resonance, *Physics Letters* **1**, 298, 1962.
- [15] A. M. Clogston, Relaxation phenomena in ferrites, *Bell Syst. tech. J.* **34**, 739, 1955.
- [16] B. H. Clarke, Rare-earth ion relaxation time and G tensor in rare-earth-doped yttrium iron garnet, II. Neodymium, *Phys. Rev.* **139**, A 1944, 1965.
- [17] B. H. Clarke, Resonance relaxation in $\text{Mn}_x\text{Fe}_y\text{O}_4$ by slow relaxing manganese and ferrous ions, *J. Phys. Chem. Solids* **27**, 353, 1965.
- [18] R. C. Le Craw, W. G. Nilsen, J. P. Remeika and J. H. Van Vleck, Ferrimagnetic relaxation in europium iron garnet, *Phys. Rev. Letters* **11**, 490, 1963.
- [19] B. H. Clarke, R. F. Pearson, R. W. Teale and K. Tweedale, Rare earth ion relaxation in ferrimagnetic resonance, *J. appl. Phys.* **34**, 1269, 1963.
- [20] R. W. Teale and B. H. Clarke, Ferrimagnetic resonance anomalies in cobalt-substituted manganese ferrite crystals, *J. appl. Phys.* **34**, 1248, 1963.
- [21] R. W. Teale, unpublished work.
- [22] R. F. Pearson and R. W. Cooper, Magnetic crystals that manipulate light, *New Scientist* **32**, 92, 1966.

can be relevant both to losses in microwave resonance and domain wall phenomena. The migration of cation vacancies, a comparatively slow process compared with ionic relaxation, is seen to be a cause of disaccommodation and a means of producing square hysteresis loops. Methods of growth of single crystal materials, essential to the above investigations, are mentioned in the article, together with the trend to novel single crystal applications.

Crystals for studies in electronics



This collection of crystals was grown at Mullard Research Laboratories by members of the crystal growth section. Synthetic single crystals are playing an increasingly important role in the electronics industry both as parts of commercial devices and for the investigation of the fundamental properties of materials which may ultimately be used in a polycrystalline form. Crystals of semiconducting materials have been grown and used in transistor manufacture for many years. However, more recently, devices like the maser and the laser have been developed which require crystals of nonconducting materials containing paramagnetic ions as deliberately added impurities. These ions can impart characteristic colours to the crystals — blue for cobalt, red for chromium, yellow for nickel, etc. Most of the crystals in the collection are

colourless when pure. The majority of them were grown for studies on maser systems but some were for experiments on quantum counters and for use as ultrasonic transducers for television delay lines.

Some of the crystals — in general the longer ones — were grown from the melt by the Verneuil or Czochralski techniques. The melting points of these materials range from about 1100 °C to over 2000 °C. The other crystals — the ones with better developed faces — were grown from solution either in lead oxide-lead fluoride mixtures at temperatures in the range 900 °C to 1300 °C or in water at about 50 °C. The collection includes zinc tungstate, lithium niobate, yttrium aluminium garnet, ruby, magnesium aluminium spinel and potassium cobalticyanide.

The physics of maser materials

J. W. Orton, D. H. Paxman and J. C. Walling

Introduction

The three level solid state microwave maser was first proposed by Bloembergen in 1956 [1]. During the following six years there was intense activity in the field culminating in the successful development of travelling wave maser amplifiers for satellite communication systems.

The masers used by the British Post Office in its receiving station at Goonhilly Down, Cornwall, were developed in the Mullard Research Laboratories and have been described previously in Philips Technical Review [2]. The active material employed was ruby (Cr³⁺-doped alumina) but there has been considerable interest in other materials including Cr³⁺-doped zinc tungstate (ZnWO₄) and emerald (Cr³⁺-doped Be₃Al₂Si₆O₁₈) in all of which the Cr³⁺ ions are responsible for the maser action. These ions are usually referred to as the "spin system".

The important material parameters relevant to maser operation are the splittings of the Cr³⁺ ground state levels (which may be adjusted by an applied magnetic field), the probabilities and linewidths of transitions between them induced by a radiation field and the inverted population difference produced by a particular pumping scheme. It is this latter parameter on which we shall concentrate here.

Population inversion

Consider the set of three energy levels shown in fig. 1. In thermal equilibrium their relative populations are governed by the Boltzmann relation:

$$n_i \propto \exp\left(-\frac{E_i}{kT}\right), \dots \dots (1)$$

illustrated by the dashed curve in fig. 1: This distribution is maintained by relaxation processes due to interaction of the spin system with the lattice. Application of pump radiation at the frequency $f_{13} = (E_3 - E_1)/h$ sufficiently intense to saturate the transition $1 \rightarrow 3$ (i.e. making $n_1 = n_3$) results in the population distribution represented by the heavy lines; assuming the population of level 2 remains unchanged it is now apparent that the populations of levels 1 and 2 are inverted with respect to their thermal equilibrium values. For effi-

cient maser action $n_2 - n_1$ should be made as large as possible.

By solving rate equations for the populations n_1 , n_2 and n_3 under steady state conditions it is easy to derive an expression for $n_2 - n_1$ as follows [2]:

$$\frac{n_2 - n_1}{N} = \frac{w_{23} \left[\exp \frac{hf_{23}}{kT} - 1 \right] - w_{12} \left[\exp \frac{hf_{12}}{kT} - 1 \right]}{w_{23} \left[2 + \exp \frac{hf_{23}}{kT} \right] + w_{12} \left[2 \exp \frac{hf_{12}}{kT} + 1 \right]}, \dots (2)$$

where $N = n_1 + n_2 + n_3$ and the w_{ij} are relaxation rates measuring the strength of "spin-lattice coupling" for the transition $i \rightarrow j$. For the frequently occurring case where hf_{12} and $hf_{23} \ll kT$ equation (2) simplifies to:

$$\frac{n_2 - n_1}{N} = \frac{h}{3kT} \frac{w_{23}f_{23} - w_{12}f_{12}}{w_{23} + w_{12}}, \dots (3)$$

revealing the dependence of $n_2 - n_1$ on the f_{ij} and w_{ij} in a very simple form. It is possible to increase the inversion by selecting a material with large zero-field

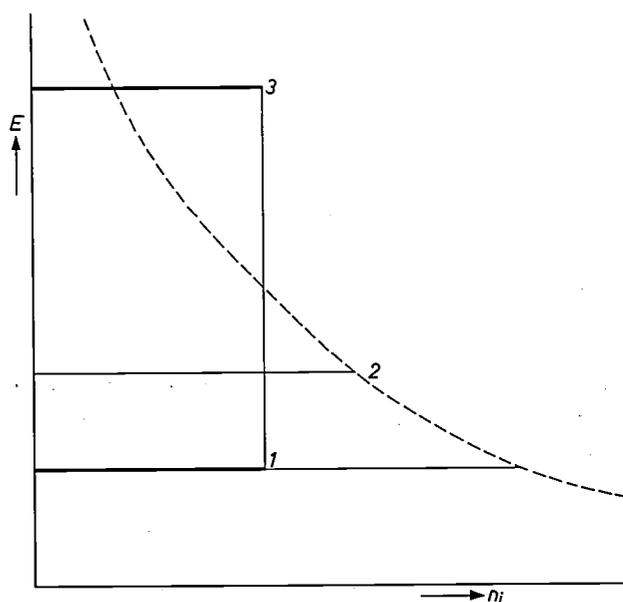


Fig. 1. Diagram of energy E against population n_i showing the inversion of populations 1 and 2 as a result of a pump saturating the transition $1 \rightarrow 3$.

J. W. Orton, D. Phil., D. H. Paxman, M. A., and J. C. Walling, Ph. D., are with Mullard Research Laboratories, Redhill, Surrey, England.

splitting, thus making $f_{23} \gg f_{12}$ but the behaviour of the w_{ij} is less obvious. In the next section we shall discuss this spin-lattice coupling in more detail.

The inverted population difference is given by equation (2) only in the approximation that interactions between paramagnetic ions may be neglected, i.e. assuming a fairly low concentration of these ions. At concentrations employed in practical maser materials it is necessary also to consider the phenomenon of cross relaxation which arises through magnetic dipole coupling between neighbouring magnetic ions.

Fig. 2 illustrates a particular type of cross relaxation (known as harmonic cross relaxation) which may occur when two of the energy separations $E_i - E_j$ bear a

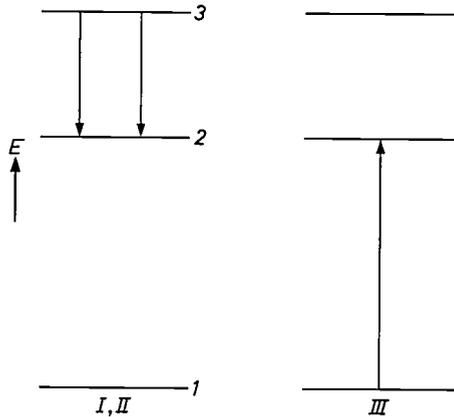


Fig. 2. Diagram to illustrate a harmonic cross relaxation process in a three level system. Separations between levels are in the ratio $(E_2 - E_1) : (E_3 - E_2) = 2 : 1$ so that ions *I* and *II* make a downward transition while ion *III* makes an upward transition in which energy is conserved.

simple harmonic relation to one another, e.g. in this case $f_{12} = 2f_{23}$. In the process illustrated, two ions *I* and *II*, originally in level 3 make a transition to level 2 while a third ion *III*, originally in level 1 is raised also to level 2. Because of the 2 : 1 frequency ratio, energy is obviously conserved within the spin system as a whole.

The significance of such processes is that they redistribute ions over the available levels and may therefore effect considerable changes in the inversion obtainable with any particular pumping scheme. Their importance may be measured by a cross relaxation rate w_{cr} which is to be compared with the spin-lattice rates w_{ij} .

In these Laboratories we have investigated 2 : 1 harmonic cross relaxation in ruby and emerald [3] [4]. However, we shall not discuss this work in detail but concentrate on spin-lattice relaxation.

Spin-lattice relaxation

In the case where the paramagnetic ion has a ground state consisting of only two levels, it is possible to define a single time constant τ_1 (the spin-lattice relaxation time) describing the approach to thermal equilibrium following a disturbance of the populations n_1 and n_2 . Thus:

$$n_1 - n_2 = \Delta n = \Delta n_0 \left[1 - \exp\left(-\frac{t}{\tau_1}\right) \right], \quad (4)$$

where

$$\tau_1 = \frac{1}{w_{12} + w_{21}} = \frac{1}{w_{12} \left[1 + \exp\left(\frac{hf_{12}}{kT}\right) \right]}. \quad (5)$$

τ_1 may be measured by saturating the transition $1 \rightarrow 2$ with a pulse of power from a magnetron or high power klystron and measuring the recovery with a paramagnetic resonance spectrometer. In fig. 3 are shown some

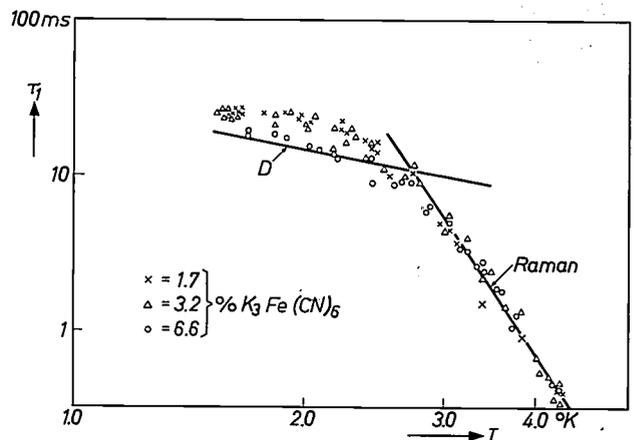


Fig. 3. Spin-lattice relaxation time τ_1 for the ion Fe^{3+} in single crystals of $\text{K}_3\text{Co}(\text{CN})_6$ as a function of temperature T . These results illustrate the occurrence of direct (*D*) and indirect (Raman) relaxation processes for a simple two level system.

results obtained in these Laboratories [5] on Fe^{3+} ions in $\text{K}_3\text{Co}(\text{CN})_6$ which show a sharp change in temperature dependence at approximately 2.7 °K. This may be interpreted as illustrating a change in the dominant relaxation mechanism. At low temperatures τ_1 is determined by the direct process interaction where the spin system exchanges a single phonon with the lattice

[1] N. Bloembergen, Phys. Rev. 104, 324-327, 1956.

[2] J. C. Walling and F. W. Smith, Solid state masers and their use in satellite communication systems, Philips tech. Rev. 25, 289-310, 1963/64.

[3] P. T. Squire, Proc. Phys. Soc. 86, 573-586, 1965.

[4] P. T. Squire and J. W. Orton, Proc. Phys. Soc. 88, 649-657, 1966.

[5] D. H. Paxman, Proc. Phys. Soc. 78, 180-184, 1961.

at the resonant frequency f_{12} . Theoretically, such a process may be shown to give a temperature dependence $\tau_1 \propto T^{-1}$, as is found experimentally. At higher temperatures the "Raman" process becomes more important. This is a scattering process involving two lattice phonons whose frequency difference is equal to f_{12} and it may be shown that τ_1 depends on temperature much more strongly, i.e. $\tau_1 \propto T^{-9}$ for the case considered here. The rapid increase in spin-lattice coupling at higher temperatures is one reason for operating masers at low temperature as it becomes increasingly difficult to saturate the pump transition as τ_1 decreases.

In a multi-level system such as the Cr^{3+} ground state (having four levels) it is no longer possible to define a unique spin-lattice relaxation time. If a particular transition is saturated at time $t = 0$, the recovery to thermal equilibrium is now given by an expression of the form:

$$\Delta n = \Delta n_0 \left[1 - A_1 \exp\left(-\frac{t}{\tau_1}\right) - A_2 \exp\left(-\frac{t}{\tau_2}\right) - A_3 \exp\left(-\frac{t}{\tau_3}\right) \right], \dots (6)$$

where the time constants τ_i and amplitudes A_i all depend on the w_{ij} . It is apparent that values of the individual w_{ij} 's cannot be obtained from a single relaxation measurement and the only way to make progress is to seek theoretical assistance. Theoretical expressions for w_{ij} may be obtained from the work of Van Vleck [6] assuming the direct process to be operative and these may then be used to predict the relaxation behaviour for any particular case. Precise numerical values are difficult to calculate but relative values suffice to determine maser inversion (see equation 2) and allow predictions of, for example, the temperature dependence of relaxation behaviour.

Fig. 4 shows a set of results obtained on Cr^{3+} -doped zinc tungstate [7]. For this case, where there is a large zero-field splitting, the four ground state levels occur for magnetic fields of a few kilo-oersteds as two closely spaced pairs designated as the lower and upper doublets. Effective relaxation times were measured for both doublets and are recorded as experimental points in fig. 4. The curves represent theoretically predicted times. Each curve is shown as solid and dashed to indicate that on recovery the curve is not quite exponential. There is a fairly satisfactory agreement between theory and experiment which suggests that the calculated values of w_{ij} provide at least a good first approximation. Similar work has been done on Cr^{3+} in emerald [4].

These calculated w_{ij} were used to predict inversion ratios ($I = -\Delta n$ (pump on)/ Δn (pump off)) for a partic-

ular operating point in Cr^{3+} - ZnWO_4 . Comparison with experimental values showed good agreement with regard to temperature dependence but suggested the relative magnitudes of the w_{ij} may be in error by as much as 50%. However, even this degree of concordance is far from discouraging in what is a very difficult enterprise.

A rather similar attempt to predict the relaxation behaviour of Cr^{3+} in ruby has been made by Donoho [8] and independently in these Laboratories [9] but has been far less successful when compared with measurements on many ruby crystals. In only one or two cases do the experimental results show reasonable agreement with theory and there is still much which is not understood. It is something of an irony that this should be so when ruby has proved the most generally useful of practical maser materials.

Summarizing, we may say that in cases where the concentration of paramagnetic ions is low and only two paramagnetic energy levels are involved, current

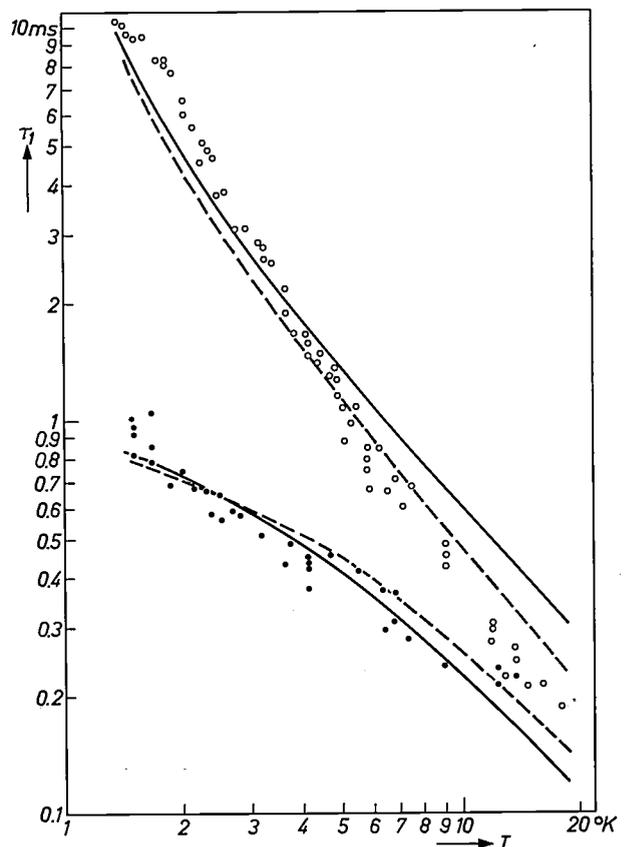


Fig. 4. Temperature dependence of relaxation behaviour. The results shown are those obtained for Cr^{3+} -doped zinc tungstate. In this case, where there is a large zero-field splitting, the four ground state levels occur as two closely spaced pairs, designated as the lower and upper doublets (open circles and black dots respectively). The solid and dashed curves represent theoretical predictions for measurements made at different parts of the recovery curves.

theories of spin-lattice relaxation predict relaxation times in satisfactory agreement with experimental results. In a practical maser material, however, the presence of further levels and increased magnetic ion concentration leads to complicating factors which make theoretical prediction more difficult.

The work discussed here constitutes a useful step

- [6] J. H. van Vleck, *Phys. Rev.* **57**, 426-447, 1940.
 [7] J. W. Orton, A. S. Fruin and J. C. Walling, *Proc. Phys. Soc.* **87**, 703-716, 1966.
 [8] P. L. Donoho, *Phys. Rev.* **133**, A1080-1084, 1964.
 [9] P. T. Squire and J. W. Orton, unpublished.

toward the better understanding of relaxation phenomena in maser materials but a great deal more work will be required before they can be understood completely.

Summary. The performance of a maser material is governed by the population inversion obtainable for a pair of energy levels. This parameter depends on relaxation processes taking place within the material such as spin-lattice and cross relaxation. The agreement between theoretical and experimental studies of spin-lattice relaxation is illustrated for Fe^{3+} in $\text{K}_3\text{Co}(\text{CN})_6$ and Cr^{3+} in ZnWO_4 . Calculated relaxation rates have been used to predict inversion ratios for Cr^{3+} in ZnWO_4 with reasonable success.

Studies of anomalous diffusion of impurities in silicon

K. H. Nicholas

The diffusion of group III and group V elements into silicon is a basic process in the fabrication of silicon planar devices and integrated circuits. These diffusions have shown a number of anomalies which can have important practical consequences. A possible explanation of the anomalies is put forward.

Experimental techniques

The impurity distribution in diffused silicon can be determined as follows [1]. Layers of the diffused silicon parallel to the silicon surface were removed by anodic oxidation and etching off the oxide. The depth of silicon removed was obtained from the interference colour of the oxide. From this depth and the decrease in conductivity of the diffused region the impurity density in the layer of silicon removed was obtained. By subsequent oxidations, etchings, and conductivity measurements the impurity profiles in the diffused regions were obtained, and an example is given in *fig. 1*.

The accuracy of the depth measurement was checked by comparison of the measurements of conductivity obtained in this way at various depths with those of another part of the same slice where the silicon was removed by etching and the depth measured by using an interference microscope.

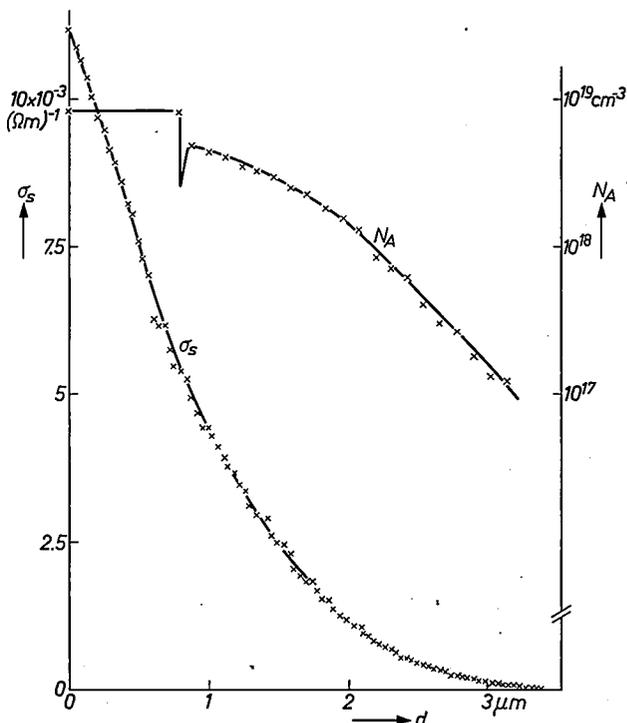


Fig. 1. The production of a plateau of constant concentration by mechanically polishing a slice. The variation of sheet conductivity σ_s and boron concentration N_A with depth d after diffusion at 1200°C for one hour into a mechanically damaged slice is illustrated.

K. H. Nicholas, Ph.D., is with Mullard Research Laboratories, Redhill, Surrey, England.

[1] E. Tannenbaum, *Solid-State Electronics* **2**, 123, 1961; S. Maekawa and T. J. Oshida, *J. Phys. Soc. Japan* **19**, 253, 1964.

theories of spin-lattice relaxation predict relaxation times in satisfactory agreement with experimental results. In a practical maser material, however, the presence of further levels and increased magnetic ion concentration leads to complicating factors which make theoretical prediction more difficult.

The work discussed here constitutes a useful step

- [6] J. H. van Vleck, *Phys. Rev.* **57**, 426-447, 1940.
 [7] J. W. Orton, A. S. Fruin and J. C. Walling, *Proc. Phys. Soc.* **87**, 703-716, 1966.
 [8] P. L. Donoho, *Phys. Rev.* **133**, A1080-1084, 1964.
 [9] P. T. Squire and J. W. Orton, unpublished.

toward the better understanding of relaxation phenomena in maser materials but a great deal more work will be required before they can be understood completely.

Summary. The performance of a maser material is governed by the population inversion obtainable for a pair of energy levels. This parameter depends on relaxation processes taking place within the material such as spin-lattice and cross relaxation. The agreement between theoretical and experimental studies of spin-lattice relaxation is illustrated for Fe^{3+} in $\text{K}_3\text{Co}(\text{CN})_6$ and Cr^{3+} in ZnWO_4 . Calculated relaxation rates have been used to predict inversion ratios for Cr^{3+} in ZnWO_4 with reasonable success.

Studies of anomalous diffusion of impurities in silicon

K. H. Nicholas

The diffusion of group III and group V elements into silicon is a basic process in the fabrication of silicon planar devices and integrated circuits. These diffusions have shown a number of anomalies which can have important practical consequences. A possible explanation of the anomalies is put forward.

Experimental techniques

The impurity distribution in diffused silicon can be determined as follows [1]. Layers of the diffused silicon parallel to the silicon surface were removed by anodic oxidation and etching off the oxide. The depth of silicon removed was obtained from the interference colour of the oxide. From this depth and the decrease in conductivity of the diffused region the impurity density in the layer of silicon removed was obtained. By subsequent oxidations, etchings, and conductivity measurements the impurity profiles in the diffused regions were obtained, and an example is given in *fig. 1*.

The accuracy of the depth measurement was checked by comparison of the measurements of conductivity obtained in this way at various depths with those of another part of the same slice where the silicon was removed by etching and the depth measured by using an interference microscope.

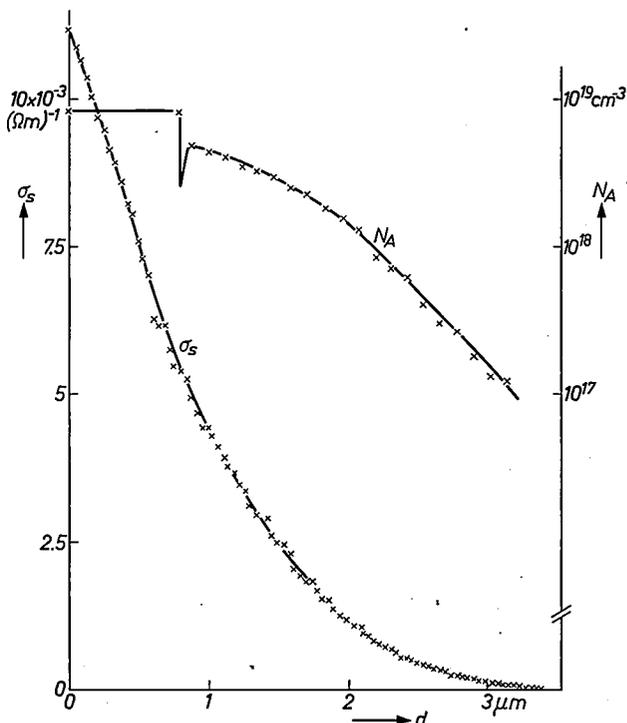


Fig. 1. The production of a plateau of constant concentration by mechanically polishing a slice. The variation of sheet conductivity σ_s and boron concentration N_A with depth d after diffusion at 1200°C for one hour into a mechanically damaged slice is illustrated.

K. H. Nicholas, Ph.D., is with Mullard Research Laboratories, Redhill, Surrey, England.

[1] E. Tannenbaum, *Solid-State Electronics* **2**, 123, 1961; S. Maekawa and T. J. Oshida, *J. Phys. Soc. Japan* **19**, 253, 1964.

Observations of the anomalies

As a result of the diffusion and oxidation processes used in the fabrication of silicon planar devices two diffusion anomalies were observed. Both are associated with the formation of the high concentration emitter region.

1) The diffusion of impurities at high concentrations in this region was much faster than that for the same impurity at lower concentrations and the profile produced was not consistent with a single diffusion coefficient [1] [2]. Fig. 2 gives an example where the line *Theor* represents the distribution obtained using the low concentration value of diffusion constant and the line *Exp* that observed experimentally. At one time it was thought the apparent anomaly might be due to most of the impurity atoms near the surface being electrically inactive but radioactive tracer measurements have confirmed that the fast diffusion does occur [1].

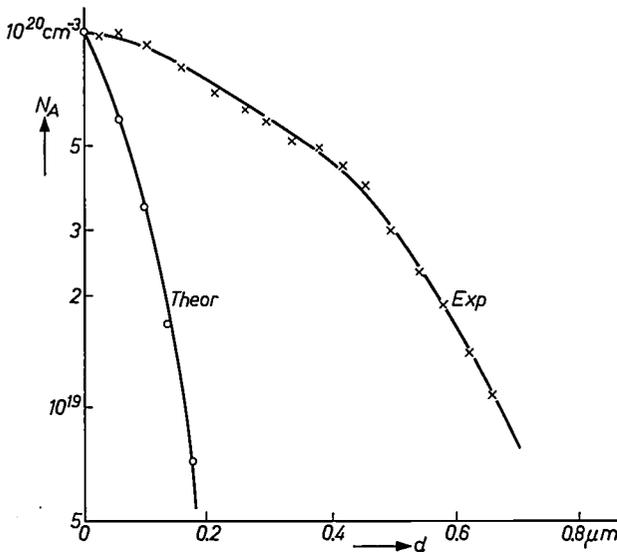


Fig. 2. The fast diffusion of impurities at high concentrations. The diagram shows the variation in boron concentration N_A with depth d after diffusion at 1000 °C for one hour. It shows that the diffusion at high impurity concentrations (curve *Exp*) is much faster than is expected theoretically from low concentration observations (curve *Theor*).

2) As the emitter of a diffused transistor is of smaller area than the base, only part of the base area has the emitter diffused into it. When narrow base width devices, e.g. high frequency devices, were sectioned and stained to show up the P and N type regions it was often found that the base collector junction in front of the emitter had diffused in further than elsewhere. This effect is shown in fig. 3 and is called the push out, snow plough, or emitter dip effect [3].

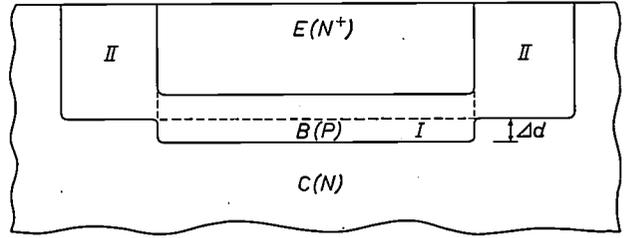


Fig. 3. The cross-section of an N-P-N transistor showing the push out effect. The region I of the boron doped base B(P) which is in front of the heavily phosphorus doped emitter E(N⁺) extends further forward by Δd into the lightly doped collector C(N) than region II.

The actual impurity profiles were measured during these investigations (see fig. 4). The base concentration in front of the emitter shows the effect to be an increase in the diffusion rate in the base in that region rather than rejection of base impurity atoms from the emitter region.

3) A third anomaly was observed during the present work on a mechanically polished crystal. Fig. 1 shows the impurity distribution after diffusion into a silicon slice with a mechanically polished surface. The anomaly is shown by the line N_A in fig. 1 where a region of constant impurity concentration was observed near the surface. The length of the plateau was found to be proportional to the square root of the diffusion time showing it was not just a fixed region of deformed silicon near the surface.

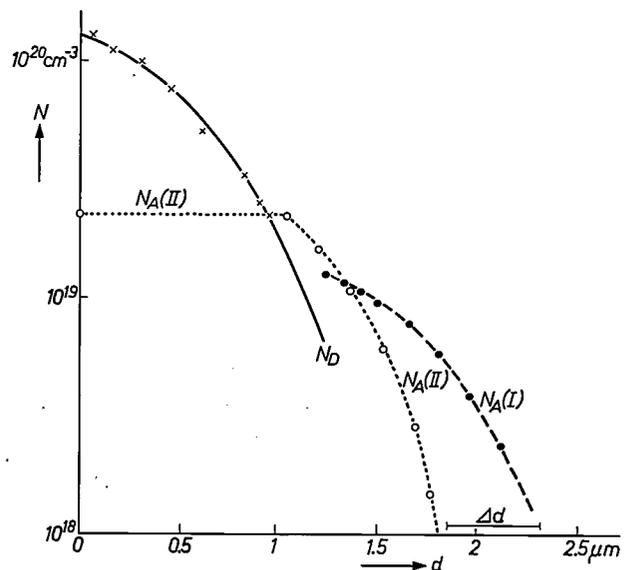


Fig. 4. The impurity profiles (N versus d) due to the push out effect. The impurity profiles in the emitter region (N_D) and in the base regions I and II ($N_A(I)$ and $N_A(II)$) are illustrated together with the extent of the push out Δd .

Experiments

To check whether mechanical polishing was responsible for the third anomaly, slices were etched before diffusion and it was found that the plateau was reduced in length and height, see *fig. 5*, but to remove it completely it was necessary to etch the back of the slice also. This indicated that strain is an important factor in the process.

Fig. 6 shows the effect of wet oxidation on a boron diffused slice. Curve 1 is the profile obtained by diffu-

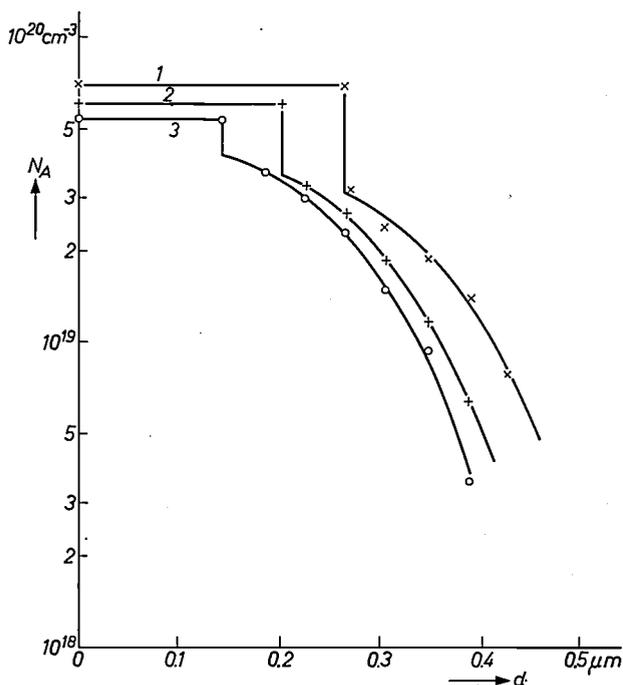


Fig. 5. The reduction of the mechanical polishing anomaly by removal of damaged regions. The plateau of constant concentration observed in a mechanically polished slice after boron diffusion is shown by curve 1. The reduction in the plateau length due to the removal of mechanical damage by etching the slice to a depth of 15 microns before diffusion is shown by curve 2. The further removal of strain by etching the back of the slice to a depth of 80 microns reduces the plateau again as shown in curve 3.

sion into a mechanically polished slice. Curve 2 is the profile obtained when the same diffusion is followed by five minutes wet oxidation at 1000 °C. Clearly the high constant concentration region near the surface has lost boron rapidly during the short period of oxidation, confirming that the plateau is associated with a region of fast diffusion.

If a mechanically polished slice was wet oxidized for a long time (e.g. 2 hours at 1200 °C) and the oxide removed before impurities were diffused into the slice, subsequent diffusions did not show either the mechani-

cally polished anomaly or the rapid loss on oxidation illustrated by curves 1 and 2. The profiles in such a slice are shown by curves 3 and 4 in *fig. 6*.

It is known that wet oxidation after polishing produces dislocations and probably relieves strain thereby^[4]. Thus the removal of the anomaly by a long wet oxidation prior to diffusion shows fast diffusion is not caused by the presence of dislocations.

The growth of dislocations (and hence the relief of strain) is much slower during dry oxidation than during

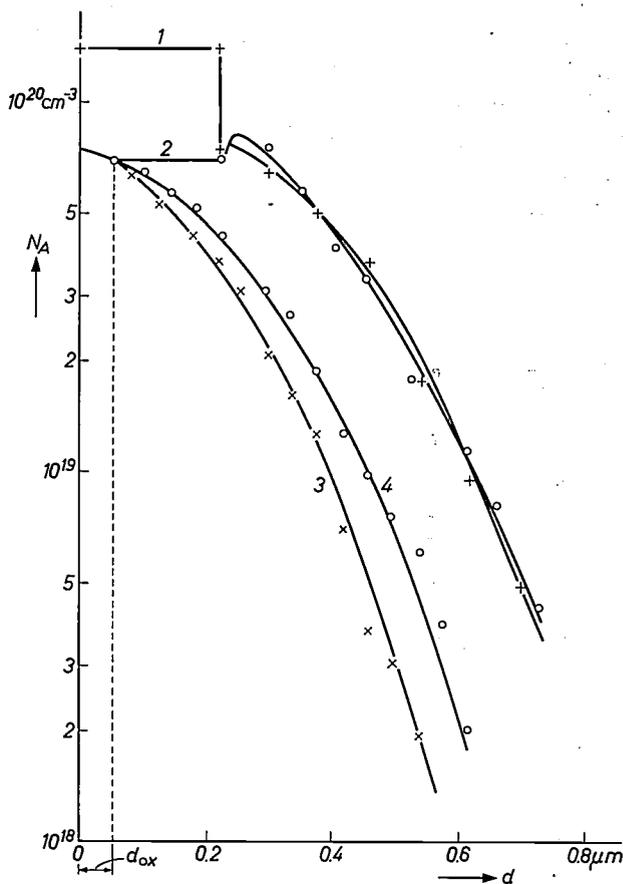


Fig. 6. The effect of a wet oxidation prior to and after boron diffusion. Comparison of curves 1 and 3 shows the removal of the plateau region associated with the mechanical polishing anomaly by wet oxidation of the slice prior to diffusion. The difference between curve 1 and curve 2 represents the rapid loss of boron from a mechanically polished slice when it is wet oxidized after diffusion. Comparison of curves 3 and 4 shows there is little loss of boron on wet oxidation in the absence of mechanical strain. The depth of silicon removed during the wet oxidation after diffusion is d_{ox} .

[2] V. K. Subashiev, A. P. Landsman and A. A. Kukharskii, *Sov. Phys. Solid State* 2, 2406, 1961; P. A. Iles and B. Leibenhaut, *Solid-State Electronics* 5, 331, 1962; E. Kooi, *J. Electrochem. Soc.* 111, 1383, 1964.

[3] P. Baruch, C. Constantin, J. C. Pfister and R. Saintesprit, *Disc. Faraday Soc.* 31, 76, 1961; Y. Sato and H. Arata, *Japanese J. appl. Phys.* 3, 511, 1964; R. Gereth, P. G. G. van Loon and V. Williams, *J. Electrochem. Soc.* 112, 323, 1965.

[4] H. J. Queisser and P. G. G. van Loon, *J. appl. Phys.* 35, 3066, 1964.

wet oxidation. A mechanically polished slice diffused with boron showed fast outward diffusion as was illustrated in fig. 6 but outward diffusion of boron was slow during dry oxidation. The presence of mechanical strain during dry oxidation does not cause very fast diffusion. However, when the strain was being relieved during wet oxidation fast diffusion did occur. This means it must be the relief of strain by dislocations growing rather than the strain itself which produces the mechanical polishing anomaly.

The other anomalies of fast diffusion at high concentration and the push out effect can similarly be explained. In both these cases the strain is caused by the high concentration gradient of impurity atoms in the emitter region instead of by the mechanical polishing damage. The mismatch of these atoms in the silicon lattice distorts the lattice and the production of dislocations relieving this strain has been widely reported [5].

Sato and Arata recently observed dislocations beyond the emitter in the pushed out base region of a transistor structure (see the second article cited in [3]). This was in good agreement with this theory of push out being due to fast diffusion occurring as strain is relieved and dislocations are growing.

To confirm that the relief of strain, not just strain itself, produced the fast diffusion in the high concentration anomaly a deposition of phosphorus was carried out in an inert atmosphere. Separate parts of the slice were then heated in nitrogen and wet oxygen. The phosphorus in those heated in wet oxygen, where strain was fast relieved, diffused about four times as fast as in those heated in nitrogen, confirming relief to strain is the important factor.

This relieving of strain leading to fast diffusion explains why particularly rapid outward diffusion of boron occurs from slices diffused at high concentration when they are subjected to wet oxidation. The resulting profile shows as expected a long plateau at an intermediate concentration level as shown in the base region in fig. 4. The loss of boron is a serious difficulty in making the emitters of *P-N-P* transistors.

Finally, to observe the effect of relief of strain directly,

a slice with a diffused impurity layer was strained mechanically and wet oxidized simultaneously at high temperature. The resultant dislocation density was taken as an indication of how much strain was relieved in a particular region and the results plotted in fig. 7. The

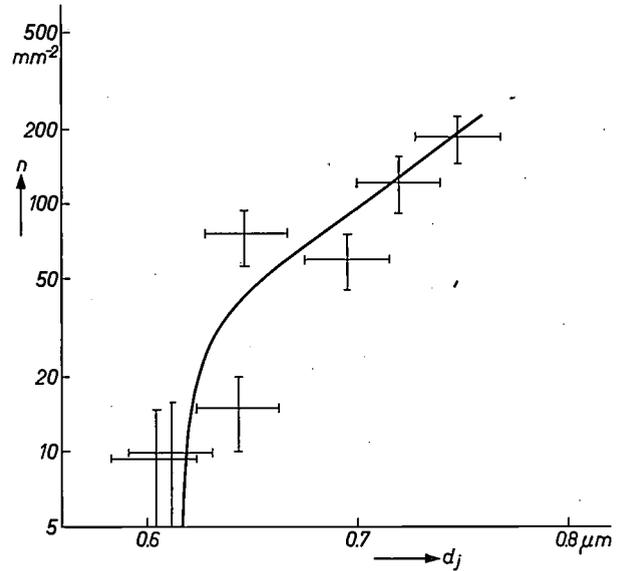


Fig. 7. The increase in junction depth due to relief of strain during boron diffusion. The diagram illustrates the effect of applying strain during the oxidation of boron diffused sample. The junction depth d_j is plotted against the magnitude of strain as represented by the number of dislocations observed per mm^2 area (n).

junction depth was about $0.5 \mu\text{m}$ before the strain was applied so that there is a great increase in the diffusion constant in the strained region.

One practical consequence of these effects that has not received much attention is that wet oxidation, as commonly used in the production of planar devices, has a major effect on impurity distributions.

Summary. The diffusion of impurities into silicon is used in the manufacture of silicon planar devices and integrated circuits. Three anomalies have been observed: fast diffusion in a mechanically polished slice, at high impurity concentrations, and in front of an emitter (the push out effect). They can be explained by fast diffusion during dislocation growth as strain is relieved. The results of experiments supporting this theory are described.

[5] H. J. Queisser, J. appl. Phys. 32, 1776, 1961; S. Prussin, J. appl. Phys. 32, 1876, 1961; H. Strack, J. appl. Phys. 34, 2405, 1963.

Some applications of contour deposition

A. F. Beer, J. B. Coughlin and P. J. Daniel

Recently, details of a new integrated circuit isolation technique have been published [1]. This technique, which is called contour deposition, consists of etching recesses in the surface of the silicon wafer, depositing epitaxial material to a thickness greater than the depth of the recess, and finally polishing and etching down to slightly below the original surface (fig. 1). This leaves isolated regions of the epitaxial material in the original

layers. An epitaxial layer of *N*-type silicon is then deposited on the surface. Boron (an acceptor impurity) is deposited (i.e. diffused shallowly at high concentration) round the buried arsenic layer and in a subsequent operation is driven in through the epitaxial layer. In this way isolated regions of *N*-type silicon are formed, surrounded by *P*-type material. The calculated values of junction capacitance between the isolated

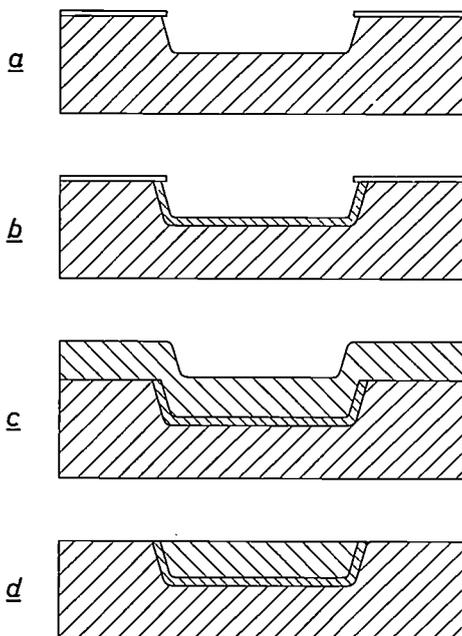


Fig. 1. Contour deposition process. *a*) Recess etched in oxide-masked silicon. *b*) Optional buried layer diffused. *c*) Oxide removed and epitaxial layer deposited. *d*) Surface polished flat and etched. This process leaves isolated regions of the epitaxial material in the original recesses and at the same time preserves a flat surface.

recesses and, at the same time, preserves a flat surface compatible with normal planar technology. If required, a buried layer can easily be formed by a diffusion step prior to the epitaxial deposition. Such buried layers are commonly used in integrated circuits to provide low resistance collector contacts but, in contrast to conventional isolation techniques, this layer comes to the surface at the edges of the isolation area.

A conventional isolation process is shown diagrammatically in fig. 2. A high concentration of arsenic is diffused locally into a *P*-type wafer to produce *N*⁺

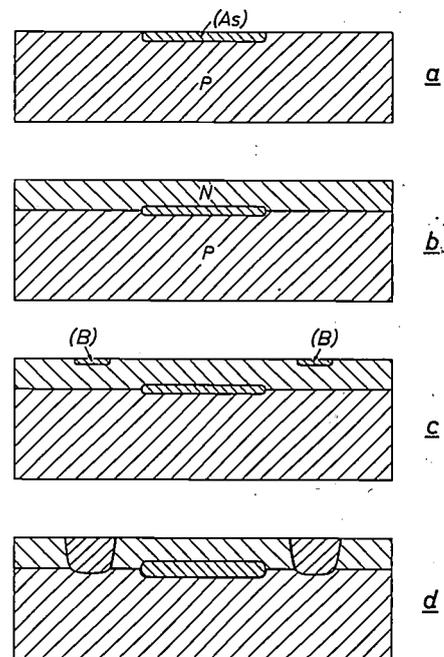


Fig. 2. Conventional isolation technique. *a*) Layers of heavy arsenic doping are diffused locally in a *P*-type wafer. *b*) An epitaxial layer of *N*-type silicon is then deposited on the surface. *c*) Boron is deposited round the buried layer which subsequently is driven in through the epitaxial layer. *d*) Isolation diffusion is completed.

region and the substrate are shown to be up to 4 times lower for practical cases of contour deposited regions than for the conventional process [1].

A number of applications for contour deposition have been suggested and the purpose of this paper is to give some results of the use of the technique in practice. Two applications have received attention:

- 1) the fabrication of complementary pairs of MOS transistors;
- 2) the production of isolated ultra high frequency (UHF) transistors.

A. F. Beer, B.A.Sc. Toronto, J. B. Coughlin, B.Sc., and P. J. Daniel, Ph.D., are with Mullard Research Laboratories, Redhill, Surrey, England.

[1] T. Klein, Solid-State Electronics 9, 959, 1966.

Complementary pairs of MOS transistors

Complementary pairs of *P*-type and *N*-type MOS transistors are useful components in switching circuits since they combine relatively high speed with low power consumption. MOS transistors are best made on relatively lightly doped substrates. To integrate *P*- and *N*-type MOS transistors in a monolithic circuit therefore requires, ideally, a substrate with *P*- and *N*-type regions with acceptor and donor concentrations N_A and N_D of $10^{15}/\text{cm}^3$ respectively.

Contour deposition is unique in being able to achieve this, *N*-type regions being deposited in a *P*-type substrate. In the present case the regions were approximately $110\ \mu\text{m} \times 140\ \mu\text{m}$ in area and $10\ \mu\text{m}$ deep. The *N*-type MOS transistors were made in the *P*-type areas by forming source and drain regions each $120\ \mu\text{m}$ long and $25\ \mu\text{m}$ wide, separated typically by $12\ \mu\text{m}$ before diffusion. The *P*-type MOS transistors were similarly formed by depositing boron in the *N*-type regions.

Drive-in of the dopants was carried out at $1200\ ^\circ\text{C}$ for 1 hour in dry oxygen and this simultaneously formed the gate oxide. The resulting structure is shown diagrammatically in *fig. 3*. Phosphorus glass was applied in order to improve the oxide stability.

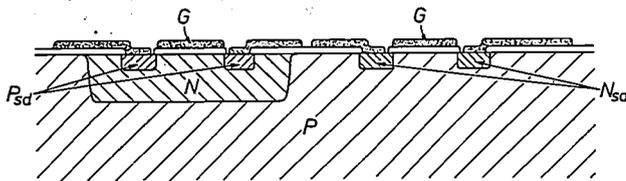


Fig. 3. Section of an integrated complementary pair of MOS transistors. The *P*-type and *N*-type source and drain are indicated by P_{sd} and N_{sd} . G points to the aluminium gate electrodes N shows the *N*-type deposited region. The substrate is *P*-type

In each case two transistors were made with common source to allow them to be connected in a simple complementary bistable circuit. Gold diffusion from the back of the wafer was used to control the threshold voltage V_{T0} (the value of gate voltage at which current can start to flow between source and drain) since it was required that both *P*- and *N*-type MOS transistors should be enhancement types, i.e. the devices should be "off" with zero gate to source voltage. Difficulties were experienced in obtaining reproducible results with this technique but examples of the characteristics of the four MOS transistors T_{N1} , T_{N2} , T_{P1} and T_{P2} of one circuit are shown in *fig. 4*. This shows the source-drain current I_{SD} plotted against drain voltage V_{SD} as a function of the gate voltage V_G , for each of the four devices. Values of the threshold voltage V_{T0} and

the gain parameter β (defined by the equation $I_{SD} = \frac{1}{2}\beta(V_G - V_{T0})^2$) are also shown. The measured characteristics are in agreement with those that would be expected from the dimensions of the devices.

Isolated UHF transistors

Contour deposition has geometrical advantages over conventional isolation techniques which should allow lower collector-substrate capacitances C_{cs} to be obtained. The fact that the buried layer is brought to the surface at the perimeter of the isolation region allows lower collector series resistance to be obtained or a smaller isolation area to be used for the same resistance and, therefore, permits higher operating speeds to be used.

In the present experiments the contour deposited regions were approximately $75\ \mu\text{m} \times 110\ \mu\text{m}$ in area and 8-10 μm deep in 20-40 Ωcm *P*-type substrates. Arsenic diffused buried layers of 50 Ω/square were used. Lower values can be obtained if required, and phosphorus diffusion can sometimes be substituted in spite of its higher diffusion coefficient. The epitaxially deposited silicon was 0.5 to 1.0 Ωcm *N*-type.

Probe tests of the isolation regions have shown that the majority have breakdown voltages in excess of 100 V and the yield of regions with breakdown voltages above 50 V exceeds 95%. Subsequent processing can degrade these voltages by producing conditions susceptible to surface breakdown.

The transistors made in these regions were similar in geometry to a commercial UHF transistor, the Mullard BFY 90, except for having a collector contact on the top surface, *fig. 5*. Processing also was basically similar to the BFY 90 schedule.

The resulting isolated transistors have characteristics very similar to those of the BFY 90 and have not been significantly degraded by the isolation. Typical values from a good wafer are shown in *Table I*.

Table I. Characteristics of resulting isolated transistors.

f_T	(current gain bandwidth product)	> 1100 MHz (2 mA, 5 V)
$N.F.$	(noise figure)	5.5 dB (800 MHz, 2 mA, 10 V)
r_{bb}'	(base resistance)	20-30 Ω
r_{cc}'	(collector series resistance)	100 Ω
C_{cb}	(collector base capacitance)	0.3 pF (10 V)
C_{re}	(feedback capacitance)	0.3 pF (10 V)
C_{cs}	(collector substrate capacitance)	1.4 pF (10 V) including 0.3 pF pad and 0.5 pF header strays
BV_{CBO}	(collector base breakdown voltage)	30 V
BV_{CEO}	(collector emitter breakdown voltage)	20 V
BV_{CSO}	(collector substrate breakdown voltage)	50 V

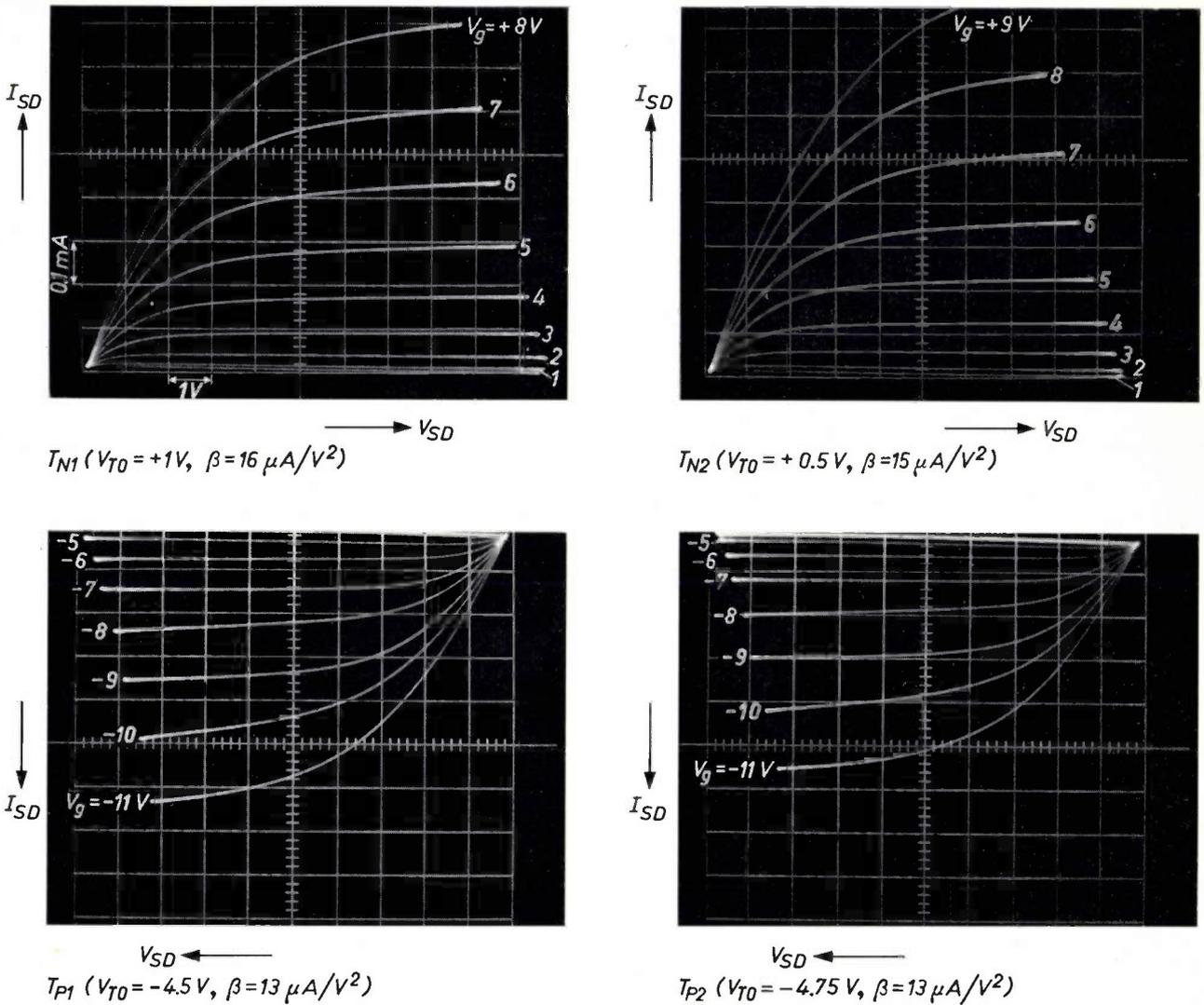


Fig. 4. Characteristics of four MOS transistors T_{N1} , T_{N2} , T_{P1} and T_{P2} of one circuit. Source drain current I_{SD} is plotted against drain voltage V_{SD} as a function of the gate voltage V_g , for each of the four devices. Values of threshold voltage V_{T0} and the gain parameter β are also shown. The measured characteristics are in agreement with those expected from the dimensions of the devices.

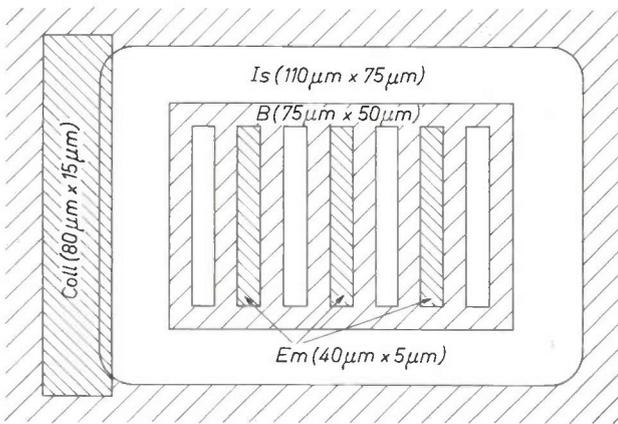


Fig. 5. An isolated UHF transistor. Transistors made in these regions were similar in geometry to the BFY 90 except for having a collector contact *Coll* on the top surface. *Is* isolated region, *B* base area, *Em* emitter stripes.

The collector to substrate isolation capacitance C_{cs} exceeds the calculated value by about 0.3 pF, but this appears to be due to a surface effect associated with the aluminizing process. It is still less than half the value that might be expected from conventional isolation.

The collector series resistance $r_{cc'}$ could be greatly reduced by the use of a buried layer of lower sheet resistance. Conventional isolation, for the same geometry and buried layer, would have an additional resistance of 50-100 Ω due to the fact that the buried layer does not come to the surface. To demonstrate the effect of this extra resistance some special structures were made. These had the same geometry as the contour deposited transistors but were made on normal N on N^+ epitaxial material. The transistors could then be measured using either the substrate or the top contact

as collector terminal. Fig. 6 shows the current gain bandwidth product f_T plotted against collector current I_C for the same transistor with the alternative

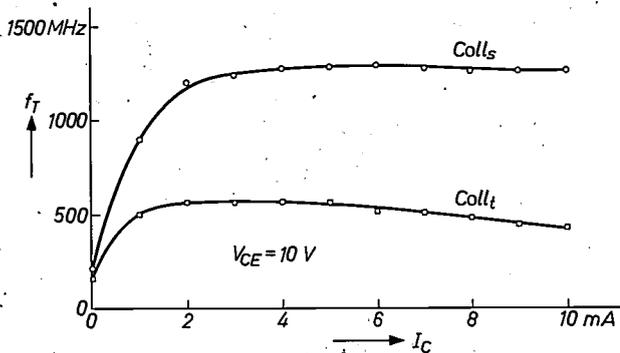


Fig. 6. The current gain bandwidth product f_T plotted against collector current I_C for a collector-emitter voltage V_{CE} of 10 V. The curves show the effect on gain when using the alternative collector contacts, substrate collector contact $Coll_s$ and top collector contact $Coll_t$. The loss of gain when using the latter is very pronounced and is probably due not only to the magnitude of the collector resistance but also to the location of a large part of it under the contact and remote from the collector junction.

collector contacts. The loss of gain when using the top contact is very pronounced and this is probably due not only to the magnitude of the collector resistance but also to the location of a large fraction of it under the contact, relatively remote from the collector junction. The whole of the collector base capacitance is therefore involved in providing feedback to the detriment of the f_T .

Some items of this work were partially supported by the United Kingdom Ministry of Technology.

Summary. Contour deposition is a technique in semiconductor device technology for making isolated regions of one type of semiconductor in a substrate of a different type with well controlled dimensions and properties. Two applications of this technique in the field of silicon integrated circuits are described, namely the construction of monolithic complementary pairs of MOS transistors and isolated UHF bipolar transistors.

The channel electron multiplier, a new radiation detector

J. Adams and B. W. Manley

The channel electron multiplier is a radiation and particle detector which meets the stringent requirements of versatility, sensitivity, ruggedness and simplicity, necessary in space experimentation [1]. In addition it offers new possibilities in image intensification.

The channel multiplier (fig. 1) is a distributed dynode multiplier which combines the functions of the dynode structure of the conventional photomultiplier and the resistor chain which divides the potential among the separate dynodes. It consists of a glass tube having a length between 50 and 100 diameters; the inside surface is coated with a semi-insulating layer, and the electrical resistance between the electrode connections made to each end of the tube is between 10^9 and $10^{11} \Omega$.

The device operates in vacuum — in space the environmental vacuum is sufficient, so that no input window is needed — with a potential difference of about 3000 volts applied between the electrodes. Energetic particles or radiation entering the low potential input end of the multiplier liberate electrons from the channel

wall. These electrons are accelerated axially by the applied electric field. They strike the wall of the channel after gaining considerable energy from the field, to produce secondary electrons. These in turn are accelerated by the applied field to produce further secondaries at subsequent collisions. This process is repeated many times along the length of the channel, and the resulting electrons finally emerge from the high potential end of the channel.

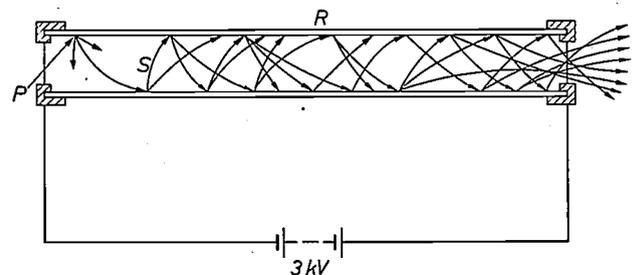


Fig. 1. A channel electron multiplier. Primary radiation P (electrons, ions, U.V., X-rays) entering the low potential end of the resistive tube R generates secondary electrons S . These are accelerated and multiplied along the tube and emerge from the high potential end.

as collector terminal. Fig. 6 shows the current gain bandwidth product f_T plotted against collector current I_C for the same transistor with the alternative

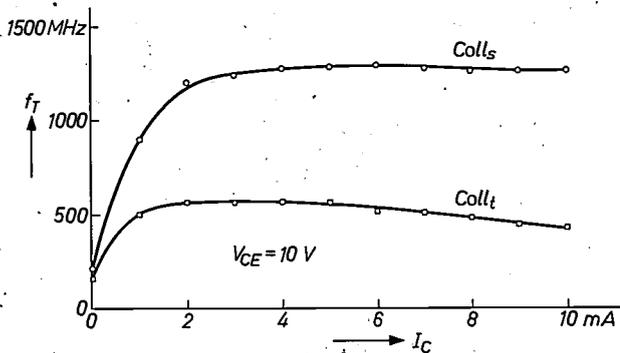


Fig. 6. The current gain bandwidth product f_T plotted against collector current I_C for a collector-emitter voltage V_{CE} of 10 V. The curves show the effect on gain when using the alternative collector contacts, substrate collector contact $Coll_s$ and top collector contact $Coll_t$. The loss of gain when using the latter is very pronounced and is probably due not only to the magnitude of the collector resistance but also to the location of a large part of it under the contact and remote from the collector junction.

collector contacts. The loss of gain when using the top contact is very pronounced and this is probably due not only to the magnitude of the collector resistance but also to the location of a large fraction of it under the contact, relatively remote from the collector junction. The whole of the collector base capacitance is therefore involved in providing feedback to the detriment of the f_T .

Some items of this work were partially supported by the United Kingdom Ministry of Technology.

Summary. Contour deposition is a technique in semiconductor device technology for making isolated regions of one type of semiconductor in a substrate of a different type with well controlled dimensions and properties. Two applications of this technique in the field of silicon integrated circuits are described, namely the construction of monolithic complementary pairs of MOS transistors and isolated UHF bipolar transistors.

The channel electron multiplier, a new radiation detector

J. Adams and B. W. Manley

The channel electron multiplier is a radiation and particle detector which meets the stringent requirements of versatility, sensitivity, ruggedness and simplicity, necessary in space experimentation [1]. In addition it offers new possibilities in image intensification.

The channel multiplier (fig. 1) is a distributed dynode multiplier which combines the functions of the dynode structure of the conventional photomultiplier and the resistor chain which divides the potential among the separate dynodes. It consists of a glass tube having a length between 50 and 100 diameters; the inside surface is coated with a semi-insulating layer, and the electrical resistance between the electrode connections made to each end of the tube is between 10^9 and $10^{11} \Omega$.

The device operates in vacuum — in space the environmental vacuum is sufficient, so that no input window is needed — with a potential difference of about 3000 volts applied between the electrodes. Energetic particles or radiation entering the low potential input end of the multiplier liberate electrons from the channel

wall. These electrons are accelerated axially by the applied electric field. They strike the wall of the channel after gaining considerable energy from the field, to produce secondary electrons. These in turn are accelerated by the applied field to produce further secondaries at subsequent collisions. This process is repeated many times along the length of the channel, and the resulting electrons finally emerge from the high potential end of the channel.

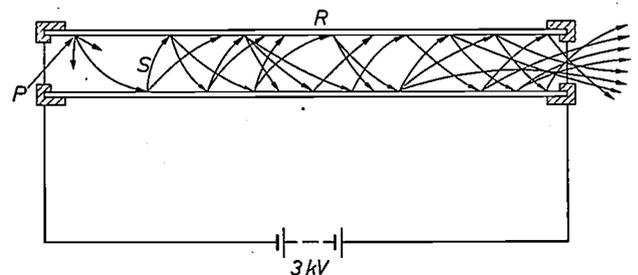


Fig. 1. A channel electron multiplier. Primary radiation P (electrons, ions, U.V., X-rays) entering the low potential end of the resistive tube R generates secondary electrons S . These are accelerated and multiplied along the tube and emerge from the high potential end.

The gain depends upon the applied voltage and the ratio of length to diameter of the channel, but not upon the absolute dimensions. The maximum gain which can be achieved is about 3×10^8 , at which level the output pulses can be readily counted using conventional techniques. Thus multipliers can be made with input apertures which range from several millimetres down to a fraction of a millimetre. In this latter form they are particularly convenient in a two-dimensional array for mapping radiation impinging on an area — image detection and intensification for example.

the multiplier input. Because positive ions need to be accelerated over most of the channel length to acquire sufficient energy to generate a significant number of secondary electrons, only a small curvature is necessary to eliminate the effect [2]. However, it is clearly an advantage to be able to wrap the multiplier up and so reduce the space which it occupies; it is usually this consideration which determines the curvature.

A range of experimental types for space exploration, made at Mullard Research Laboratories, is shown in *fig. 2*.

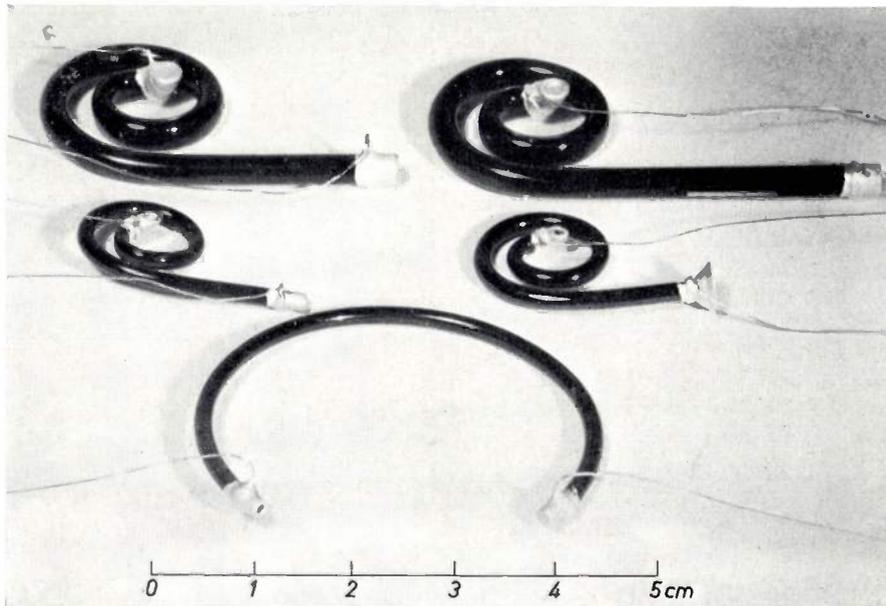


Fig. 2. A range of multipliers made at the Mullard Research Laboratories for space exploration. Included are multipliers in which the detecting area has been enlarged by forming the input aperture into a cone, or by using a slit in the channel wall close to the input electrode. The Mullard B310 multiplier mentioned in this article is the lower of the two spiral multipliers in the left of the picture. The diameter of the input aperture is 1.25 mm.

The channels may be straight tubes, and usually the smaller multipliers are of this form. A disadvantage of straight multipliers in the larger sizes is that they suffer from positive ion feedback. Positive ions are generated by collision of the electron stream within the multiplier with residual gas atoms, and the ions will be accelerated back down the multiplier to reach the input. There they can liberate electrons from the walls and so initiate new output pulses. These "after-pulses" continue for a time which depends upon the gas pressure and the channel dimensions. To avoid this pressure dependence of gain, multipliers having a diameter above 0.5 mm are curved to prevent positive ions getting back to

Operating characteristics of channel multipliers

The gain G of a curved channel multiplier as a function of voltage V_0 applied between the electrodes of the multiplier is given [3] approximately by the equation:

$$\ln G = \left(\frac{V_0 \beta^2}{36 V_{em}} \right)^{1/3} \ln \left\{ K \left(\frac{36 V_{em} V_0^2}{\beta^2} \right)^{1/3} \right\}, \quad (1)$$

where eV_{em} is the initial energy of emission of the secondary electrons, β is the ratio of the length of the channel

[1] J. Adams and B.W. Manley, *Electronic Engng.* **37**, 180, 1965.

[2] D. S. Evans, *Rev. sci. Instr.* **36**, 375, 1965.

[3] J. Adams and B. W. Manley, *IEEE Trans. on nuclear science* **NS-13**, No. 3, 88, 1966.

to its radius of curvature, K is the secondary emission coefficient per electron volt of collision energy of electrons with the channel wall. K is presumed to be approximately constant for the electron energies occurring in a channel.

The theoretical characteristic for a Mullard B310L channel is shown in *fig. 3* together with the observed results. There is good agreement over several orders of magnitude until the channel begins to saturate at a gain of about 10^8 . This gain saturation occurs at about the same level in all curved multipliers. It is due to the distortion of the electric field within the channel by the space charge of the electrons. As the space charge cloud resulting from a single input electron grows, the transverse velocities of new secondary electrons are reduced. The charge density finally reaches the point where the secondary electrons are reflected back to the wall without gaining sufficient energy to increase the number of electrons in the cloud. This state of dynamic equilibrium is maintained and the charge cloud remains at a constant level of about 3×10^8 electrons as it proceeds along the channel to the output.

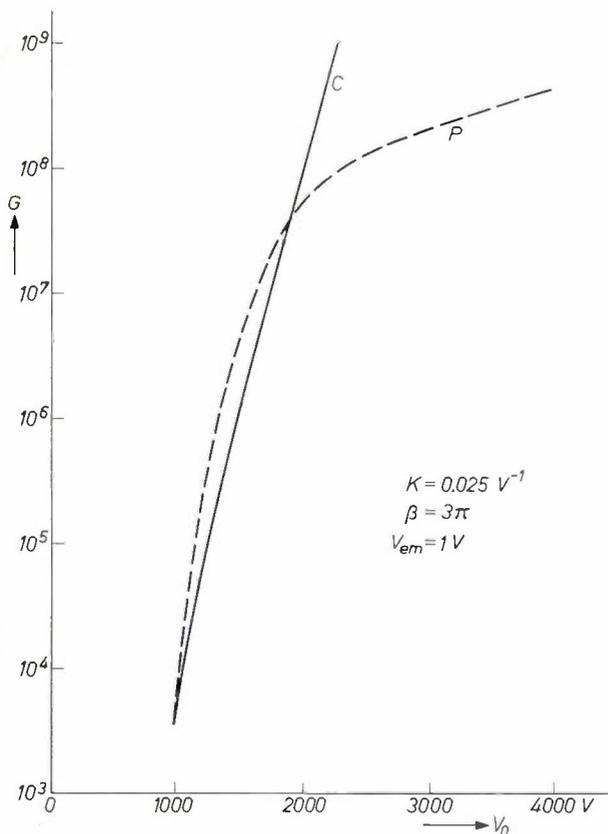


Fig. 3. Electron gain G versus applied voltage V_0 for Mullard B310L multiplier. P practical curve showing saturation due to space charge, C curve calculated from equation (1) with the parameter values indicated.

Pulse amplitude distribution

In a curved channel the maximum amplitude of an output pulse resulting from a single electron input is limited by the space charge and this shows itself in the pulse amplitude distribution. *Fig. 4* shows the distribution obtained from a Mullard B310 multiplier. The resolution of the distribution, defined as the ratio of the full width of the distribution at half peak height to

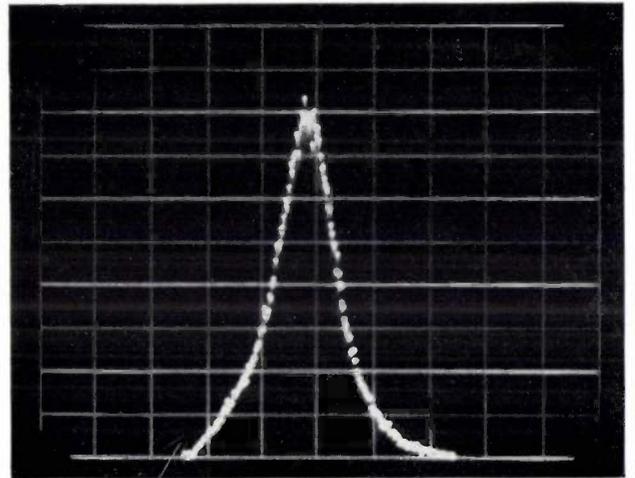


Fig. 4. A typical pulse amplitude distribution of a B310L multiplier operated in the space charge saturated mode. This photograph from the screen of a multi-channel analyser shows gain G as abscissae and frequency of occurrence of a particular gain as ordinate. The modal gain is 2.25×10^8 and the resolution of the distribution is 0.5.

the gain at the peak height, improves with gain in the range 10^7 to 10^8 but then worsens at higher gains. The best resolution is about 0.45, which is to be compared with about 1.5 obtained with conventional multi-dynode multipliers. The resolution in a conventional multiplier is limited by the statistical variation of the gain at each stage, while the narrow distribution in the channel multiplier results from the progressive constraint which space charge imposes on the gain process when the gain exceeds 10^7 .

For pulse counting, it is desirable to operate a channel multiplier in the space charge saturated mode so that the pulse amplitude distribution is narrow. This ensures that a large fraction of output pulses are above the threshold of the counting system.

Speed of response

The output current pulse resulting from a single input electron continues for a time dependent upon the geometry of the multiplier and the applied voltage. In a Mullard B310 the rise time of the pulse is about 5 ns and the total pulse width about 10 ns.

Due to the finite resistance of the channel multiplier, there is a recovery period after the passage of a pulse while the potential distribution on the wall of the channel is restored to its original form. While this is not a "dead-time" like that observed in a Geiger tube, it does mean that as the electron input rate is increased, the mean level of the output pulses diminishes. Thus if a threshold is set for counting pulses, there is a maximum rate which can be detected depending upon the resistance of the channel multiplier. As an example for a multiplier with a resistance of $10^9 \Omega$, a count rate of 10^5 pulses per second can be achieved above a threshold of 2×10^7 electrons per pulse.

Noise

A channel multiplier will produce spurious pulses unrelated to the input. This background count rate may be due to cosmic radiation, or radioactivity within the experimental system, or it may be generated within the multiplier by the action of high fields on faults in the multiplying surface. The number of spurious pulses depends upon the applied voltage, but typically in a Mullard B310 multiplier will be less than one pulse in ten seconds. Because of the low frequency of spurious pulses the channel multiplier can be used for particle counting rates extending over six orders of magnitude.

Sensitivity of channel multipliers

Channel electron multipliers have been applied to the detection of electrons having energies above 100 eV, positive ions of similar energies, X-rays in both the diagnostic range and the soft X-ray region (1-10 nm), and vacuum ultraviolet radiation.

The detection efficiency of a channel multiplier defined as the percentage of input particles or quanta producing detectable output pulses, varies considerably with the type of radiation and its energy. *Table I* collates the presently available data.

Table I. Detection efficiency of channel electron multiplier for various forms of excitation.

Type of radiation	Detection efficiency of channel multiplier
Diagnostic X radiation ^[4] (60-100 kV, 0.012-0.02 nm)	1%
Soft X radiation ^[5] (0.8-7 nm)	20%
U.V. radiation ^[6] (30-70 nm)	7%
Positive ions ^[7] (Li ⁺ at 1000 eV)	10%
Electrons ^[7] (450 eV)	60%

Multiple arrays

Because the channel electron multiplier may be scaled over a range of sizes without affecting its performance, it is possible to make parallel arrays of small multipliers to detect information in two dimensions and to produce images with point by point intensification. Arrays of this type — called channel plates — are made by fusing the separate straight channels together and polishing the parallel input and output faces of the plate. The electrodes are applied by evaporating a metal film at an oblique angle on each surface. *Fig. 5* shows a magnified part of a 10 cm diameter plate, suitable for imaging purposes, composed of channels 200 μm in diameter.

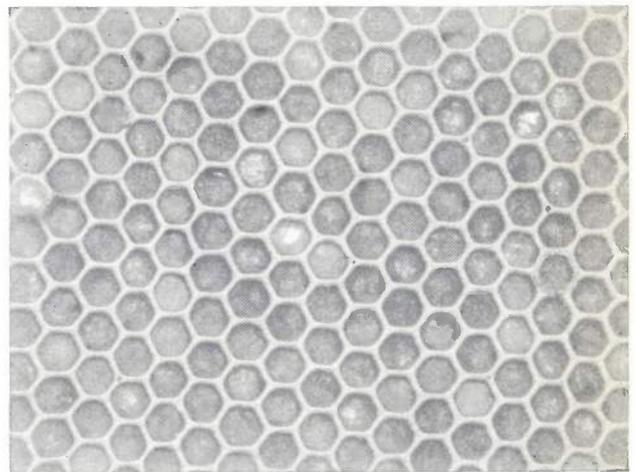


Fig. 5. View of part of a channel plate composed of channels of 200 μm diameter.

It is anticipated that plates of this type will prove especially useful for the analysis of soft X-ray radiation from the sun and from stellar sources — a subject assuming increasing importance in astronomy.

When used in the space environment, a channel plate needs no input window and the detection efficiency in the range 1-10 nm (1240 eV-124 eV) in which much stellar radiation falls is good. By placing a channel plate at the focus of a satellite borne X-ray telescope, and using the sensitivity of the input electrode of the plate to convert the X-rays into photoelectrons, an intensified image of the X-ray pattern can be obtained. The output electrons can be recorded on film or they may be used to excite a phosphor viewed by a TV camera tube, and the output transmitted to a ground station (*fig. 6*).

^[4] J. Adams, *Advances in Electronics and Electron Physics* **22A**, 139, 1966.

^[5] D. Smith, Leicester University, private communication.

^[6] R. Speer, Mullard Space Science Laboratory, University College, London, private communication.

^[7] L. A. Frank, University of Iowa Report 65-22, July 1965.

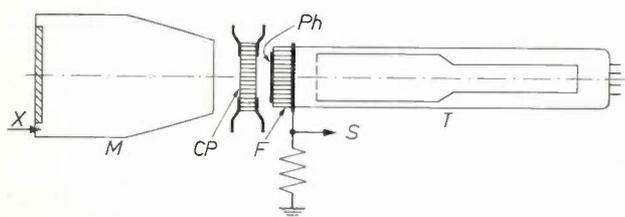


Fig. 6. A schematic diagram of a proposed soft X-ray telescope for astronomical use. The mirror *M* is composed of a paraboloidal section and a hyperboloidal section and forms an image of the incoming radiation *X* on the input surface of the channel plate *CP*. The resulting electrons emerging from the output surface of the plate are accelerated to the phosphor *Ph* deposited on the fibre optic input window *F* of the "Plumbicon" tube *T*. The electrical output signal *S* is telemetered to earth.

Channel electron multipliers maintain their sensitivity through and beyond the diagnostic X-ray spectrum from 40 keV up to at least 400 keV. At these high energies the interaction in which an X-ray quantum produces a photoelectron may take place at some depth

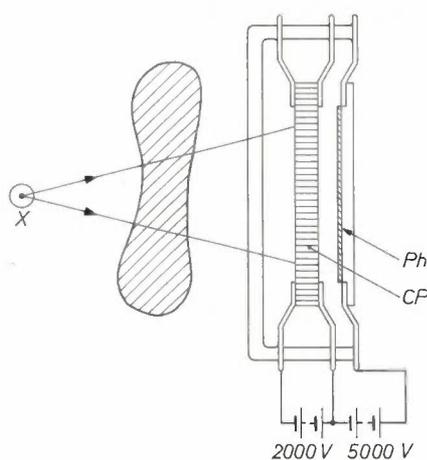


Fig. 7. A schematic diagram of an X-ray image converter based upon channel multiplication. There is no conventional photocathode in the device; the X-ray quanta are absorbed and converted into photoelectrons within the material of the channel plate. *X* X-ray source, *CP* channel plate, *Ph* phosphor.

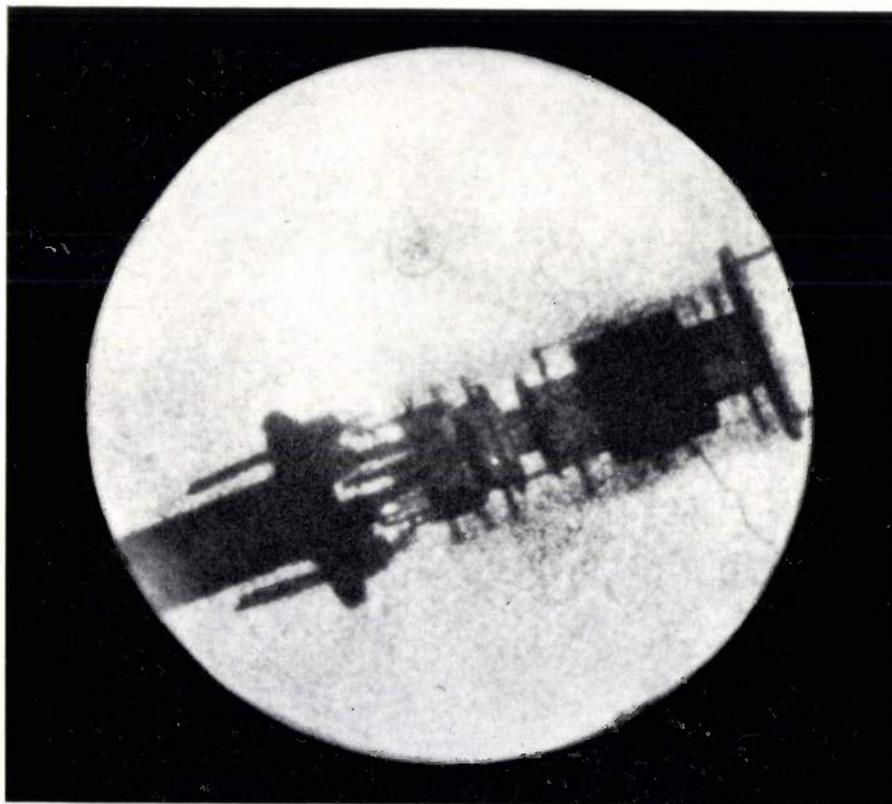


Fig. 8. A radiograph of an electron gun obtained using an image converter of the type shown in fig. 7. The channel plate was composed of channels 200 μm in diameter. The picture diameter is 8 cm.

within the channel plate. Wherever the quantum is finally absorbed however, the resulting photoelectron has only a short distance to travel in the wall in order to escape into the channel. The detection quantum efficiency of this process depends upon the material of the

channel plate, but is not strongly dependent on the X-ray energy. By suitable choice of material, detection efficiencies of about 1% can be obtained [4]. This leads to the possibility of a "panel" X-ray image converter based upon channel multiplication, and offering

the advantage, compared to a conventional image converter of a full size picture.

Fig. 7 shows a diagram of an experimental X-ray image converter incorporating a channel plate. *Fig. 8* is an X-ray image of an electron gun produced by the direct absorption of the X-rays in the channel plate, and subsequent acceleration of the multiplied photoelectrons on to the phosphor screen.

The detection efficiency obtained by X-ray absorption in the channel plate is substantially less than that achieved in conventional X-ray image converters. To achieve the same efficiency it would be necessary to incorporate an X-ray photocathode before the channel plate.

Considerable possibilities exist for intensification in other regions of the spectrum by incorporating a suitable photocathode at the input of the channel plate.

Conclusions

The usefulness of channel electron multipliers for low energy particle measurements is already being exploited on a wide scale in upper atmosphere physics. Since they are small, simple and rugged, they are ideally suited to space use, where there is as yet no competitive detector in the low energy region. Because of the wide spectrum of energies over which the device is sensitive — from 100 eV to at least 400 keV — the applications are likely to extend to other fields such as X-ray and U.V. detection, and mass spectrometry, both in space and terrestrial laboratories.

Improvements are likely to be made in the count rate capability, which is determined by the resistance of the multiplier. With the present operating voltage of around 3 kV, reducing the resistance below $10^8 \Omega$ would result in unacceptably high power dissipation for a device which loses heat by radiation alone. Thus to increase the count rate capability a more efficient secondary emitter is needed so that a lower multiplying voltage achieves the required gain.

The application of channel multipliers to image intensification and conversion offers many interesting possibilities in various parts of the spectrum. The technology of making channel plates with high resolution is a difficult one, but the resulting intensifier would require no electron optical lenses, could be physically small and would produce a high photon gain at relatively low voltages.

Summary. The channel electron multiplier is a tubular distributed dynode multiplier which responds to electrons and other charged particles and energetic quanta. With voltages between 1000 V and 4000 V applied to such multipliers they are capable of electron gains between 10^3 and 10^8 . The size of a multiplier is not important providing the ratio of channel length to diameter is between 30 : 1 and 100 : 1. It is possible, therefore, to make very small windowless radiation detectors which are useful in satellite experiments for measuring fluxes of energetic particles and cosmic X-ray and U.V. intensities. Single particles are readily detectable and the background count rate is very low. The detectors may be made so small that large numbers of them are packed into arrays for detecting radiation patterns. Such so-called channel plates may be used as intensifying elements in X-ray telescopes, and they also have applications in diagnostic radiology.

Image converter and intensifier research

P. Schagen and A. W. Woodhead

Historical aspects

The image converter tube was born in the Philips Laboratories at Eindhoven^[1] and many important contributions to its subsequent development were made within the organization. It is therefore perhaps not surprising that this type of tube has been a subject of particular interest in the Mullard Research Laboratories, almost throughout its history.

P. Schagen, Ph.D., F.Inst.P., and A. W. Woodhead, B.Sc., are with Mullard Research Laboratories, Redhill, Surrey, England.

High speed photography

In the years 1945-1950, interest was built up in image converter tubes as electronic camera shutters for high-speed photography. This was due to their freedom from mechanical inertia during pulsed operation.

The technique can be explained by reference to *fig. 1*,

[1] G. Holst, J. H. de Boer, M. C. Teves and C. F. Veenemans, An apparatus for the transformation of light of long wavelength into light of short wavelength, *Physica* 1, 297-305, 1934.

the advantage, compared to a conventional image converter of a full size picture.

Fig. 7 shows a diagram of an experimental X-ray image converter incorporating a channel plate. *Fig. 8* is an X-ray image of an electron gun produced by the direct absorption of the X-rays in the channel plate, and subsequent acceleration of the multiplied photoelectrons on to the phosphor screen.

The detection efficiency obtained by X-ray absorption in the channel plate is substantially less than that achieved in conventional X-ray image converters. To achieve the same efficiency it would be necessary to incorporate an X-ray photocathode before the channel plate.

Considerable possibilities exist for intensification in other regions of the spectrum by incorporating a suitable photocathode at the input of the channel plate.

Conclusions

The usefulness of channel electron multipliers for low energy particle measurements is already being exploited on a wide scale in upper atmosphere physics. Since they are small, simple and rugged, they are ideally suited to space use, where there is as yet no competitive detector in the low energy region. Because of the wide spectrum of energies over which the device is sensitive — from 100 eV to at least 400 keV — the applications are likely to extend to other fields such as X-ray and U.V. detection, and mass spectrometry, both in space and terrestrial laboratories.

Improvements are likely to be made in the count rate capability, which is determined by the resistance of the multiplier. With the present operating voltage of around 3 kV, reducing the resistance below $10^8 \Omega$ would result in unacceptably high power dissipation for a device which loses heat by radiation alone. Thus to increase the count rate capability a more efficient secondary emitter is needed so that a lower multiplying voltage achieves the required gain.

The application of channel multipliers to image intensification and conversion offers many interesting possibilities in various parts of the spectrum. The technology of making channel plates with high resolution is a difficult one, but the resulting intensifier would require no electron optical lenses, could be physically small and would produce a high photon gain at relatively low voltages.

Summary. The channel electron multiplier is a tubular distributed dynode multiplier which responds to electrons and other charged particles and energetic quanta. With voltages between 1000 V and 4000 V applied to such multipliers they are capable of electron gains between 10^3 and 10^8 . The size of a multiplier is not important providing the ratio of channel length to diameter is between 30 : 1 and 100 : 1. It is possible, therefore, to make very small windowless radiation detectors which are useful in satellite experiments for measuring fluxes of energetic particles and cosmic X-ray and U.V. intensities. Single particles are readily detectable and the background count rate is very low. The detectors may be made so small that large numbers of them are packed into arrays for detecting radiation patterns. Such so-called channel plates may be used as intensifying elements in X-ray telescopes, and they also have applications in diagnostic radiology.

Image converter and intensifier research

P. Schagen and A. W. Woodhead

Historical aspects

The image converter tube was born in the Philips Laboratories at Eindhoven^[1] and many important contributions to its subsequent development were made within the organization. It is therefore perhaps not surprising that this type of tube has been a subject of particular interest in the Mullard Research Laboratories, almost throughout its history.

P. Schagen, Ph.D., F.Inst.P., and A. W. Woodhead, B.Sc., are with Mullard Research Laboratories, Redhill, Surrey, England.

High speed photography

In the years 1945-1950, interest was built up in image converter tubes as electronic camera shutters for high-speed photography. This was due to their freedom from mechanical inertia during pulsed operation.

The technique can be explained by reference to *fig. 1*,

[1] G. Holst, J. H. de Boer, M. C. Teves and C. F. Veenemans, An apparatus for the transformation of light of long wavelength into light of short wavelength, *Physica* 1, 297-305, 1934.

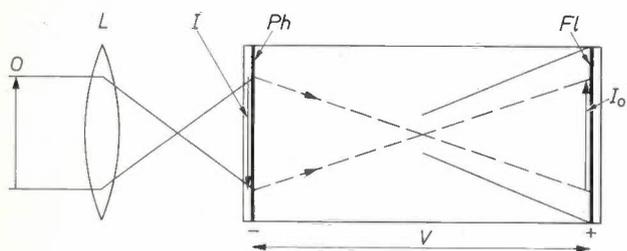


Fig. 1. Schematic diagram of a simple image converter tube. An object O is made to form an image I on the image converter photocathode Ph by the lens L . The electrons emitted from the photocathode are accelerated by the voltage V applied to the tube and form a visible image I_o on the fluorescent screen Fl .

in which a diagram is shown of a simple image converter. The scene to be photographed is imaged on to the photocathode through the glass window of the tube. Photoelectrons are released from the cathode in numbers which are proportional to the local light intensity, so that the optical image is transformed into an electron image. The electrons released from each small element on the cathode are accelerated by the electric field and focussed by an electron optical lens on to a corresponding element on the fluorescent screen. There their kinetic energy is converted into light and an optical image is once again formed.

Switching the voltage across the tube on or off, or alternatively creating or removing a negative potential barrier somewhere inside the tube with the aid of an intermediate electrode, can effectively control the flow of electrons to the fluorescent screen. This process is analogous to the operation of a conventional camera shutter as used in photography, but can be carried out far more rapidly.

The tubes which were used in some early attempts to employ this technique had been mainly designed for military applications and were not very suitable for shutter purposes. The initial results were, however, sufficiently encouraging to create a demand for an image converter tube with better characteristics in this respect. At Mullard Research Laboratories Jenkins and Chippendale designed and made such a tube, based on the electron optics of the then popular television camera tube, the image iconoscope^[2]. The resulting image converter tube, the ME 1201, is shown in *fig. 2*. It was magnetically focussed and incorporated a separate shutter electrode. Compared with the tubes which had previously been available, this tube was not only simpler to operate, but also had a much better picture quality. Comparatively simple circuitry enabled exposure times down to 10^{-7} s to be obtained. By using rather specialized techniques the photographs of a spark discharge, shown in *fig. 3*, were made with exposure times of 10^{-9} s.

Multiple-image photography

The ME 1201 soon built up an excellent reputation in single frame photography, and many hundreds were used in laboratories throughout the world. The decay time of the screen as well as the high voltage pulses (3 kV) needed to operate the tube, presented difficulties however for its application in multiple frame photo-

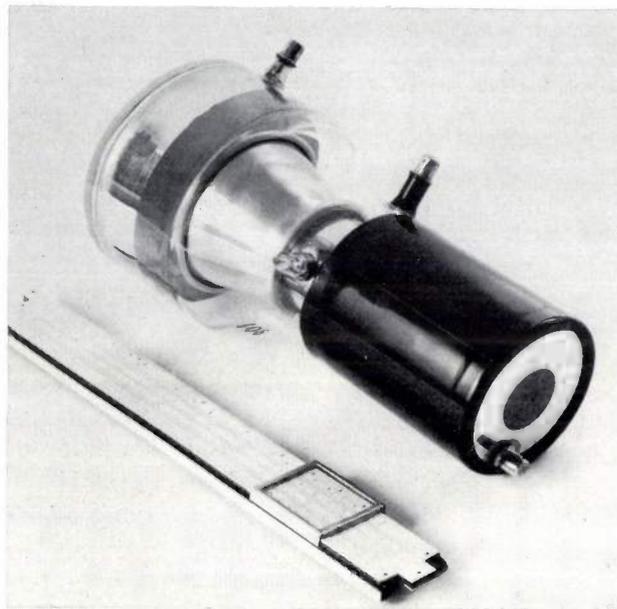


Fig. 2. The Mullard ME 1201 image converter, developed in 1951 and extensively used as an extremely fast photographic shutter.

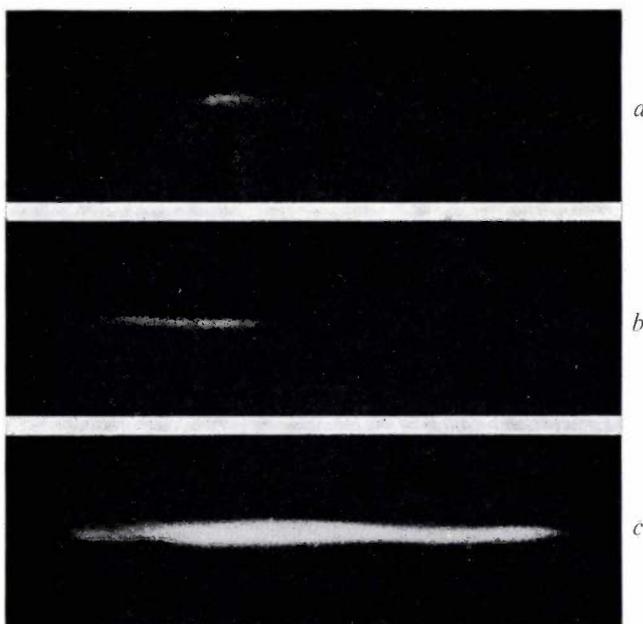


Fig. 3. Photographs of the growth of a spark, taken with an ME 1201; exposure time 10^{-9} s.

a) Initiation of the spark discharge, taken at $t = 0$.

b) Picture of the spark at $t = 1.3 \times 10^{-9}$ s.

c) Picture taken at $t = 3.4 \times 10^{-9}$ s.

graphy. These difficulties could be avoided by substituting a deflection of the image for a mere repetition of the exposure — thus at the same time eliminating the problems of very rapid film transport. In certain cases, for example spectrometry, where the object is essentially one-dimensional, the image can be swept across the screen of the tube with the aid of deflection fields. Temporal variations in the intensity of light output in the image, for instance during the development of a spark, are recorded on the screen as the image moves across it.

Alternative methods of multiple frame photography with tubes of the ME 1201 type were tried, with varying degrees of success. One method was to use an image which occupied only a small fraction of the total screen

of the undeflected image, and moving this mask along the direction of deflection.

Photographic gain

The operating voltage of the original ME 1201 was 6 kV with an electron optical magnification of 4 times. Under these conditions the luminance of the screen was of the same order as the illumination incident on the photocathode. The image on the fluorescent screen was usually recorded on the photographic emulsion in demagnified form with the aid of a high quality photographic lens with a collecting efficiency of about 1%. The photographic speed of the system was therefore low compared with direct photographic recording. Many self-luminous events did not have a sufficiently

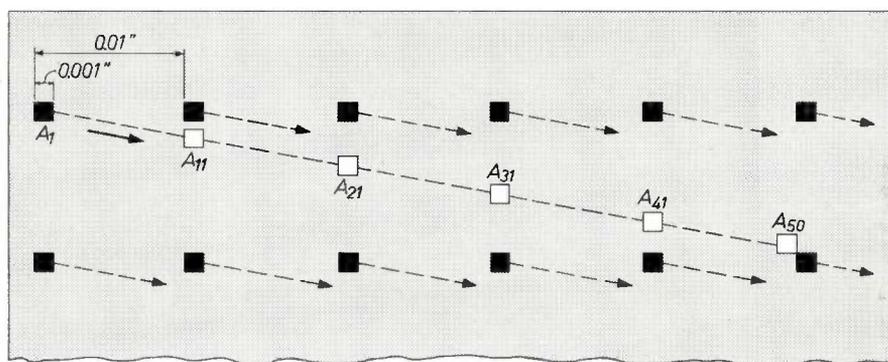


Fig. 4. Multiple picture photography with an "image dissector" tube. The photocathode is in the form of a large number of small isolated squares of 0.001" side and separated by 0.01" so that they occupy 1% of the total image area. The electron image is deflected in the direction of the arrows. The positions indicated by A_{11} , A_{21} , A_{31} , A_{41} and A_{50} correspond to frame numbers 11, 21, 31, 41 and 50 in the composite picture.

area. By deflecting the complete picture to different parts of the screen, a number of frames could be recorded on the same negative. Due to the abrupt changes in field required, this method of operation was very difficult to achieve with magnetically focussed and deflected tubes [3].

A slightly different approach was therefore made, which avoided this difficulty by employing a "dissected image" [4]. The photocathode of the image tube was constructed in the form shown in fig. 4. Light from the image only reaches the photocathode in a large number of small isolated squares which occupy approximately 1% of the total image area. The electron optical image of these squares on the fluorescent screen is deflected during the exposure to obtain a whole series of subsequent exposures which can all be recorded on the same negative. Individual frames can be examined by viewing the composite picture through a mask with apertures corresponding to the original picture points

high light output to produce a photographic record with exposure times in the order of 10^{-6} s.

In order to increase the photographic gain of the system the operating potential of the tube was increased to 15 kV. At the same time the electron optics were modified to enable a continuous variation in tube magnification between 1 and 4 times. In this way the image brightness could be further increased by up to 16 times at the expense of picture resolution.

[2] J. A. Jenkins and R. A. Chippendale, The application of image converters to high speed photography, *J. Brit. I.R.E.* **11**, 505-517, 1951; J. A. Jenkins and R. A. Chippendale, *Electronic Engng.* **24**, 302-307, 1952; J. A. Jenkins and R. A. Chippendale, *Philips tech. Rev.* **14**, 213-225, 1952/53.

[3] Recently, since electrostatically operated tubes with good picture quality have become available, excellent results have been reported by Huston with this technique: A. E. Huston, An image tube camera for photography of plasmas, *Appl. Optics* **11**, 1231-1234, 1964.

[4] G. H. Lunn, Multiple picture photography with the mosaic cathode image converter, *Proc. 3rd int. Congress on high speed photography*, pp. 102-107, 1956.

Subsequent tube developments

For some photographic applications the improvement in gain achieved by higher operating potential and decreased electron optical magnification inside the tube was still not sufficient. Methods of achieving further image intensification were therefore needed, which are described in more detail in the next section. In the meantime a number of tubes applying some of these principles to photographic shutters were developed for the United Kingdom Atomic Energy Authority, mainly for the purpose of particle reaction studies.

One of the earliest tubes of this kind, shown in *fig. 5*, incorporated two stages of image intensification inside one vacuum envelope. This was necessary to cut down

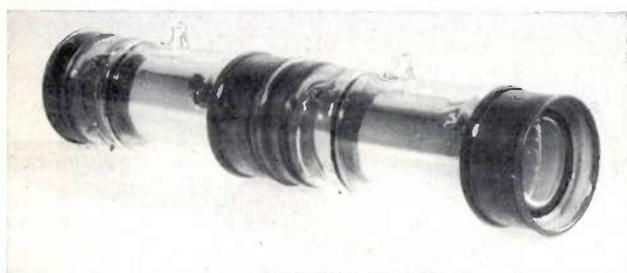


Fig. 5. An early two-stage image intensifier tube with magnetic focussing.

the light losses which occur when two completely separate image intensifier tubes are coupled via a lens which projects the image from the fluorescent screen of the first tube on to the photocathode of the second. These light losses can be reduced to a minimum if the screen of the first stage is deposited on a very thin membrane, for instance mica of less than 10 micron thick. The photocathode of the second stage can then be deposited on the other side of the membrane, and a very high coupling factor is achieved without undue loss of picture resolution.

This arrangement is particularly suitable for electromagnetically focussed image tubes, capable of employing flat photocathodes and screens. It appears to be less favourable for tubes with electrostatic focussing, when photocathode and screen should ideally be curved in opposite directions for best picture quality off-axis.

The most sophisticated tube of this kind^[5] did, however, employ electrostatic focussing. Because the tube diameter was much larger than the intermediate image size, and due to the shaping of the electrode surrounding the second photocathode, the off-axis resolution was still acceptable. A photograph of this tube is shown in *fig. 6*. The first stage had a photocathode of about 6 inches diameter, which was imaged on to the very thin fluorescent screen-photocathode sandwich. The

screen, with a decay time in the order of microseconds, provided the necessary storage time to enable the operator to select particular events for recording. The second stage incorporated a separate shutter electrode, operating at a few hundred volts. Each of the two stages demagnified electron-optically by a factor of about 3. This tube has been used as the detector and preamplifier in a system used to measure the energy of particles in the Nimrod accelerator (7 GeV) at the Rutherford High Energy Laboratory by observing the Cerenkov radiation which the particles generate^[6].

At the same time the development was initiated of an infra-red image converter tube for night viewing when employing an infra-red searchlight. The combination of night viewing applications and image intensifier studies very naturally led to an interest in passive night viewing devices, i.e. without any additional illumination. It is this area of research in particular which has been occupying most of the image tube effort at Mullard Research Laboratories during recent years.

Image intensifiers

As part of the programme on passive night vision, a thorough study was carried out of the basic principles involved in trying to improve on the perception capability of the unaided eye by electronic means^[7]. This has led to the development of a number of image tubes with different factors of intensification, which can either be used for direct viewing purposes or can be employed in conjunction with photographic recording. Alternatively they can be coupled to a television camera tube for remote viewing or electronic recording purposes.

Signal-to-noise considerations

The problem of perceiving a picture detail at low light levels can be defined as that of detecting the difference between the number of photons received from two adjacent picture elements with different brightness, during the integration time of the detector. The difficulties arise from the quantum nature of the light. This implies that the number of quanta, received from any picture element during one integration time of the instrument, will fluctuate in time as a result of the random nature of the emission process. The difference in brightness between adjacent picture elements gives rise to a "signal" at the detector, and the random fluctuations in the number of photons received cause a "noise" which is superimposed on this signal. When the signal-to-noise ratio is too small, the detector will not be able to recognize the detail in the picture.

From this consideration, it follows directly that an improvement in perception over the unaided eye by any viewing aid can only be achieved if a larger fraction of the photons emitted or reflected by picture or scene is

captured and detected by the device, or if the device has the facility to integrate the detected photons over a longer time interval.

The perception gain of an image intensifier system

The three parameters which define the basic perception ability of any instrument are therefore:

a) the *effective diameter* of the objective optics, which determines the number of photons captured from the object;

The effective diameter of the objective optics is only restricted by practical considerations of size and weight of the instrument, and will therefore vary from application to application.

It is therefore possible to achieve very considerable basic improvements in perception over the unaided eye by using an instrument with an S20 photocathode, thus gaining about a factor 10 in quantum efficiency, with an equal integration of about 0.2 s, but with a very large objective optical system, capturing many more

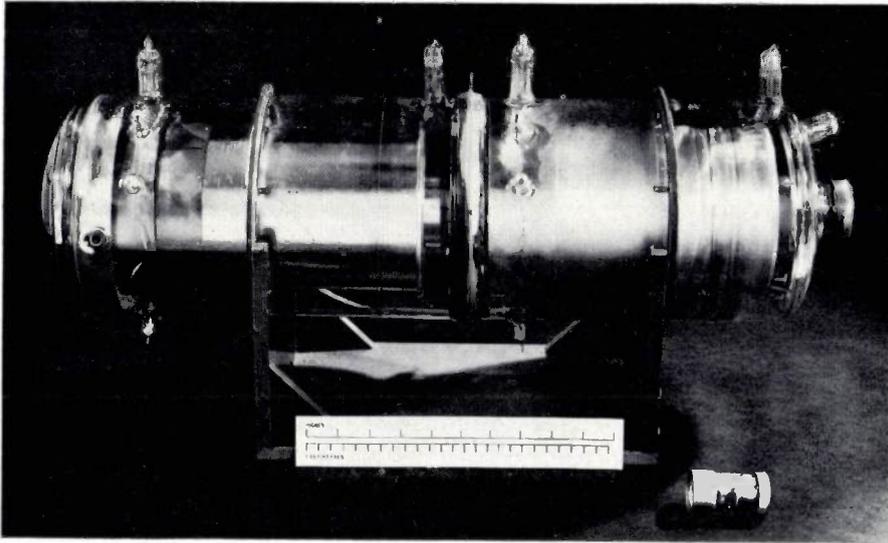


Fig. 6. Two-stage image intensifier tube with electrostatic focussing, incorporating a shutter electrode. This tube has been used as the detector and preamplifier in a system used to measure the energy of particles in the Nimrod accelerator (7 GeV) at the Rutherford High Energy Laboratory by observing the Cerenkov radiation which the particles generate^[6]. An infrared sensitive diode which might be used in an active night viewing system is shown for comparison.

b) the *quantum efficiency* of the light detecting process, which determines the fraction of the received photons which is actually detected;

c) the *integration time* of the instrument.

The eye reaches its optimum sensitivity when it is fully dark-adapted, with a maximum pupil diameter of 7-8 mm, a quantum efficiency of 1-2% for "white" light, and an integration time of nearly 0.2 s.

An image intensifier can lead to a basic improvement in perception, a *perception gain*, over the dark-adapted eye by an increase in any of the three parameters listed above, with the following practical limitations.

Any further increase in integration time would result in reducing the ability of the instrument to detect moving objects and is therefore only of limited value, even if it could be extremely useful for certain applications.

The most suitable detector of visible light for practical applications appears to be the photoemissive cathode of the multi-alkali type (S20), with a quantum efficiency of up to 10% for white light.

photons from the scene than the pupil of the eye. Fig. 7 shows such an instrument, built in the laboratory, with an overall perception gain in the order of 10^4 . With this instrument one can observe the same picture detail as with the unaided eye, but at light levels which are about 10^4 times lower.

The basic requirements for light amplification

Once the optimum perception gain of an image intensifier system has been determined by the three parameters mentioned above, the main requirement for such an instrument can be expressed as follows. It is

[5] A. W. Woodhead, D. G. Taylor and P. Schagen, A two-stage electrostatic image intensifier with a large photocathode area, *Adv. in Electronics and Electron Phys.* **16**, 105-112, 1962.

[6] P. Iredale, G. W. Hinder and D. J. Ryden, *IEEE Trans. on nuclear science NS-11*, No. 3, 139-146, 1964.

[7] P. Schagen, D. G. Taylor and A. W. Woodhead, An image intensifier system for direct observation at very low light levels, *Adv. in Electronics and Electron Phys.* **16**, 75-84, 1962; P. Schagen, *Electronic aids to night vision*, *Television Soc. J.* **10**, 218-228, 1963.

essential that the registered information in the form of recorded photons giving rise to photoelectrons should not be lost in the further processing. This means that throughout the further system the signal-to-noise ratio in the detected photons must remain the limiting factor. Since the picture information is usually again presented to the observer in pictorial form, this requirement

considerably reduced with the aid of a magnifying eye-piece.

The single-stage image intensifier tube with variable magnification

The fundamental requirement, formulated above, can be met with a single-stage image intensifier tube,



Fig. 7. Complete image intensifier system for viewing at very low light levels. Using this equipment it is possible to recognize objects when the illumination is so low that it cannot even be detected by the unaided eye.

implies a process of light amplification, or *lumen gain* in the instrument. Each photoelectron must produce more than a minimum number of photons on the final viewing screen, in order to ensure that at least one of these will be registered on the retina of the observer's eye. This minimum number depends on the state of adaptation of the observer's eye, and can of course be

provided that the observer looking at the fluorescent screen remains dark-adapted. For this purpose it is necessary to maintain the brightness of the viewing screen within a certain range. The corresponding comparatively narrow range of input brightness levels can, however, be extended by applying variable electron optical magnification in the tube [8]. This enables the

observer to spread the photoelectrons, approaching the screen, out over a smaller or larger area, without affecting the number of photons produced on the screen per incident photoelectron. The observer can thus reduce the screen brightness if this would become too high for proper dark-adaptation, without reducing the number of photons produced per photoelectron. In other words, the *brightness gain* of the tube can be varied without affecting the *lumen gain*. Fig. 8 shows a diagram of such a tube, as developed at the Laboratories.

Systems with higher lumen gain

The alternative approach is to employ such a large amount of image intensification that the observer need no longer be dark-adapted in order to ensure a registration on his retina for every photon recorded at the detection stage of the instrument. This can be achieved by coupling a number of image intensifier tubes in tandem, as described before. For night viewing applications, tubes with magnetic focussing are not very suitable due to the weight and size of the additional components and also as a result of the requirement for highly stabilized power supplies necessary to maintain correct focus. This means that electrostatically focussed tubes are preferred. As mentioned before, the difficulty

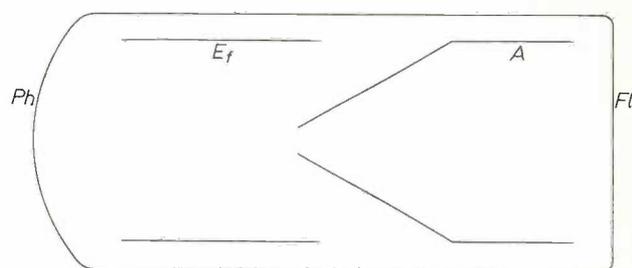


Fig. 8. A diagram of a variable magnification tube which has been developed at these Laboratories. *Ph* is the photocathode, *E_f* the focussing electrode, *A* the anode and *Fl* the viewing screen. When the anode and screen operate at the full potential of 15 kV the tube is a conventional triode with unity magnification. As the potential of the anode is reduced a highly convergent lens is formed near the screen and the image size is reduced. A small change in the voltage applied to the focus electrode is necessary to keep the image in focus on the screen. A change in magnification from 1 to 0.25 is possible.

efficiency of manufacture is bound to be reduced, thus leading to a considerably higher price. For this reason, a tube was developed with fibre optic input and output windows, flat on the outside and curved on the inside. A number of tubes like this can be stacked together in optical contact to provide as much image intensification as is needed. Fig. 9 shows diagrammatically three such tubes coupled together to give a gain of many thousands.

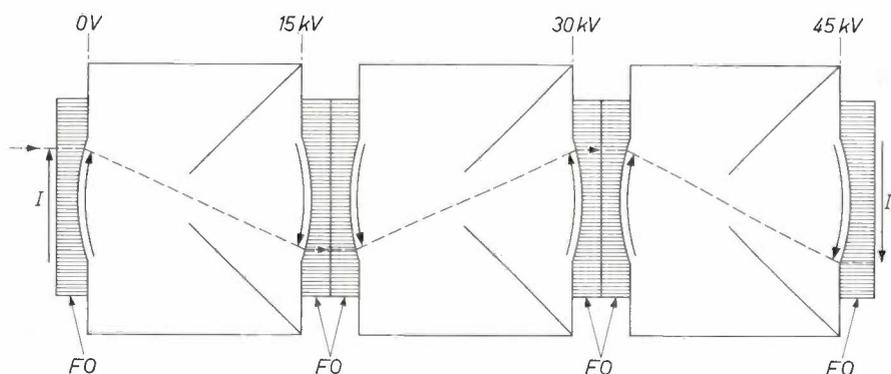


Fig. 9. The diagram shows an arrangement of three electrostatic tubes coupled fibre-optically. The fibre optic input and output windows are indicated in the diagram by *FO*. Each window is plano-concave so that the optimum conditions for good electron image formation can be achieved and the input and output optical images are flat. Thus the tubes can be simply coupled. The dashed line traces the optical and electron optical path from one point of an image *I* focussed on the input window to the image *I₀* at the output. 45 kV is applied to the whole stack, 15 kV across each stage.

when coupling such tubes is the desirability to curve the intermediate screen and photocathode in opposite directions. This difficulty can now be overcome by using a fibre optic coupling element between screen and photocathode. This can then be curved in the most suitable way to fit both sides. Another problem, however, is the manufacture of more than one photocathode in a single vacuum envelope. The resulting

The main advantage of this approach with a higher lumen gain is that it no longer requires the dark-adaptation of the observer. This makes it more comfortable to view the pictures and enables the observer to recognize picture details more rapidly. It does not, however,

[8] A. W. Woodhead, D. G. Taylor and P. Schagen, An experimental image-intensifier tube with electrostatic "zoom" optics, Philips tech. Rev. 25, 88-95, 1963/64.

provide any basic perception gain over the single-stage tube approach. It will even result in somewhat poorer resolving power for fine picture details. These can still be detectable with the single-stage tube at somewhat higher light levels.

The brightness intensification of a stacked tubes arrangement, such as shown in fig. 9, is also sufficient to allow coupling directly into a television camera tube, with the photon noise still remaining the limiting factor.

Other ways of obtaining a high brightness gain are being studied in the laboratory at present. A potentially very elegant approach is to make use of the principle of channel electron multiplication, a project which was first looked at in this laboratory in 1958, independently of the similar Russian and American efforts. This area of research is discussed in a separate paper in this issue by Adams and Manley [9].

Associated activities

The measure of success or failure of an image tube, intended to perform a specific task when incorporated in a particular instrument, depends on the combined performance of the separate system components. These are largely independent of one another. Each constit-

uent part will have its own influence on the final performance of the instrument, either in terms of sensitivity or picture quality or both.

Any research project on image tubes must therefore include system studies as soon as specific applications are considered. This implies not only a careful investigation of the most suitable tube design, photocathode preparation method, electron optical design and fluorescent screen properties, but also of the performance of the purely optical components in the system.

In the course of the years suitable facilities for these investigations have been established in the laboratory and some of these will be briefly discussed.

Optical measurements and design

An important characteristic of any imaging device is the picture resolution obtainable. In this respect the practice in photography of indicating the limiting resolution obtainable as a number of line pairs per mm just discernible, is now often considered to be insufficient in image tube practice and the much more revealing modulation transfer function (M.T.F.) is preferred [10]. This function describes the response of an instrument to sinusoidal input patterns of varying frequency. The spatial frequency response of a complete

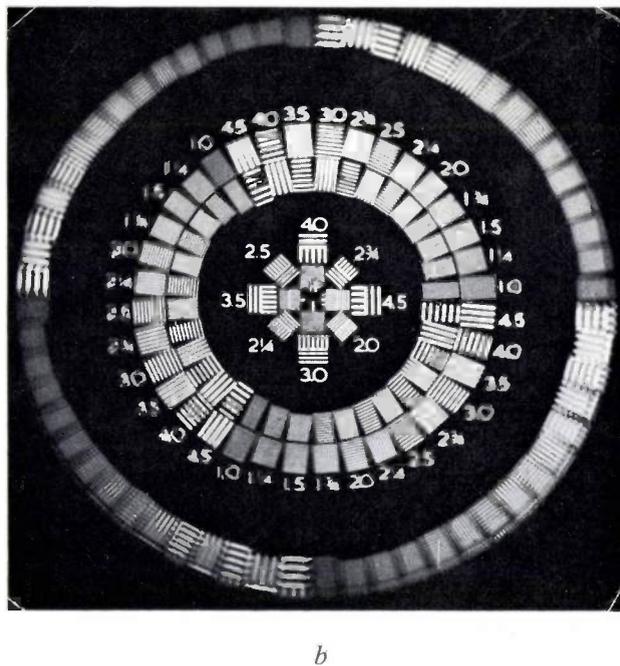
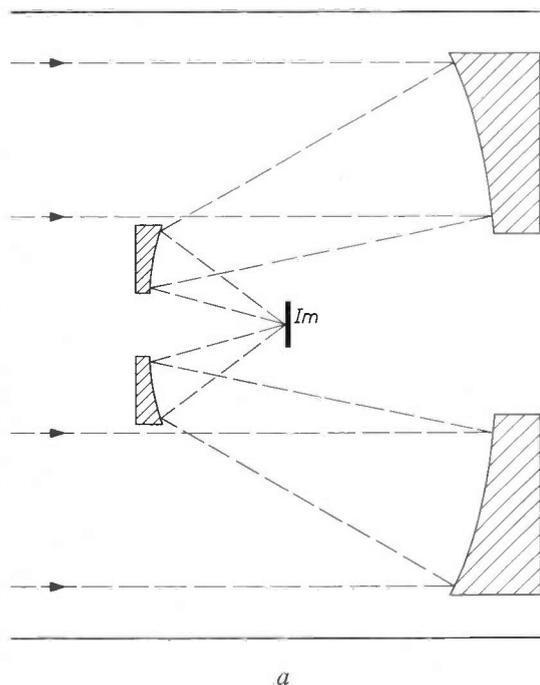


Fig. 10. Illustration of an optical system consisting of two aspheric mirrors, designed for achieving a large relative aperture and a flat image plane (*Im*).

a) Diagram of the arrangement.

b) Photograph of an image obtained by the system in visible light.

instrument can readily be found by multiplying the different functions of each of the components.

Equipment for measuring the M.T.F. of optical elements as well as that of fluorescent screens [11] and individual image tubes is in regular use in the laboratory.

For some applications, in particular when dealing with infra-red image conversion beyond 2 micron wavelength, it is very difficult to obtain suitable objective optics with a large aperture and a flat image plane. This is due to the scarcity of sufficiently homogeneous lens materials with small dispersion, which are transparent at the wavelength region concerned. For this reason a design study was undertaken, which has led to the construction of an objective system consisting of two aspheric mirrors, one of which is an ellipsoid. A diagram of this system with a photograph of an image obtained with it in visible light is shown in *fig. 10*. This mirror system can combine at an effective aperture of $F:0.76$ a flat image plane with a 6° total viewing angle. The resolution obtained in a first practical system with a reduced effective aperture $F:1$ is shown in *fig. 11*.

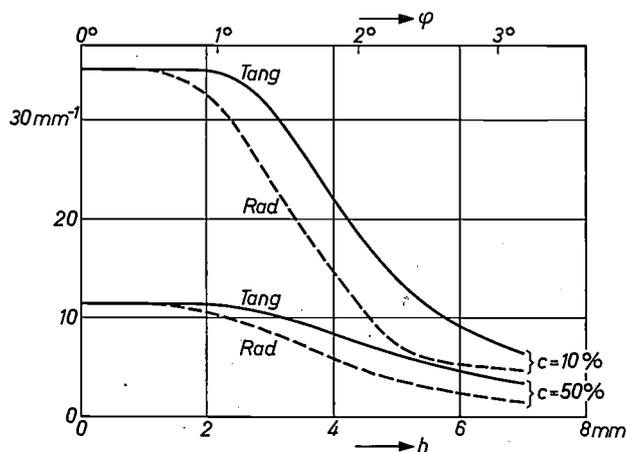


Fig. 11. Graphical illustration of the resolution obtained with a practical system consisting of two aspherical mirrors (focal length $5.0''$, effective relative aperture $F:1.0$). The value of the spatial frequencies in mm^{-1} at which the image contrast c is reduced to 50% and 10% are plotted against the radial distance h from the axis. This shows how the resolution of the system varies across the field of view. Curves for objects in both the radial and tangential direction are shown.

Photocathodes

The photocathodes employed in the various image tubes are normally either of the silver-oxygen-caesium type (S1) for near-infra-red image conversion, or of the antimony-sodium-potassium-caesium type (S20) for most other applications. Typical spectral response curves of these cathodes are shown in *fig. 12*.

Tubes used for high-speed photographic shuttering

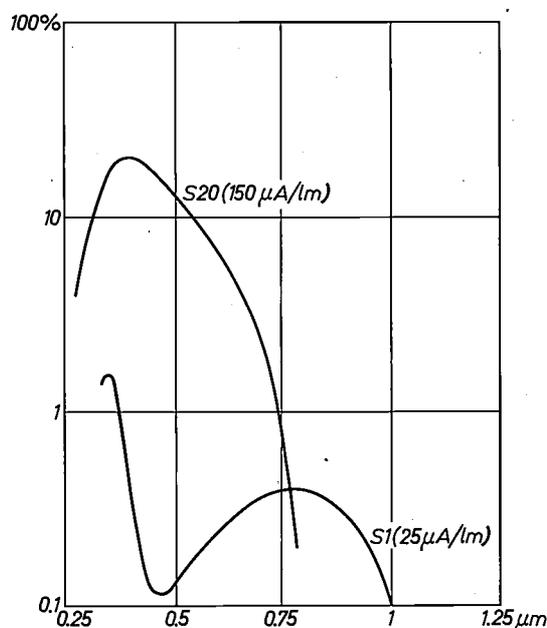


Fig. 12. Spectral response of typical S1 and S20 cathodes. The absolute photon efficiency (that is the average number of electrons emitted as a percentage of the number of incident photons of a particular wavelength) is plotted against the wavelength of the incident radiation.

usually needed a much higher conductivity than the standard cathodes could provide, in order to enable them to deliver the high current densities required during the very short pulsed operation. For this purpose a modification of the antimony-caesium (S11) photocathode was developed, with much lower internal resistance resulting from the addition of a percentage of palladium to the antimony.

The radiation from the night sky usually contains a large amount of energy in the red and near-infra-red part of the spectrum, originating in the night glow. Modifications to the processing of a standard S20 photocathode have provided an enhanced sensitivity to this radiation.

More recently the research in the photocathode area has also been extended to work on GaAs. The aim is to obtain semitransparent layers of this material with similar properties to those realized in single crystals in the Philips Laboratories at Eindhoven [12]. If successful, such cathodes or their further derivatives might well replace the S20 photocathodes in some future image tubes.

[9] J. Adams and B. W. Manley, The channel electron multiplier, a new radiation detector, Philips tech. Rev. **28**, 156-161, 1967.

[10] R. L. Lamberts, Application of sine-wave techniques to image-forming systems, J. SMPTE **71**, 635-640, 1962.

[11] D. G. Taylor, The measurement of the modulation transfer functions of fluorescent screens, Adv. in Electronics and Electron Phys. **22A**, 395-405, 1966.

[12] J. J. Scheer and J. van Laar, GaAs-Cs, a new type of photo-emitter, Solid State Comm. **3**, 189-193, 1965.

Electron optics

The early image tubes of the ME 1201 type were magnetically focussed. This was because this type of system was capable of giving a good uniform picture quality with a flat photocathode and fluorescent screen. The magnification of this tube was made variable by introducing a highly converging lens very close to the

statically focussed tubes. In this case optimum focus could be maintained by adjusting the potential of an additional cylindrical electrode situated between photocathode and anode [8].

In the preceding sections the relative merits of magnetic and electrostatic focussing have been discussed. For most applications the simplicity, low

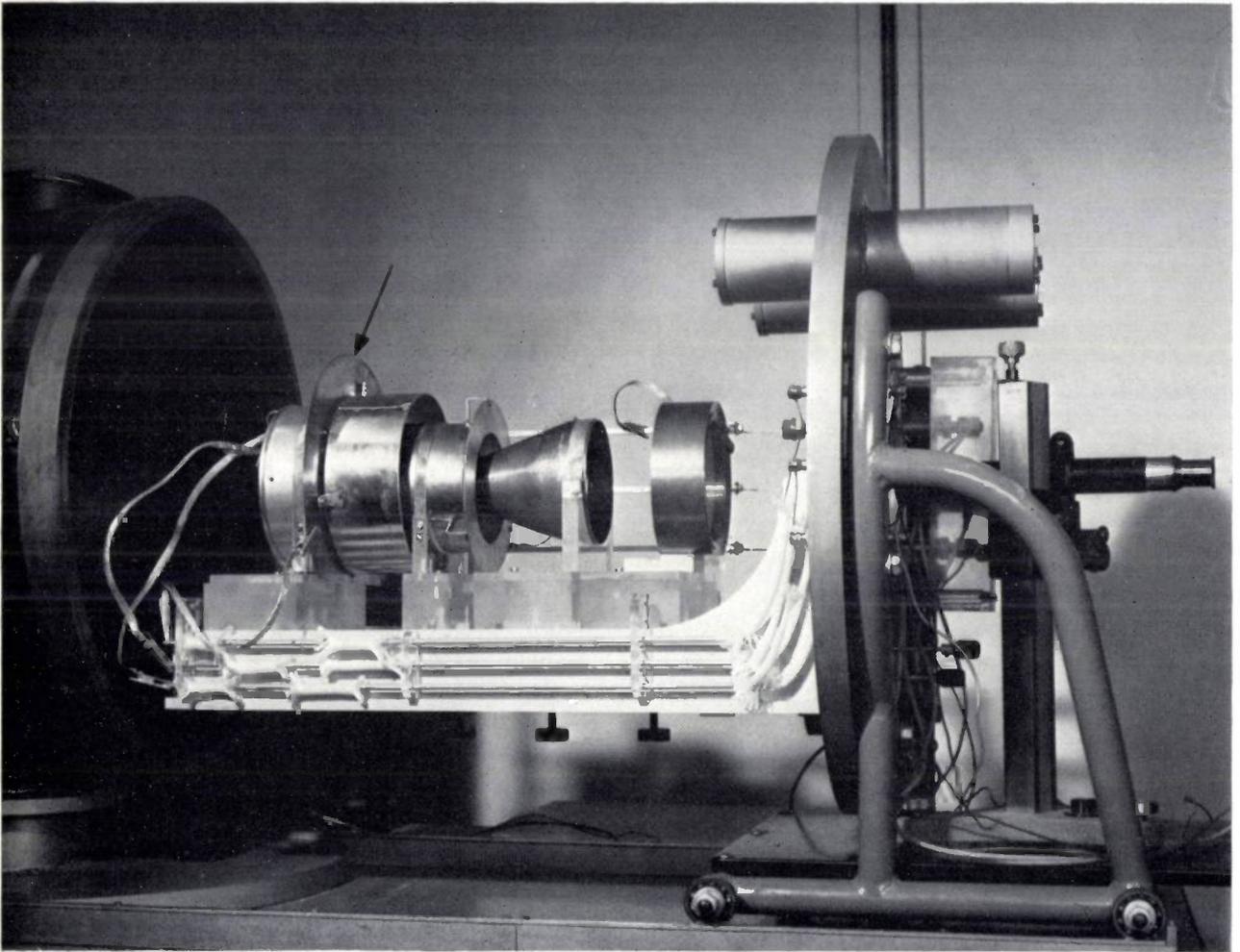


Fig. 13. Demountable system used to design electron optical arrangements for image converter tubes. The photocathode is simulated by a number of thermionic emitters placed behind the curved copper plate indicated by the arrow. Small holes are made in the plate so that the electrons can escape into the space beyond. The electrodes of a typical electrostatic image tube are shown. The fluorescent screen, which exhibits the imaged pattern of the isolated emitters, can be viewed through the window in the end plate. This structure can be moved into the bell jar which is then evacuated. The tube electrodes are adjusted by external controls to give optimum focus and picture quality.

screen of the tube. By varying the potential of the modified shutter electrode from 1 kV to the full screen potential, and adjusting the magnetic field for optimum focus, the magnification could be varied between 0.8 and 4.

The same basic principle was adopted for electro-

static systems. The light weight and small volume of the power supplies favour electrostatic systems.

Nearly all electrostatic image tubes of recent years have been based on the concentric spheres model [13]. In the ideal case the electric fields and hence the electron trajectories can be determined theoretically. In practice

deviations from the ideal model occur because photocathode and anode are only partial spheres closed by a cylindrical envelope. The boundary conditions which this geometry imposes are such that the problem becomes incapable of an exact solution.

Two approaches to the solution of these electron optical problems are possible: firstly by way of an analogue and secondly by the use of the electronic computer.

In the analogue approach a photocathode is simulated by a number of thermionic emitters placed behind a copper plate. Small holes are made in the plate so that the electrons can escape into the space beyond. The simulated cathode is situated in a large evacuated bell jar in which the other tube electrodes are also mounted.

tron trajectories show only a very small divergence beyond the point at which they focus. The resulting off-axis resolution on a flat viewing screen placed as shown in the drawing can therefore still be quite acceptable for most applications. The technique of calculating electron paths does not only aid the design of new tubes considerably, but also enables the effects of small changes in geometry to be studied. It is hoped that in this way it will be possible to establish the essential manufacturing tolerances more precisely.

Fluorescent screens

The fluorescent screen of an image tube consists of a thin granular layer of phosphor in which the kinetic energy of the impinging electrons is converted into

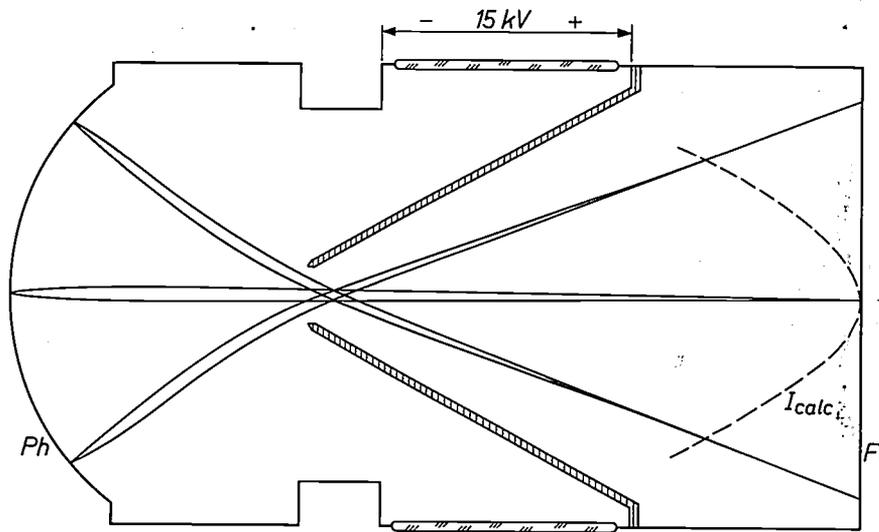


Fig. 14. Electron trajectories in a typical image tube, obtained with the aid of a computer. The calculated image plane I_{calc} shown by the broken line is curved towards the photocathode Ph . Because of the small angle of divergence of the electron beamlets the loss of resolution, at the edge of the image, even on a flat screen Fl , is small.

The fluorescent screen exhibits the imaged pattern of the isolated emitters and can be viewed through a window in the wall of the bell jar. The tube electrodes are moved by external controls and adjusted to give optimum focus and picture quality for the particular arrangement. An equipment of this kind is shown in *fig. 13*.

With the development of large, high-speed computers the calculation of the electron paths in an image tube is now possible. Appropriate methods have been described by Weber^[14] and have been developed to a high degree of accuracy at the Mullard and Philips Research Laboratories. *Fig. 14* shows some electron trajectories in a typical image tube. The image plane is seen to be a surface curved towards the photocathode. The elec-

tron trajectories show only a very small divergence beyond the point at which they focus. The resulting off-axis resolution on a flat viewing screen placed as shown in the drawing can therefore still be quite acceptable for most applications. The technique of calculating electron paths does not only aid the design of new tubes considerably, but also enables the effects of small changes in geometry to be studied. It is hoped that in this way it will be possible to establish the essential manufacturing tolerances more precisely.

In addition to the obvious requirements of suitable colour, high efficiency and appropriate decay time, the screen should permit of the highest possible resolution.

[13] P. Schagen, H. Bruining and J. C. Francken, A simple electrostatic electron-optical system with only one voltage, Philips Res. Repts. 7, 119-130, 1952.

[14] C. Weber, Calculation of potential fields and electron trajectories using an electronic computer, Philips tech. Rev. 24, 130-143, 1962/63; C. Weber, Analogue and digital methods for investigating electron optical systems, thesis Eindhoven, 1967.

The first requirement for a high resolution screen is therefore that the grain size of the phosphor must be small. Secondly the grains must stay separate throughout the screen preparation or the advantages of the small initial particle size will be lost. This latter condition is often difficult to achieve in conjunction with the requirement to obtain adequate adhesion to the substrate upon which the phosphor is deposited.

Techniques of preparing high resolution screens differ from laboratory to laboratory, but usually include a process of settling the phosphor from a suspension, which contains a material to bind the particles to the substrate. In the Laboratories a technique has been developed based upon the method used in the manufacture of colour television screens, but modified to achieve the high resolution requirements of image tubes [15]. The phosphor is suspended in a solution of polyvinyl alcohol and applied to the screen. The resulting layer is hardened by exposure to radiation from a mercury lamp through the glass supporting the screen. In this way the screen thickness can be controlled by the exposure time. Subsequently the unhardened phosphor is washed away. The polyvinyl alcohol suspension can be made very stable, enabling the small grain size to be preserved. This technique can be applied equally successfully to both curved and flat screens. Fig. 15 shows the measured modulation transfer function of a screen made by this technique where the particle size ranged between 0 and 5 micron. The limiting resolution as measured using a black and white bar pattern was 130 line pairs/mm.

Final remark

The research programme on image devices in the Mullard Research Laboratories, as discussed in this article, has gradually been extended to its present scope. There are as yet no indications that the interest in image conversion and intensification will fade in the near

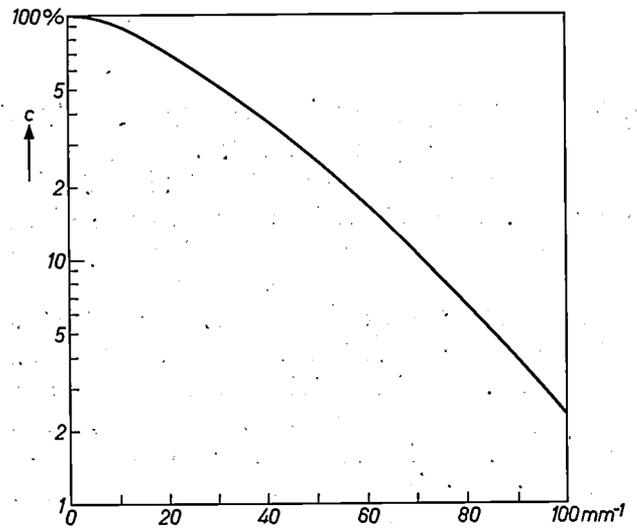


Fig. 15. Modulation transfer function of a fluorescent screen, produced with the "polyvinyl alcohol method" from particles ranging in size between 0 and 5 micron. The image contrast c is plotted against the spatial frequency expressed in mm^{-1} .

future. On the contrary, the latest developments in X-ray image conversion, passive night vision and thermal imaging techniques all seem to indicate that a further expansion of this field still lies ahead.

It is hoped that the research team in the Laboratories will also be able to make further contributions to this fascinating area of the electronics industry.

Summary. The first part of this article describes some of the image converter research carried out in the Mullard Research Laboratories in the past, when the main emphasis was on photographic shutter aspects of image tubes. This is followed by a discussion of the parameters which are important for image intensification. These considerations have had an important bearing upon the programme and have led to developments which include a 2-stage tube for detecting Cerenkov radiation and night viewing systems. In the latter field the development of tubes with variable magnification and of cascaded tubes which are fibre-optically coupled to provide a very high lumen gain are of particular interest. The final section deals with some of the associated activities in the Laboratories and their significance to image converter and intensifier tubes. This includes a brief discussion of the work on optics, photocathodes, electron optics and fluorescent screens.

[15] H. D. Stone, Preparation of high resolution phosphor screens, *Adv. in Electronics and Electron Phys.* 22A, 565-570, 1966.

Ultrasonic radar



Photograph Walter Nürnberg

The navigation of vehicles on the surface of an airfield in conditions of bad visibility is much more difficult than along a road. When traffic on the road can still travel at 40 m.p.h., movement on the airfield may well have ceased altogether. This raises difficulties when it is desired to land aircraft "blind", for if an accident occurs under these conditions the emergency services (fire, ambulance, etc.) will not be able to reach the aircraft. If they do, they may well mow down any survivors!

A short range radar device is required for these vehicles to overcome the problem. A range resolution of only a few metres is necessary and this cannot be obtained with conventional techniques. An ultrasonic radar system would be able to achieve this

resolution easily, and an experimental model (shown in the photograph) was built to assess its suitability for this application. The target area was scanned by a rotating reflector illuminated by a 40 kHz transducer, and a radar display of the area was formed. A disadvantage of this system is that air attenuates the sound and the maximum practicable range is less than 100 metres. Also sound travels relatively slowly and the target area cannot be scanned quickly enough compared with the speed of the moving objects it is necessary to detect.

Work is now progressing at Mullard Research Laboratories on an advanced radar technique to enable a high definition picture to be obtained.

Electron beam processes

H. N. G. King

Thermal machining

Work started on the use of the electron beam as a tool in the Mullard Research Laboratories in 1961. At this time our interest was concentrated on the "machining" possibilities where the energy of the electron beam is used to raise the temperature of a target to the point where local vaporization occurs. By applying suitable deflection to the beam in two planes at right angles and also, if necessary, mechanically moving the workpiece, holes of various shapes can readily be cut in all materials. The beam is normally pulsed on and off and the "on" time is chosen so that the size of the heated spot is not seriously increased by heat conduction. The pulse repetition frequency is usually limited by the maximum allowable temperature rise of the target as a whole.

A machine was designed and constructed to investigate this technique with the following specification:

voltage	10-140 kV
pulse current	0- 20 mA
mean power	up to 150 W
pulse length	1 μ s to d.c.

A tungsten hairpin emitter was used in a simple gun, the current from which was controlled by the potential of the Wehnelt electrode. A single magnetic lens focussed the beam giving under typical conditions a spot size of 5 μ m at a beam current of 1 mA and voltage of 100 kV.

The machine was made so that it could be used as a simple low resolution scanning electron microscope at a suitably reduced beam current. This facility has proved very useful for positioning the workpiece and inspecting the holes produced.

The equipment has proved to be quite versatile and a considerable amount of experimental work has been done with it both to improve the performance of the machine itself and to assess the usefulness of the technique in various applications. Of these the drilling of diamonds for wire drawing dies has been studied most fully [1].

Micro-electronic applications

More recently we have become interested in the micro-electronic applications of electron beam processes and here it seems likely that the ability of energetic electrons to cause chemical changes is more

useful than the heating effect of the beam. Thus one can expose photographic emulsion or photoresist with an electron beam instead of a light beam. The possibility of very high resolution then occurs because the beam current required is very much smaller, resulting in a much smaller spot size than that used for thermal machining. There is also, of course, no longer a limitation due to the wavelength of light, and the speed and ease with which an electron beam may be deflected makes the process very amenable to automation.

There are many other chemical reactions which electron irradiation can activate and we have been investigating the direct deposition of materials under the action of the beam [2]. Initially the work was concentrated on the making of diffusion barriers of silica deposited on silicon. It was felt that the high resolution possible could be used here to make transistors capable of operation at higher frequencies, and that the elimination of the etching involved in the photoresist technique should also help in attaining reproducible high resolution.

The technique is as follows. Some tetraethoxysilane vapour is fed into the workchamber of the electron beam machine adjacent to the target. Molecules of the vapour will then be continually condensing and re-evaporating from all surfaces but will have a certain time of stay there. If during this time they receive energy from an incident electron, this may cause re-evaporation, dissociation, or cross-linking with other molecules. In either of the last two cases a non-volatile product may be formed (silica or a polymer) and a film will be deposited wherever electrons strike a surface.

Early experiments with the machine already described showed that useful deposition rates and sub-micron resolution were in fact possible. The work has continued with a study of the deposition process and of the physical and chemical nature of films produced. To this end equipment has been set up which enables us to deposit large area films. From these it has been found that although the film as deposited contains up to 20% carbon it masks satisfactorily against the diffusion of boron, phosphorus and arsenic into silicon.

As a vehicle to demonstrate the potential of this technique we are making an MOS transistor where the diffusion mask separating the source and drain is deposited by the electron beam. A separation between source and drain of about 1 μ m is aimed at and working transistors have now been made (see fig. 1a and b).

H. N. G. King, B.Sc., is with Mullard Research Laboratories, Redhill, Surrey, England.

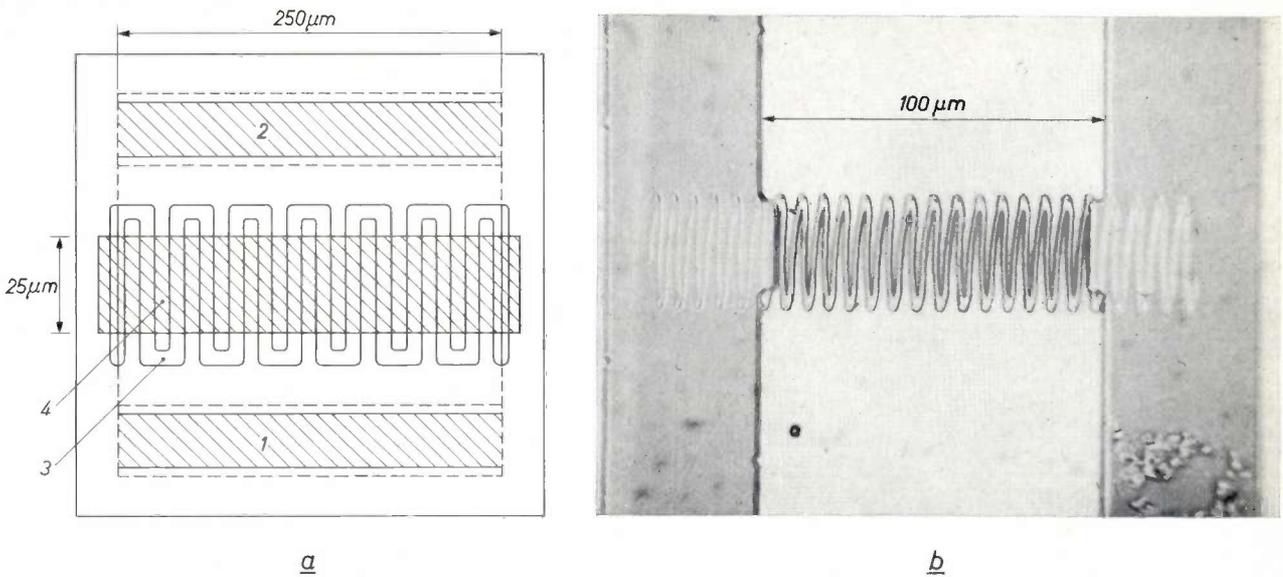


Fig. 1. a) A schematic diagram of the MOS transistor. 1 source contact, 2 drain contact, 3 channel, 4 gate electrode.
b) The electron beam deposited diffusion barrier for the transistor.

The diffusion barrier used is 200 nm thick and takes one minute to deposit. A high resolution machine has been constructed for this technique which is provided with differential pumping of the column and gun so that a good vacuum may be maintained in these regions in spite of the admission of vapour near to the target. This machine may be used with up to four magnetic lenses. Fig. 2 shows two modes of operation; in the first the pattern required is produced by suitably deflecting the focussed spot, while in the second the bottom two lenses form a demagnified image of a mask introduced half-way up the column. The top two lenses then form a condenser system to suitably illuminate the mask. This machine also has provision for scanning electron microscopy to be carried out. This facility allows precision registration of the electron beam deposited film with a reference mark on the target. A photograph of the machine is shown in fig. 3.

Although most of our work has been concerned with the deposition of diffusion barriers, it is believed that this is only one of many similar processes which may be useful. Thus we are investigating the deposition of metal films and we have achieved low resistance thin

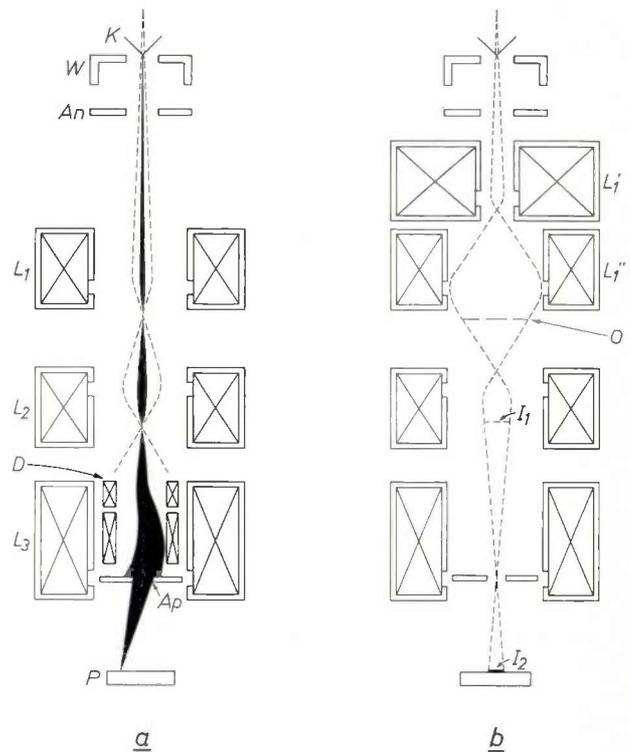


Fig. 2. A schematic diagram of the electron beam machine showing the two modes of operation, viz. a) the pattern required is produced by suitably deflecting the focussed spot and b) the bottom two lenses form a demagnified image of a mask introduced half-way up the column. *K* cathode, *W* Wehnelt electrode, *An* anode, *L₁'* condenser lens, *L₁''* second condenser lens, *L₂* intermediate lens, *L₃* objective lens, *O* object mask, *I₁* first image, *I₂* final image, *D* deflection system, *A_p* aperture, *P* workpiece.

[1] I. H. Lewin, Drilling of diamonds using electron beams, Philips tech. Rev. 28, 177-178, 1967.

[2] R. Ford, H. N. G. King, E. D. Roberts and J. M. S. Schofield, The preparation of high resolution silica diffusion barriers by an electron beam simulated reaction, to be published in Suppt. Vol. Proc. Joint IERE-JEE Conf. on Applications of thin films to electronic engineering, London, 1966.

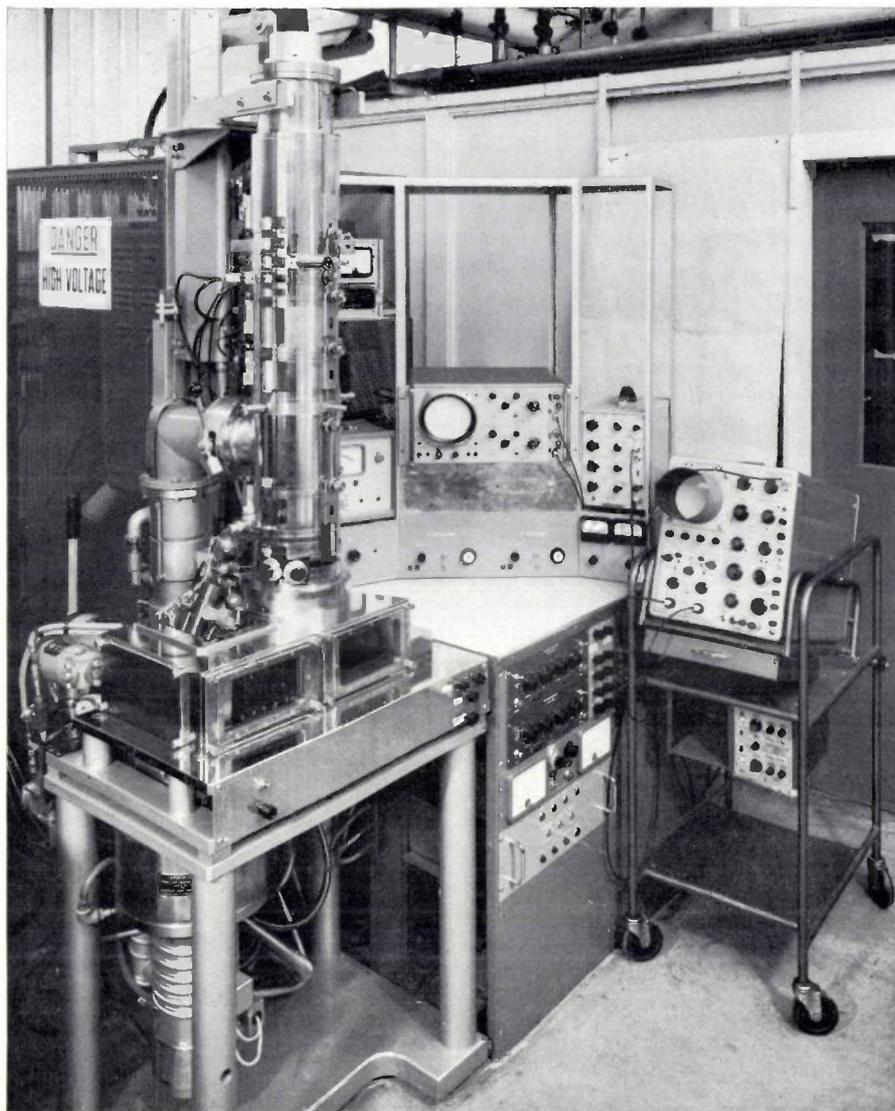


Fig. 3. The high resolution electron beam machine built at Mullard Research Laboratories for the investigation of micro-electronic applications.

films by the electron bombardment of a surface while stannous chloride is being evaporated on to it.

Automation

In order to utilize the automation potential of the electron beam, a further machine is under construction in which the position and intensity of the beam will be controlled by a small "on-line" process control computer. In order to extend the area of scan possible with a given accuracy, this machine will have the capability of positioning the beam with respect to registration marks laid down on the target. Automatic focus correction will also be provided. It is envisaged that the machine will be used to expose photoresist since this is a rapid process in keeping with the potential speed

of movement of the beam. Thus it will be possible to make masks, or work on circuits directly, the pattern produced being controlled by the tape input of the computer.

Summary. An electron beam may be used either as a heat source or to cause some chemical reaction to occur on the surface of the bombarded target in a very precisely defined region or — by controlled deflection of the beam — in a very accurate pattern. These possibilities have been investigated and the equipment to carry them out designed and built. The effect obtained can be checked by using the same set-up at a lower beam intensity as a scanning electron microscope. In particular thermal machining has been studied and the deposition of high resolution diffusion barriers for planar silicon active devices is being investigated.

Drilling of diamonds using electron beams

I. H. Lewin

Diamond dies are used for wire drawing where their hardness gives good die life and accurate control of diameter.

Within the Philips concern there is a considerable demand for dies primarily for the production of filament wire. Dies for this purpose, and also for export, are produced by the Philips die factory at Valkenswaard, Netherlands.

The three commonly used methods for diamond drilling have different advantages and limitations:

- 1) Mechanical drilling, in which a simultaneously rotating and oscillating needle is used in conjunction with a diamond paste — versatile but slow.
- 2) Electrolytic drilling, in which a low voltage is set up between the diamond and a needle whilst immersed in an electrolyte — good hole shape but limited to below 100 μm bore.
- 3) Spark erosion, in which a high voltage discharge takes place between a needle and the diamond — quite fast but produces a rough surface.

In all cases a subsequent polishing stage is required to obtain the final geometry.

Electron beams have the ability to machine hard materials thus providing another technique potentially suitable for drilling diamonds. This possible application of the electron beam machine has been investigated in conjunction with the Philips die factory at Valkenswaard and the Mullard Radio Valve Company at Blackburn.

Electron beam machine drilling technique

The section of a typical diamond die is shown in *fig. 1*. It will be seen that it is formed from three intersecting cones; this is very convenient as an electron beam drilled hole is naturally conical due to the attenuation of the mean surface energy density, through projection, down the slope of the hole. The rate of drilling is governed by the mean surface energy density and when this falls below a certain minimum value no further machining occurs.

A 70 kV, pulsed, electron beam is used for the drilling process and the pulse length, repetition frequency and beam current are adjusted for optimum drilling conditions for each individual diamond. Most drilling is done with beam currents of 3-6 mA and pulse lengths of approximately 100 μs . The pulse frequency is set to

keep the diamond well below red heat and can be within 30-500 Hz dependent upon diamond size. The 30-40 μm diameter spot is electromagnetically deflected to give a circular scan pattern, biased towards the centre where maximum machining is required.

The diamond is mounted in a slowly rotating holder to eliminate any effects of a non-circular spot, and centred in the beam axis using the machine as a scanning electron microscope. The first operation is to drill the passage cone until breakthrough occurs on the lower face of the diamond. The scan pattern is then readjusted to concentrate the energy on the walls of the hole and machining continued until the outlet diameter of the passage cone is correct. As the depth of focus over which machining takes place is only 800 μm adjustments are necessary, dependent on the relative depth of hole, to the beam focus to maintain a satisfactory drilling rate. The bell is then drilled, the scan pattern being set to the required diameter, using shorter pulses at a higher pulse rate frequency to reduce roughness due to cratering. The back relief is then drilled the diamond having first been inverted in its holder and re-centred on the passage outlet diameter.

All stages of the drilling process are monitored by direct viewing and measurements made where necessary by using the machine as a scanning electron microscope.

At a later stage the diamond is worked mechanically to obtain the final geometry.

Drilling results

Several hundred diamonds of 1-3 mm thickness have now been drilled with holes from 40 μm to 2 mm bore.

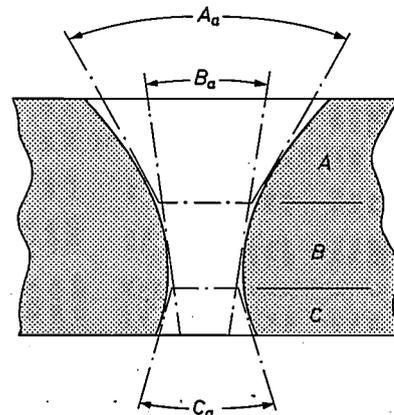


Fig. 1. Section of a typical diamond die. *A* bell. *B* passage. *C* relief. A_a bell cone. B_a passage cone. C_a relief cone.

Typical cross-sections of drilled diamonds are shown in *fig. 2* and *fig. 3*. Average drilling times for complete holes are as follows:

Nominal 100 μm bore — 8 minutes	} Time variation within each size range $\pm 50\%$
400 μm „ — 11 „	
1 mm „ — 21 „	
2 mm „ — 27 „	

These times compare favourably with the hours or days taken by conventional techniques and could be significantly reduced, particularly with the larger diamonds, by using a control system which modulated the focus in phase with the spot deflection so that the spot was in focus wherever it fell on the sloping sides of the hole.

For normal holes the minimum cone angle which can be readily achieved is 14° - 16° . The minimum size die hole which can be drilled at the present moment is about $60 \mu\text{m}$, though blind holes and breakthrough diameters (on the bottom face of the diamond) of $20 \mu\text{m}$ have been achieved. There is a limitation on maximum diameter of 2.1 mm which is due to the present spot deflection system — there are no technical difficulties associated with increasing this if required.

The surface finish of the hole is dependent on the local cone angle; within the drawing cone the roughness is 6 - $10 \mu\text{m}$ and within the bell and relief it is 15 - $20 \mu\text{m}$, generally getting worse as the diameter is increased. The bell roughness can be reduced if longer drilling times are permitted.

Two types of drilling failure, producing cracks, have been observed; major cracks which are predominantly radial and sometimes split the diamond in half, and micro-cracking which is usually a regular pattern of 5 - $10 \mu\text{m}$ cracks. Major cracks are induced by an excessive drilling rate, bad diamond clamping or inherent faults in the diamond. Micro-cracking only occurs on small holes ($< 400 \mu\text{m}$) usually at or near a change of section. Suitable drilling techniques have been developed to reduce cracking to less than 10% and this will no doubt be further reduced with more experience.

At least 50 electron beam machine drilled diamonds (600-1500 μm bore) are in use now at Mullard Radio Valve Company Blackburn, and are to date giving wear rates and tonnage per die comparable with conventionally drilled diamonds.

Conclusions

The primary advantages of the electron beam machine are its speed, flexibility of drilling and good hole shape. Against these advantages must be set its high capital cost and complexity. It is obvious that part of the drilling process can be automated, though some degree of operator control or supervision may well be

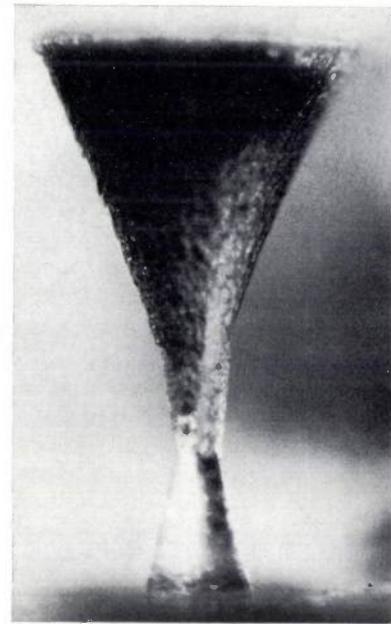


Fig. 2. Photograph of a $900 \mu\text{m}$ thick diamond; drilling time $5\frac{1}{2}$ minutes, diamond bore $68 \mu\text{m}$ for finishing to $100 \mu\text{m}$ nominal.

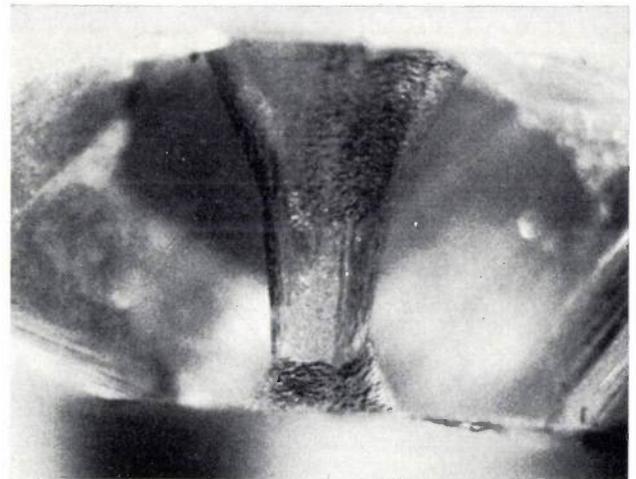


Fig. 3. Photograph of a $1200 \mu\text{m}$ thick diamond; drilling time 12 minutes, diamond bore $270 \mu\text{m}$ for finishing to $400 \mu\text{m}$ nominal.

desirable. After relatively little development the unit cost per drilled hole (allowing for the saving in diamond grit) is comparable with existing techniques so that there is every reason to hope that the use of the electron beam machine for diamond drilling purposes will be a viable economic production process.

Summary. The ability of electron beams to machine hard materials has prompted an investigation into their suitability for drilling diamond dies. An experimental 70 kV electron beam machine has been used (having a spot size of 30 - $40 \mu\text{m}$). Several hundred diamonds have been drilled ranging in thickness 1 - 3 mm and with bore $40 \mu\text{m}$ - 2 mm dia. At least 50 of these diamonds are now being used for wire drawing and are giving results comparable to conventionally drilled dies. In comparison with existing techniques the electron beam process offers a considerable reduction in drilling time and with further development should be competitive in other respects.

An experimental electronically controlled car transmission

R. W. Lindop

Introduction

Automatic transmission systems usually employ an epicyclic gearbox which enables different ratios to be obtained by actuating separate clutches or brake bands. A gear change can be effected without discontinuing the drive by actuating the new brake band or clutch before the previous one has been completely released. In addition a torque converter is often used and this transmits the drive from the engine to the gearbox, its torque speed characteristic enabling the vehicle to be started smoothly from rest. Where ample power is available, these "power shift" systems are likely to be preferred because acceleration is maintained during gear shifts resulting in a smoother ride.

This article describes a method used to convert a conventional manual-change gearbox and clutch into an electronically controlled automatic transmission.

Hydraulic actuators are made to operate the gears and clutch of a conventional car. These actuators move in response to signals from a control circuit which in turn receives and processes information of road speed, engine speed, throttle position, gear position and clutch position from a number of transducers.

Most automatic transmissions make use of the two parameters, road speed and throttle position, to select the gear ratio but the use of electronic circuits enables these parameters to be easily used in a more sophisticated way. For example, all conventional systems use some form of mechanical governor which can only indicate instantaneous values of road speed, whereas acceleration signals or simple time delays can be introduced into electronic systems giving better control.

Although epicyclic gearbox plus torque converter systems are potentially smoother in operation than the system described here, the performance of present types deteriorates as the characteristics of the brake bands change with age. Control circuits in the electronic system enable the clutch engagement to be incorporated in a feedback loop which will compensate for changes in the characteristics of the clutch and should

maintain the same performance throughout the life of the clutch.

When driving a car fitted with a conventional gearbox and clutch, a good driver can start smoothly and quickly from rest and also achieve consistently fast gear changes which are relatively free from jerks. The "electronic driver" described here has been installed in an experimental car. It takes complete control of the gear lever and clutch pedal and also control of the throttle during gear shifts. To perform as well as a good human driver the electronic control circuits must simulate or copy his techniques as closely as possible, particularly with respect to clutch control.

In the experimental car the driver is provided with a lever mounted on the steering column with which he can select the mode of operation. The lever operates a four position switch giving four conditions arranged in the sequence:

- 1) Reverse.
- 2) Neutral.
- 3) Drive.
- 4) Hold second.

"Neutral" is the position used for starting and stopping the engine and parking, etc. For normal driving, the lever is moved to "Drive" and this position gives fully automatic operation. "Reverse" is the corresponding position for backward manoeuvring. The "Hold second" position selects second gear independently of road speed or throttle position. The purpose of the "Hold second" position is to enable the engine to be used as a brake.

Clutch control and gear changing are mechanical operations and a source of controllable mechanical energy must therefore be provided. A hydraulic system was chosen because of the ease with which hydraulic power can be controlled by simple fluid valves. The hydraulic supply is obtained from a piston pump driven by a belt drive from the engine.

The gear shifting mechanism consists of two hydraulic rams, the "main jack" and the "gate jack". Oil is supplied to both jacks via a master valve which removes pressure between gear shifts and allows the gearbox internal locating mechanism to position the gears correctly.

Gear selection

An engine torque curve can be reproduced as tractive effort curves for the three gears, as shown in *fig. 1*. With a given throttle setting the excess tractive effort over the drag is available for acceleration, and the maximum speed occurs when the drag (which, of course, depends on the road surface and gradient, etc.)

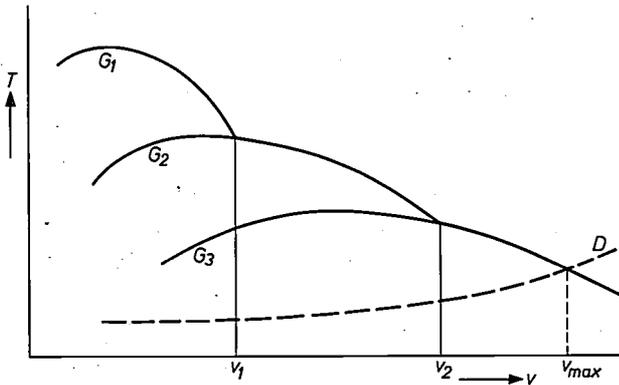


Fig. 1. Ideal points at which gear changes should take place for maximum acceleration. T tractive effort. v road speed. D drag. G_1 , G_2 and G_3 show tractive effort in first, second and third gears.

becomes equal to the tractive effort available. To accelerate to this maximum speed in the shortest possible time, the gear shifts should occur at the points where the tractive effort curves intersect (v_1 , v_2). (With some engines these points of intersection may occur above the safe maximum speed of the engine, which must then become the shift point.)

A set of curves such as that shown in *fig. 1* exists for all throttle settings, with the points of intersection occurring at progressively lower road speeds for smaller throttle openings.

To achieve rapid acceleration through the gears and also the highest road speed when climbing gradients (where the drag curve is displaced vertically from the position shown in *fig. 1*) both up and down shifts should be arranged to occur at the cross-over points for a given throttle setting. However, the vehicle speed generally falls during a gear change (with a conventional gearbox) and if up and down change conditions are made to coincide, there will be continual "hunting" between the two gears. The up changes must therefore be selected at a higher road speed than the down changes for the same throttle setting.

In the experimental car, which is provided with a gearbox for three gear positions, the two parameters, road speed and throttle position, are used to select the desired gear ratio. Road-speed signals are obtained

from a toothed wheel rotating in a magnetic field. The pulses derived from a coil in the magnetic circuit are integrated to provide a voltage proportional to speed. Throttle position is derived from a variable resistance rotated by the throttle linkage.

Road-speed and throttle-position signals are then used in the following way. The voltage proportional to road speed is modified as a function of throttle position. Thus, on light throttle the voltage increases rapidly with speed but this rate is progressively decreased with higher throttle opening. The gear selector circuit to which this voltage is applied has four critical input levels V_1 , V_2 , V_3 and V_4 (*fig. 2*). Below V_1 , bottom gear is always selected, and the change to second takes place when the input voltage rises above V_2 . To reselect first gear, the input voltage must now fall back below V_1 . This prevents any tendency to hunt between first and second gear. In the region between V_2 and V_3 , second gear is always engaged; the change to third gear takes place when the input voltage rises

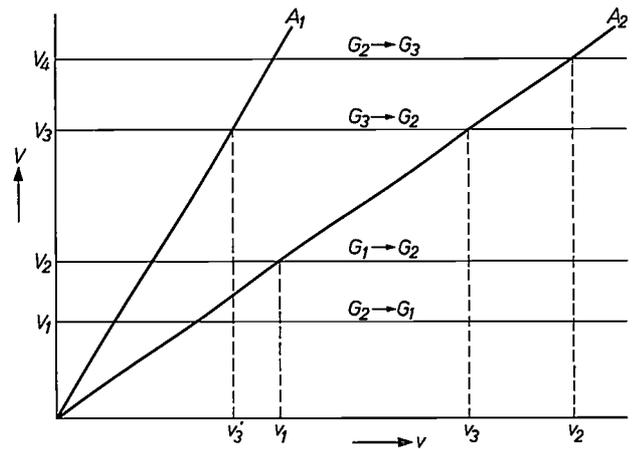


Fig. 2. The manner in which road speed and throttle positions are used to select the three gears. A_1 and A_2 show voltage curves at light and full throttle respectively. V_1 , V_2 , V_3 and V_4 are the critical voltage levels at which changes occur.

above V_4 , and similarly, to reselect second gear, the input must fall below V_3 .

A driver wishing to accelerate to a high speed in the shortest possible time will depress the accelerator to its full extent and the vehicle will follow the full throttle line A_2 in *fig. 2*. The first-to-second gear change will then occur at road speed v_1 , and the second-to-third gear change at road speed v_2 .

For lower throttle settings, v_1 and v_2 will be reduced so as to match the change in the position of the intersections of the tractive effort curves.

The speed of the vehicle is rarely constant when driving normally in the speed range where gear shifts

are necessary, and the driver is either accelerating through the gears or slowing down. When accelerating, most drivers hold a constant throttle setting and only change up when they feel the acceleration of the car fall to a low value; consequently long term variations in the performance of the engine have negligible effect on the smoothness of their driving. When gear changes are controlled by an automatic system, however, a change which occurs too early is very unpleasant due to the interruption of the relatively high acceleration. A change which occurs too late is less unpleasant, but in any case the driver is able to induce the gear shift by reducing the throttle setting. It follows that if instantaneous values of road speed and throttle position are used to select the gears, the change points should be arranged to be high (i.e. late) when the engine is correctly tuned and performing at its best. However, when the engine is cold (or old) the performance may then fall off so drastically as to make readjustment desirable. This difficulty can be overcome by setting the change speed levels suitable for a poorly performing engine (which would then normally make the changes early for a correctly tuned engine) and reducing the road-speed signal by a factor dependent upon vehicle acceleration. The "apparent" road speed is then reduced until the acceleration of the vehicle falls off; the signal then rises rapidly bringing about the shift at a more favourable point. A simple method of achieving such an effect is to introduce a long "time constant" into the circuit producing the road-speed signal so that the signal lags behind the true road speed when the vehicle is accelerating.

It is also necessary to allow the engine to be used as a brake, otherwise the performance in traffic could be embarrassing. For example, if a driver finds he is gaining ground too quickly on the vehicle in front and he reduces his throttle setting when in second gear, then the system could select third gear. In this case, as the driver takes his foot off the accelerator to reduce his speed, the up shift which results may even accelerate the car due to the excess stored energy in the engine. This effect makes it very difficult to adjust the road speed by means of the throttle at speeds below about 30 m.p.h. although this facility is very desirable when driving in tight traffic formation.

By inhibiting up shifts when the throttle setting is reduced to a level sufficiently low to give vehicle deceleration (i.e. just above the idling position), this difficulty can be almost entirely eliminated without upsetting the required law (fig. 2) and without removing the ability to induce up shifts with slight movements of the throttle.

It was seen earlier that it is desirable to delay the road-speed signal artificially. If this delay is applied

after the throttle-position information has been added, it also delays induced up shifts and allows the throttle to be eased back to the "hold" position before an up shift occurs.

Referring again to fig. 2, it is evident that third gear might be engaged at any road speed above v_3' and that, between v_3' and v_3 , a sudden opening of the throttle can cause a shift down to second gear. This "kick down" effect is useful in overtaking but, to avoid an embarrassing delay in obtaining the desired acceleration, the gear shift must be very rapid and the artificial delay of the gear selecting voltage must only operate when the signal is increasing. Unfortunately, although the driver can anticipate the need for a lower gear to give increased acceleration, the automatic control tends to prevent a lower gear from being obtained until the acceleration is actually required. If the driver kicks down and then immediately returns to a reduced throttle setting, the higher gear could be reselected (after a short delay) unless, as often happens in practice, he returns to a sufficiently low throttle level for the gear to be held.

The backlash, required between the up and the down change levels to prevent hunting, reduces the range of the kick down. It was found in practice that this backlash has to be quite large to allow for up changes on moderate inclines, and this reduces the kick down range below a satisfactory level. To improve the kick down range, the input to the selector circuit is momentarily reduced below the level appropriate to the particular throttle setting and the road speed (fig. 2) when the accelerator pedal is depressed rapidly. Thus the driver can also influence the gear selection by moving the throttle quickly or slowly.

To perform all the above operations the electronic control must be given information, and in the block diagram (fig. 3) blocks 1 to 6 represent the transducers and circuits which obtain and process this information and convert it into d.c. signals.

The output from the gear selector (block 8, fig. 3) consists of two signals in addition to the supplies to the solenoid valves which produce the gear shift. One of the signals is fed to the "gate jack comparator", the other to the "main jack comparator".

In the comparator circuits, these desired main jack and gate jack position signals are compared with signals giving the actual positions and give rise to error signals in the event of a mismatch. If the gate jack is not in the correct position therefore, a signal is fed back to the gear selection circuit which modifies the output to the main jack solenoids so as to give neutral position on the main jack. An error signal from the main jack comparator remains until the main jack has reached the new gear position. This error signal switches on the

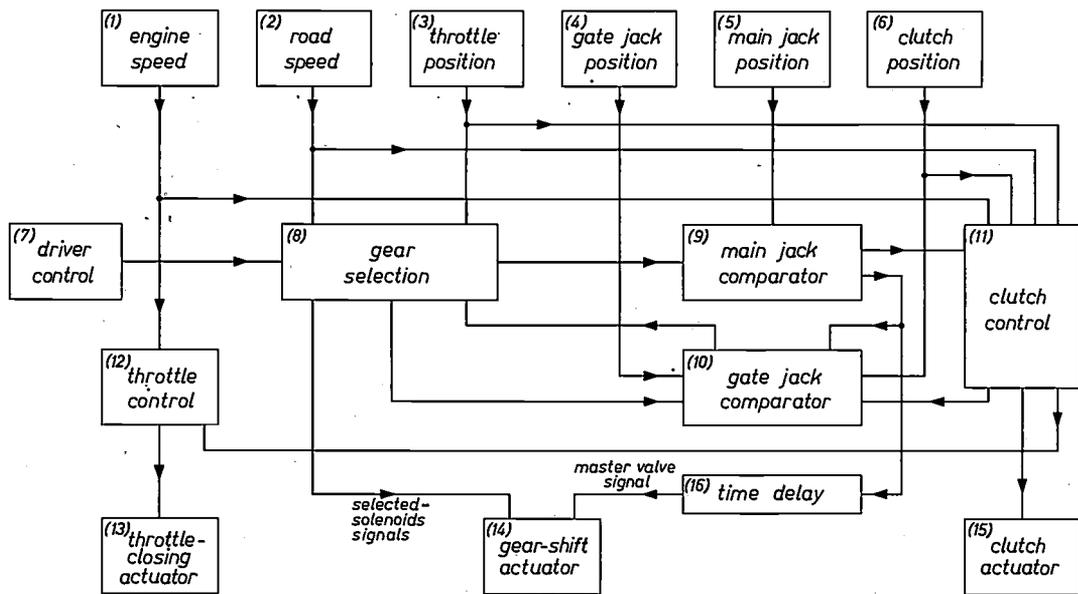


Fig. 3. Blocks 1-6 represent the transducers and circuits which provide information for the control circuits. Blocks 8, 9, 10, 11 and 12 are the control circuits operating the throttle, gear and clutch actuators (blocks 13, 14 and 15).

master valve to supply hydraulic pressure to the jacks. It also causes the clutch to disengage and the throttle to close.

Thus, if the gear selection circuit selects a new gear, the following sequence of operations occurs:

- 1) The signals to the comparator circuits change.
- 2) The comparators give error signals because the gears are now in the wrong positions.
- 3) The error signal from the main jack comparator opens the clutch, closes the throttle and energizes the master valve.
- 4) The hydraulic rams shift the gears to their new positions (via neutral and across the gate when necessary) and the outputs from the comparators fall to zero.

To complete the gear shift, the clutch must now be re-engaged and the throttle opened.

Sequential operation in this way makes possible the fastest gear shift, and position indicators enable some electrical interlocks to be added which reduce the risk of damage to the gearbox in the event of a failure. Full use is made of the existing synchromesh and no attempt is made to bring about synchronization by a double de-clutching action, although engine speed is adjusted by the throttle control as necessary.

Sequential operation in this way makes possible the fastest gear shift, and position indicators enable some electrical interlocks to be added which reduce the risk of damage to the gearbox in the event of a failure. Full use is made of the existing synchromesh and no attempt is made to bring about synchronization by a double de-clutching action, although engine speed is adjusted by the throttle control as necessary.

Clutch control

Very fine control of the clutch is necessary both during the start from rest and also during clutch re-engagement after a gear shift. Most modern clutches have a high hysteresis characteristic. This is illustrated in fig. 4, which shows a typical graph of transmitted

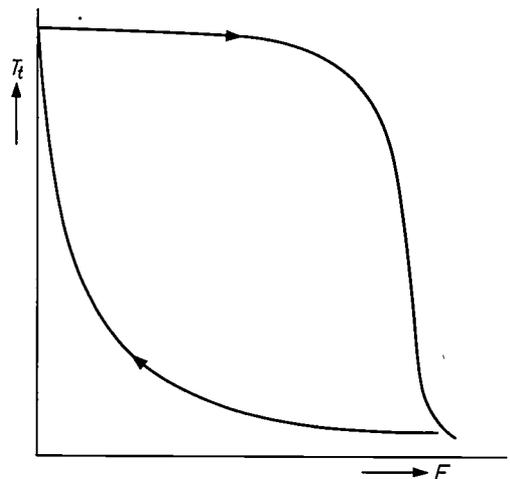


Fig. 4. Typical graph of transmitted torque T_t against clutch lever force F . The high hysteresis characteristic will be noted.

torque against clutch lever force. However, a graph of clutch lever position plotted against transmitted torque has a curve of the general shape shown in fig. 5. This characteristic has very little hysteresis, and it follows that, to achieve a fine control of the transmitted torque, the controlling element should ideally be a position-controlling rather than a force-controlling device.

The electronic circuits produce an electrical signal defining the desired torque, and this must be converted to a given position of the clutch lever by the clutch-operating mechanism. There are two parts to the clutch-operating mechanism:

- a) A moving coil servo which accurately controls the position of a small push rod, as dictated by the level of an electrical input voltage.
- b) A hydraulic power amplifier which magnifies the small force available from the moving coil servo to a level capable of operating the clutch lever.

The two sections form a mechanism which allows an input voltage to control the position of the clutch lever — and consequently the transmitted torque — to a high accuracy.

Clutch control must deal with two actions: starting from rest and re-engaging the clutch after a gear shift.

A successful start from rest can be obtained by using the basic action of a centrifugal clutch, that is, for steady state conditions, the clutch lever moves towards the engaged position as the engine speed is increased above the idling rate. It is, however, necessary to include modifications which reduce the initial rate of engagement to a relatively low level in order to avoid a jerk as the clutch plates first make contact, and control the subsequent clutch engagement rate as a function of engine deceleration.

For starting on steep hills, or for rapid acceleration on level ground, the clutch should become fully engaged at an engine speed near the maximum torque value (e.g. 2000 r.p.m.). However, under some driving conditions the clutch should remain fully engaged as the engine speed drops to as low as 1000 r.p.m. To satisfy both of these conditions, the road speed must also influence the clutch position or transmitted torque.

A low rate of clutch engagement with increase in engine speed is also helpful during manoeuvring, because it is difficult to achieve a fine control of engine

speed with the usual throttle linkage and carburettor valve arrangement.

During a gear shift the clutch must be disengaged and, at the end of the shift, re-engaged as quickly as possible. The level of the disturbance that will be caused by the clutch re-engagement will depend on the degree of mismatch of the engine speed and the rate of engagement of the clutch. The amount of jerk that can be tolerated varies with the mode of driving; if the driver is demanding a high acceleration, he will expect a greater rate of change of acceleration (i.e. jerk) at each gear shift; a driver who is using only a light throttle would be disturbed even by a relatively low transmission jerk. To accommodate these varying requirements the basic clutch re-engagement rate is made to vary with the throttle setting.

During the gear change from a low to a higher gear (for example, first to second) the engine speed is falling towards the speed corresponding to the new ratio. As the clutch re-engages, the engine deceleration increases. The deceleration is measured by an engine speed-differentiating circuit whose output reduces the rate of engagement of the clutch. This feedback enables a higher basic rate of re-engagement to be used and consequently shortens the time taken to engage the clutch at the end of a shift, without increasing the jerk experienced by the car.

To enable this differentiating circuit to be used for down changes (for example, third to second), the engine must be allowed to accelerate to a speed greater than that required for the new gear while the gear shift is taking place. This will automatically occur unless the driver has removed his foot from the accelerator. To allow for down changes when the driver is not accelerating, the above feedback is suppressed and the clutch engagement rate reduced to a low level if the engine speed is below the new gear speed.

The electronic driver type of automatic transmission described here gives a very good and comfortable performance. Naturally, it does not have the smooth gear transitions of the epicyclic gear box plus torque converter but neither does it potentially have the losses normally involved with that type of transmission.

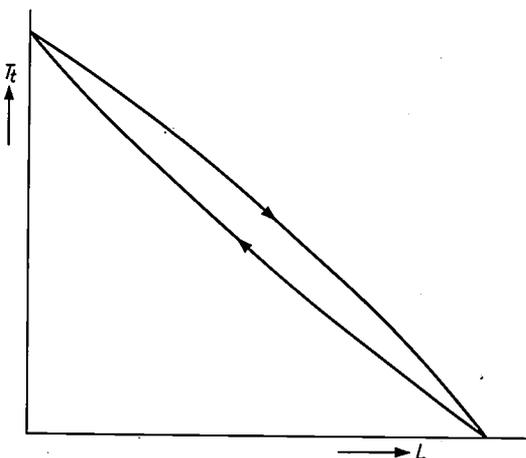


Fig. 5. Clutch lever location L plotted against transmitted torque. The lower hysteresis characteristic illustrates the need for a position controlling device.

Summary. This article describes a method used to convert a conventional manual change gearbox and clutch into an electronically controlled automatic transmission. Hydraulic actuators are made to operate the gears and clutch of a conventional car. These actuators move in response to signals from a control circuit which in turn receives and processes information of road speed, engine speed, throttle position, gate jack position, main jack position and clutch position from a number of transducers. The various control functions are described and a block diagram of the circuitry used is given.

Analysis of variability in ferrite cored inductors

C. V. Newcomb and E. C. Snelling

Introduction

Ferrite cored inductors are used extensively in the telecommunication industry as filter elements [1]. It is desirable to make the cut-off characteristics of these filters as sharp as possible for a given number of inductors; this implies that the circuit elements must have the highest possible Q factors. As ferrite material development progressively reduces the magnetic losses in ferrites, higher Q factors become attainable (Q factors of 1000 are in common use today) [2] and better filter performance becomes possible. However, high Q factor and sharp cut-off must be accompanied by a high degree of constancy of the element values during service. A filter with a sharp cut-off characteristic that changes frequency position excessively with temperature or time may have an overall performance that is inferior to a filter that is less sharp but more constant. The attention of the Ferrite Section at Mullard Research Laboratories has recently been concentrated on these problems of variability. Variability may be defined as any unwanted variation of inductance.

There are many processes which may cause the inductance to change. The changes may be reversible with temperature, i.e. they may be described in terms of a temperature coefficient, or they may be irreversible and constitute a permanent change with time. Either form may arise from the ferrite or from other sources such as the clamping parts, cemented joints in the core, the winding and the inductance adjusting mechanism. In general an observed change of inductance will be the net result of a number of processes. The ferrite contributions, both reversible and irreversible, are fairly well understood; they are usually studied by measurements on ring specimens rather than on complete inductors. The identification of the non-ferrite contributions is more difficult and requires large numbers of high resolution inductance measurements. The general procedure is to subject the inductor to a long series of temperature cycles such as that shown in *fig. 1* and to measure the inductance at each of the steady temperatures in each cycle. The temperature cycle permits the mean temperature coefficient to be deduced and it accelerates time changes. A plot of

successive inductance measurements made at the same temperature is indicative of the inductance drift that may be expected in service.

In order to isolate and identify individual processes of inductance variation it is desirable to make the measurements on samples which have been stabilized with respect to all other processes. This is not always possible and it may be necessary to make many comparative observations. It is imperative that any spurious variations introduced by the measuring technique

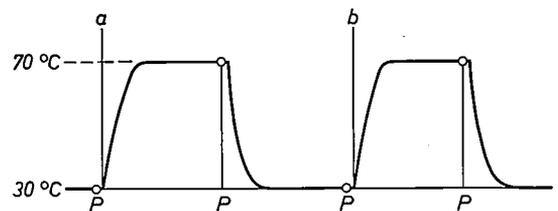


Fig. 1. The temperature cycle of ferrite cored inductors under test. P indicates the points at which the inductance is measured. The time for the complete cycle $a-b$ is 4 hours.

should be made so small that they will not mask the required results. Special apparatus has been developed for automatically performing the measurements and making possible extensive investigations which could not be contemplated by the previous manual techniques. Before describing this apparatus it is useful to note that present day inductors may be required to have a constancy better than 500 p.p.m. during a 20 year service life, therefore contributions to the overall drift as small as 100 p.p.m. are significant and if these are to be positively identified the resolution of the apparatus must be about 30 p.p.m. This requires a high standard of control; particularly important is the reset accuracy of the temperature.

The variability recording apparatus

A programmer controls the temperature cycling of a measuring chamber which contains the ferrite cored inductors under test. The temperature cycle is shown in *fig. 1*. At each temperature the inductors have at

C. V. Newcomb, A.M.I.E.E., and E. C. Snelling, B.Sc. (Eng.), F.I.E.E., are with Mullard Research Laboratories, Redhill, Surrey, England.

[1] E. C. Snelling, Ferrite-cored inductors and transformers, Mullard tech. Comm. 9, 30-42, 1966 (No. 82).

[2] E. C. Snelling, Q contour charts, Philips tech. Rev. 28, 186-187, 1967 (No. 5/6/7).

least 1 hour to acclimatize. Each inductor is then sequentially selected by the programmer and connected as the inductive element of a parallel resonant oscillator. The frequency of oscillation is automatically measured; the result is printed out and also punched on tape. When all the inductors have been so measured, the oven temperature control unit is switched to the other temperature level by the programmer. The complete cycle time is 4 hours, permitting a total of 42 cycles a week; the maximum number of inductors that may be simultaneously tested is 22. Fig. 2 shows a block diagram of the apparatus.

thermistor bridge controls the heaters through a phase sensitive detector circuit. The control accuracy is $\pm 0.1^\circ\text{C}$.

The oscillator is a two-port, parallel resonant, resistance stabilized type. The resonating capacitors are of the silvered mica type and are mounted inside a small oven maintained at $50\pm 0.5^\circ\text{C}$. One of five capacitors may be pre-selected to cover a range of inductances from $200\ \mu\text{H}$ to $55\ \text{mH}$ within the frequency range of 10 to 50 kHz.

The programmer switches on the tape punch, controls the stepping rate of the inductor selector and

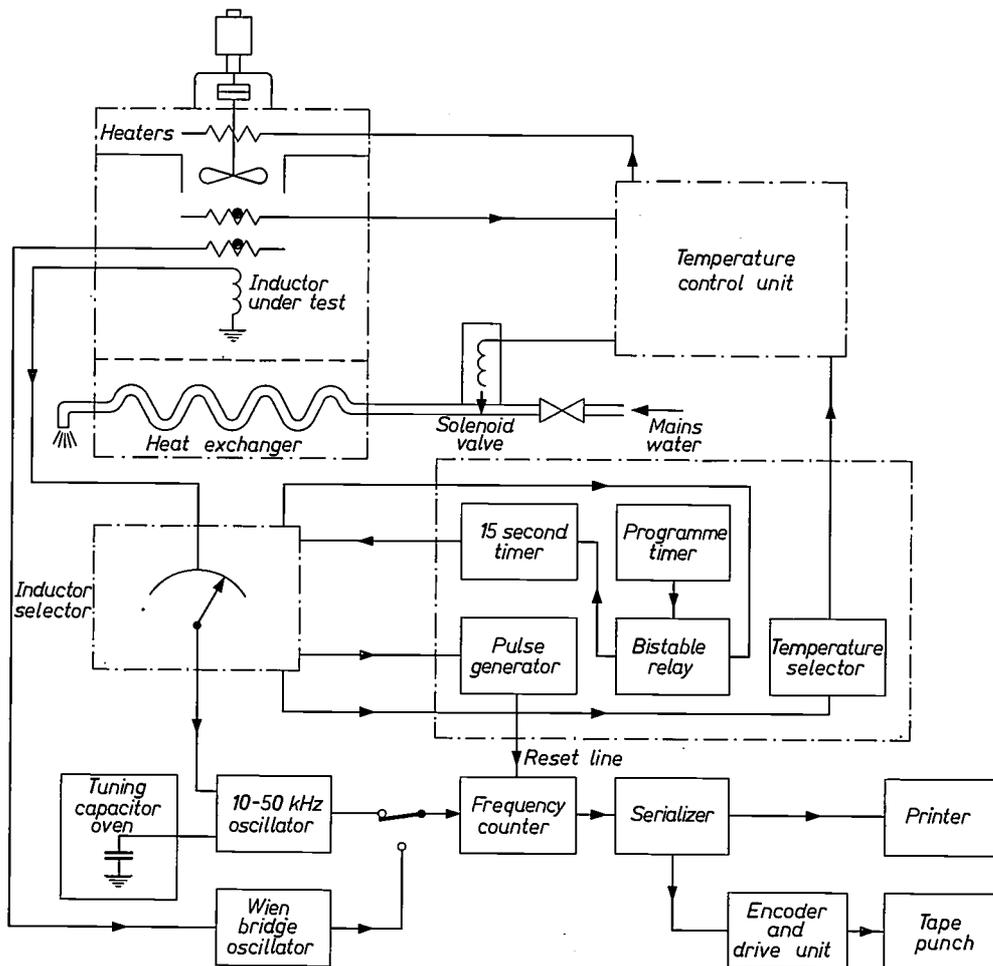


Fig. 2. Block diagram of the variability recording apparatus.

The temperature controlled chamber has a working volume of about $0.1\ \text{m}^3$. This is part of a closed circuit path around which air is rapidly circulated. In the return channel round the back and base of the chamber the air passes over the heaters and a water-cooled heat exchanger. The latter is activated at the end of the higher temperature period to increase the cooling rate. The error signal from an a.c. energized

resets the frequency counter. The 25-way inductor selector pauses for 15 seconds at each position to provide adequate time for the integrating period of the frequency counter; this has been made 10 seconds to obtain sufficient accuracy. The oven temperature is monitored by a thermistor controlled Wien bridge oscillator and the frequency of this oscillator is also recorded.

A computer programme has been provided to pro-

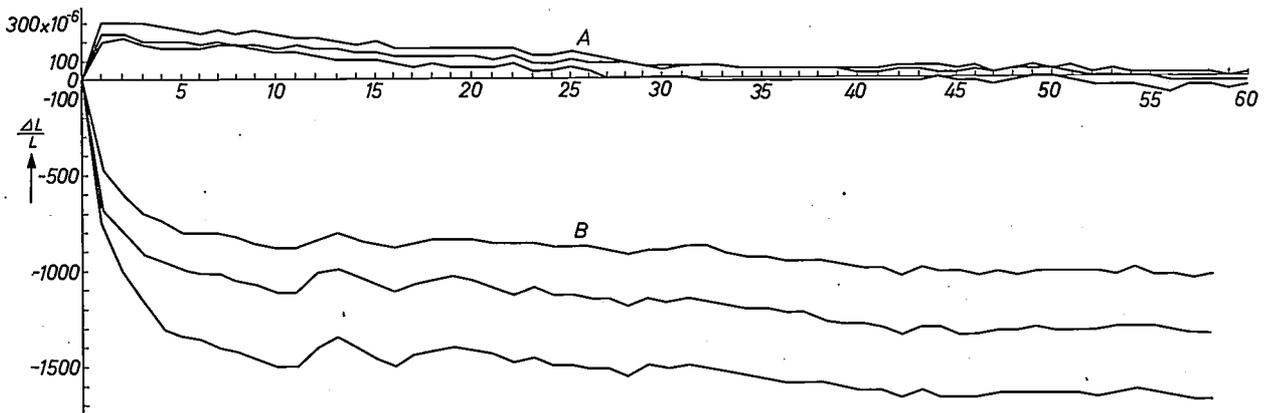


Fig. 3. A typical computer-plotted result showing the drift of inductance due to the cemented joint in 25 mm ferrite pot cores. The inductance change recorded at successive 30 °C points is plotted against the number of temperature cycles. A good adhesive, *A*, is compared with a poor adhesive, *B*.

cess the data on the tape. Basically it computes inductance change from the frequency observations and plots drift and temperature coefficient results for each inductor. It can also provide results averaged over a group of nominally identical inductors.

This apparatus has been in use for about 18 months and during this time has been engaged on the study of the contributions to inductance variability arising from non-ferrite sources. In particular it has isolated the contributions from the inductance adjusting mechanism and, as a result, an improved design has been introduced. It has also been used in an extensive assessment of adhesives used to cement the ferrite core halves together. *Fig. 3* shows a typical set of results as plotted by the computer.

A more versatile measuring chamber has recently been constructed and will be operated in addition to the one already described, the control, sequencing and recording apparatus being common. The new chamber

will allow a great variety of temperature programmes to be selected or pre-set; they may incorporate as many as twelve steps and span the temperature range -50 to $+100$ °C.

Future investigations will include a comparison of mechanical methods of fastening the pot core halves together, the effect of pressure and its point of application and the effect of the assembly technique on the permeability-temperature curve over a wide temperature range.

Summary. As ferrite development enables the Q factor of filter inductors to be progressively raised, a corresponding improvement in inductance constancy is required. There are many mechanisms by which unwanted changes of inductance can occur and it requires large numbers of high-resolution measurements to isolate and identify them. An automatic apparatus that has been specifically designed for this purpose is described and some typical results are shown.

Q contour charts

The Ferrite Section of the Mullard Research Laboratories has for some time been studying the subject of inductor Q factor analytically. *Fig. 1* shows the way in which the total loss tangent and the contributory loss tangents vary with frequency for a particular inductor design. The Q factor is the inverse of the total loss tan-

gent. If all the component loss tangents may be expressed as functions of frequency for any pot core size, material, effective permeability, type of conductor and the number of turns, the corresponding Q factor curve may be determined. Using a computer to calculate the Q factor for every combination of the above data, Q

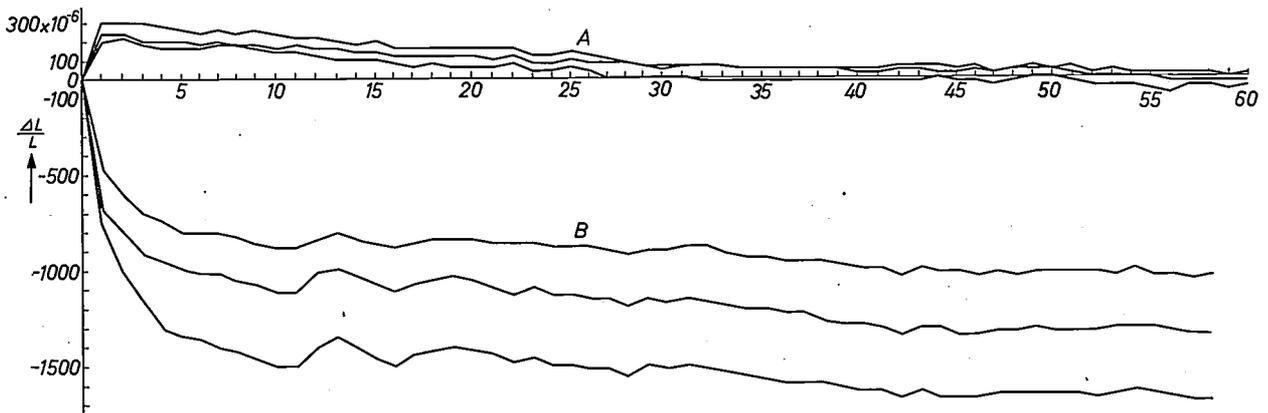


Fig. 3. A typical computer-plotted result showing the drift of inductance due to the cemented joint in 25 mm ferrite pot cores. The inductance change recorded at successive 30 °C points is plotted against the number of temperature cycles. A good adhesive, *A*, is compared with a poor adhesive, *B*.

cess the data on the tape. Basically it computes inductance change from the frequency observations and plots drift and temperature coefficient results for each inductor. It can also provide results averaged over a group of nominally identical inductors.

This apparatus has been in use for about 18 months and during this time has been engaged on the study of the contributions to inductance variability arising from non-ferrite sources. In particular it has isolated the contributions from the inductance adjusting mechanism and, as a result, an improved design has been introduced. It has also been used in an extensive assessment of adhesives used to cement the ferrite core halves together. *Fig. 3* shows a typical set of results as plotted by the computer.

A more versatile measuring chamber has recently been constructed and will be operated in addition to the one already described, the control, sequencing and recording apparatus being common. The new chamber

will allow a great variety of temperature programmes to be selected or pre-set; they may incorporate as many as twelve steps and span the temperature range -50 to $+100$ °C.

Future investigations will include a comparison of mechanical methods of fastening the pot core halves together, the effect of pressure and its point of application and the effect of the assembly technique on the permeability-temperature curve over a wide temperature range.

Summary. As ferrite development enables the Q factor of filter inductors to be progressively raised, a corresponding improvement in inductance constancy is required. There are many mechanisms by which unwanted changes of inductance can occur and it requires large numbers of high-resolution measurements to isolate and identify them. An automatic apparatus that has been specifically designed for this purpose is described and some typical results are shown.

Q contour charts

The Ferrite Section of the Mullard Research Laboratories has for some time been studying the subject of inductor Q factor analytically. *Fig. 1* shows the way in which the total loss tangent and the contributory loss tangents vary with frequency for a particular inductor design. The Q factor is the inverse of the total loss tan-

gent. If all the component loss tangents may be expressed as functions of frequency for any pot core size, material, effective permeability, type of conductor and the number of turns, the corresponding Q factor curve may be determined. Using a computer to calculate the Q factor for every combination of the above data, Q

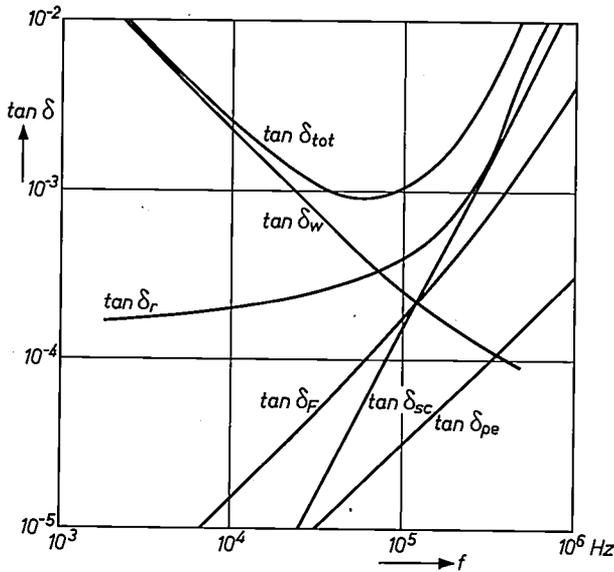


Fig. 1. Total and component loss tangents as functions of frequency f for an inductor consisting of a 25 mm pot core, effective permeability 160, fully wound with 62 turns of a bunched conductor having 140 strands of 0.04 mm diameter, inductance 2.1 mH. $\tan \delta_{tot}$ is the tangent of the total loss angle, $\tan \delta_w$ is due to the winding resistance, $\tan \delta_r$ to the residual loss in the core, $\tan \delta_{sc}$ to the loss in the winding self capacitance, $\tan \delta_F$ to the eddy current loss in the core, $\tan \delta_{pe}$ to eddy current loss in the winding.

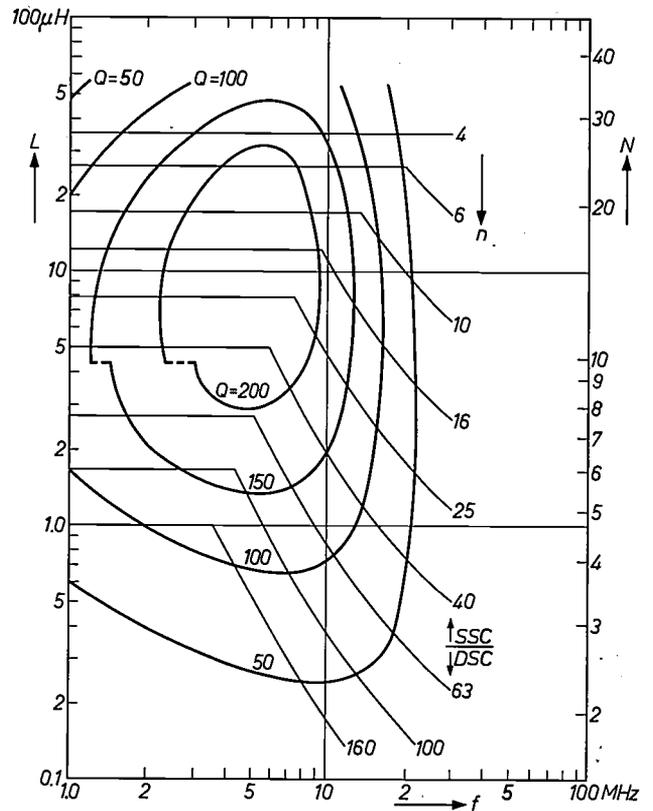


Fig. 3. Q contour chart for a high frequency inductor consisting of a 14 mm pot core, effective permeability 25, having a single layer winding of bunched conductor, strand diameter 0.04 mm. The required number of strands can be read from the over-laid curves.

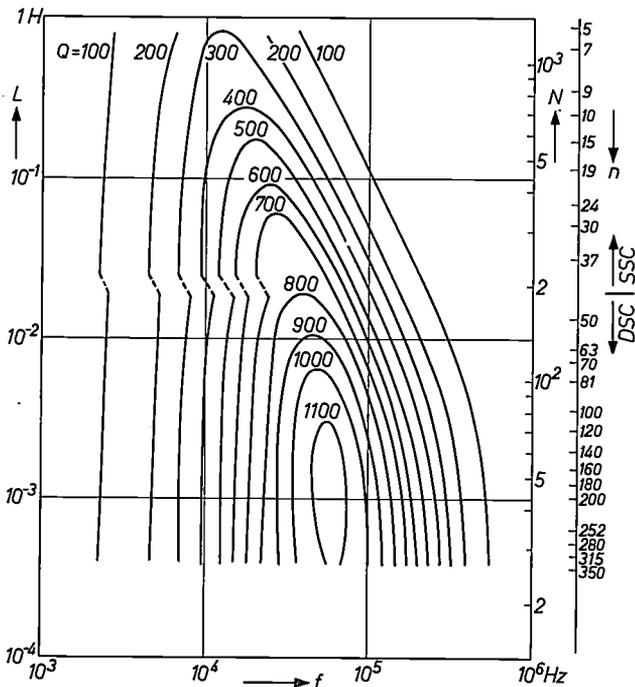


Fig. 2. Complete Q contour chart for an inductor consisting of a 25 mm pot core, effective permeability 160, fully wound with bunched conductor, strand diameter 0.04 mm. n number of strands in the conductor. N number of turns. L inductance. SSC = single silk covered, DSC = double silk covered.

curves for all likely inductor designs may be prepared. These may be presented as Q contour charts as shown in fig. 2. Over 200 such charts have been computed [1]. More recently the problem of the high frequency inductors has been studied. Postulating single layer windings, Q contour charts have been computed in which over-laid curves indicate the wire diameter, or number of strands of bunched conductors, that should be used to obtain the corresponding Q factor. Such a chart is shown in fig. 3.

E. C. Snelling

E. C. Snelling, B.Sc. (Eng.), F.I.E.E., is with Mullard Research Laboratories, Redhill, Surrey, England.

[1] E. C. Snelling, Ferrite-cored inductors and transformers, Mullard tech. Comm. 9, 30-42, 1966 (No. 82).

Critical parameters of evaporated NiFe films for fast stores

R.V. Peacock

Introduction

In the computer field much emphasis is placed on increasing the speed of machines. Significant advances in this direction can only be made if the operating speed of the memory unit is considerably improved.

The thin magnetic film is a storage element capable of intrinsic switching times at least an order of magnitude shorter than those of the best ferrite cores. Hence its use can result in a major advance in computer memory speed.

The magnetic film consists of a layer of magnetic alloy typically $1.0 \text{ mm} \times 0.5 \text{ mm}$ and about 100 nm thick. Arrays of these films are made by vacuum evaporating a nickel-iron alloy through a mask on to a heated substrate in the presence of an applied magnetic field. This magnetic field builds into the deposited film a preferred direction of magnetization, called the "easy" direction. Thus the stable direction of the magnetic vector within the film is parallel to this easy direction and is either pointing \rightarrow or \leftarrow corresponding to a "1" or a "0" state. The direction perpendicular to this easy direction is known as the "hard" direction.

The aim of this paper is to show the influence of film parameters on the store design, and to discuss the many compromises with which a store design engineer is faced.

Critical film parameters

It appears at the present time that most advantage can be made of the fast switching properties of thin magnetic films as computer storage elements if the matrix of films, which together make up the store, is operated in the "orthogonal word-address mode". This method of storage has been described in the literature^[1]; briefly the system is as follows.

In the film matrix shown in *fig. 1* each binary computer word is stored in one horizontal row of films (*fig. 2a*). Read-out of a complete word is achieved by passing current down a word conductor which passes over each film parallel to its easy direction; the resultant field causes the magnetization of each film to rotate into the hard direction (*fig. 2b*). This rotation induces a signal on a sense conductor which passes over each film parallel to the hard direction. The polarity of the signal produced depends on whether a "1" or a "0"

was originally stored. The output signal is amplified to the level required for feeding subsequent logic circuits by means of a read amplifier and stored in a staticizer.

While the films are magnetized in their hard directions under the influence of the word field, pulses of current are passed along the digit lines, which pass above the films in a direction at right angles to the word lines. The fields produced by these digit currents cause the magnetization of each film to tilt away from the hard direction (*fig. 2c*) so that when the word field is removed the magnetization relaxes leaving a "1" or "0" stored, depending on the polarity of the digit current (*fig. 2d*). It is clearly necessary to ensure that the digit pulses, applied to write information into the word selected, shall not cause switching of films in unselected words.

From this method of operation we can see that a store designer is primarily concerned with six critical parameters of the film elements:

- the minimum word current in a strip passing over the element which will allow satisfactory reading and writing (this affects his word circuit design),
- the film output for a given rise time of this word current (this influences his read amplifier design),

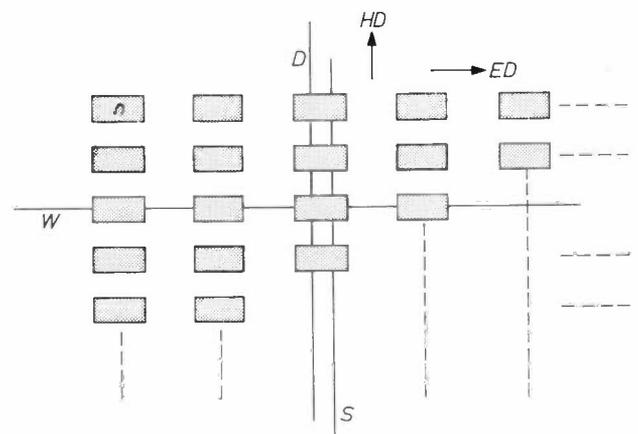


Fig. 1. Schematic diagram of array of magnetic films n , showing how each binary word is stored in one horizontal row of films. ED easy direction and HD hard direction of magnetization. Current from the word conductor W passes over a complete row of films, parallel to their easy direction to enable binary information stored in the film to be retrieved. The polarity of the signal induced in the sense conductor S determines whether "1" or "0" was originally stored in the film. Information is written into the film by passing current down the digit conductor D .

- c) the minimum digit current which will guarantee that information of either polarity can be satisfactorily written into all films in the store (this tells him the minimum current requirements of his digit driver circuit),
- d) the maximum tolerable digit current which will not destroy information in unselected films in the store during the worst possible sequence of operations on other films,
- e) the absolute and percentage difference between (c) and (d) (this influences the digit current design margins),
- f) the size of the element (as this influences not only the final size of the store, but also the time which has to be allowed for signals to pass to and from the selected element).

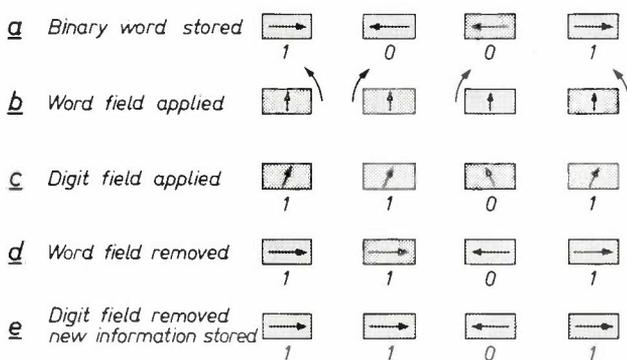


Fig. 2. Orthogonal word-address mode of storage.

These parameters are functions of:

- 1) film geometry (length, width and thickness),
- 2) the anisotropy field, H_K , a field that has to be applied in the hard direction to overcome the anisotropy,
- 3) the dispersion of the anisotropy in both direction and magnitude,
- 4) the internal demagnetization fields,
- 5) the deviation or skew of the easy direction from the intended direction,
- 6) the "creep" threshold measured with the highest hard direction pulsed field which will be encountered in the operation of the store.

As a number of the critical parameters (a) to (f) are functions of many of the above variables (for example, (c) is a function of (1), (2), (3), (4) and (5)), the design engineer is confronted with a complex set of simultaneous relationships. If the variables (1) to (6) were independent and the engineer knew to what extent he was able to compromise between the parameters (a) and (b) for example, the problem would not be too intractable. However, even for a single film composition, a number

of variables are related in complex ways. For example (2), (3), (5) and (6) are related both to the evaporation conditions (substrate temperature, vacuum, evaporation rate, deposition field) and to the substrate properties. It is, therefore, difficult to arrive at an "optimum" film.

It is the purpose of this paper to give a résumé of the more significant relationships relating the critical parameters (a) to (f) with the variables (1) to (6).

The word current

The current required to read-out a word satisfactorily is related chiefly to the film geometry and the material anisotropy. The influence of film geometry and the variations of film anisotropy with composition and evaporation conditions are discussed in the following sections.

Relationship between film geometry and word current

One might expect that, for a film of given material with a given anisotropy, the field which has to be generated by current in the word conductor would be independent of the width of the film in the hard direction. In fact this is not true for two reasons. First, it does not allow for the bending or curling of magnetization near the edges of the film which occurs when the bulk of the magnetization is turned into the hard direction [2] and secondly it does not take into account shape anisotropy.

A comprehensive series of measurements which was carried out in these Laboratories [3] on quasi-elliptical films shows that the word field required to give 80% of the maximum possible output increases above the intrinsic material anisotropy field by some oersteds for very small elements ($\approx \frac{1}{4} \times \frac{1}{2}$ mm), *fig. 3*, and the behaviour is in good agreement with the theoretical relationship derived by Bonyhard-Buckingham [2] for this curling of magnetization at the edge of a film.

Now the shape anisotropy field even for small elements with extreme aspect ratios is only around 1 oersted. This is, in the most significant cases, only about 20% of the effect due to curling of magnetization. Therefore, one can conclude that because of the non-uniform magnetization distribution in a film during read-out, the hard direction width of the film has a very significant and predictable effect on the word field necessary for satisfactory operation.

[1] J. I. Raffel, T. S. Crowther, A. H. Anderson and T. O. Herndon, Proc. IRE 49, 155, 1961.
E. M. Bradley, J. Brit. I.R.E. 20, 765, 1960.

[2] P. I. Bonyhard and I. C. Buckingham, IEEE Trans. on magnetics MAG-1, 258, 1965.

[3] M. J. Folkes and G. Winsor, Int. J. Control 3, 513, 1966 (No. 6).

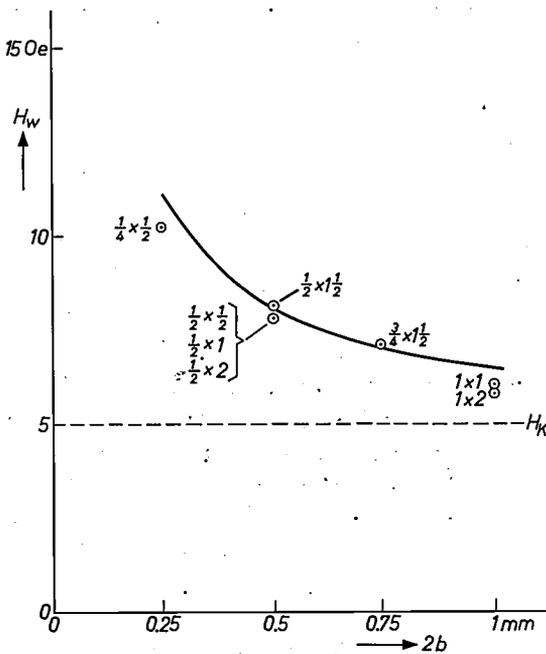


Fig. 3. Word field H_w (in Oe) required for setting a fraction 0.8 of the film in the hard direction, as a function of the hard direction film width $2b$ (in mm), for a film with an intrinsic anisotropy field of 5 Oe, saturation magnetization 800 e.m.u. and thickness 85 nm. The circles are measured values for a number of quasi-elliptical elements. The curve shown is calculated using the relation derived by Bonyhard and Buckingham [2].

Effect of composition and evaporation conditions on film anisotropy

For a given film size, the current needed to read-out the information is a function of the intrinsic film anisotropy. The variation of anisotropy with composition for NiFe alloys is now well-known [4], *fig. 4*. This indicates that one should from this point of view operate in the 80-90% Ni region. However, there are other considerations (e.g. magnetostriction and the related question of anisotropy induced by the angle of incidence of the vapour beam during deposition) which give a strong preference for the lower end of this range, and most storage element work is carried out around 80-82% Ni.

Even at a given composition the anisotropy created in the film is a function of the temperature of the substrate during the deposition process. Generally, the induced anisotropy decreases with increasing substrate temperature, but there is evidence that above about 400 °C it rises again steeply [5]. The temperature at which the anisotropy is a minimum, is a function of evaporation rate; the lower the rate, the higher the optimum substrate temperature.

For practical storage elements, however, these high optimum temperatures are not generally used as the dispersions and film skew are bad at very high substrate temperatures.

It is clear from the above that although the way of producing low anisotropy fields in evaporated NiFe films is known, the store designer is forced to compromise between having the lowest anisotropy on the one hand and the extra skew, dispersion and magnetostriction which such a low figure produces. It will be seen in later sections that there are many areas where similar compromises have to be made in magnetic thin film store design.

Film outputs

When the magnetization in a thin film is turned into the hard direction by a rapidly applied magnetic field (say with a rise time $\tau_r = 0.25$ ns) the magnetization precesses in a flat elliptical spiral to its final position [6], and the output on a loop passing round the film parallel to the hard direction shows a clear damped oscillation, *fig. 5*, which is in good agreement with the behaviour expected. For fields of the order of the anisotropy field, H_K , or a few H_K which are applied at a lower rate ($\tau_r \geq 5$ ns say) the magnetization virtually "follows" the field, and at any given field value is

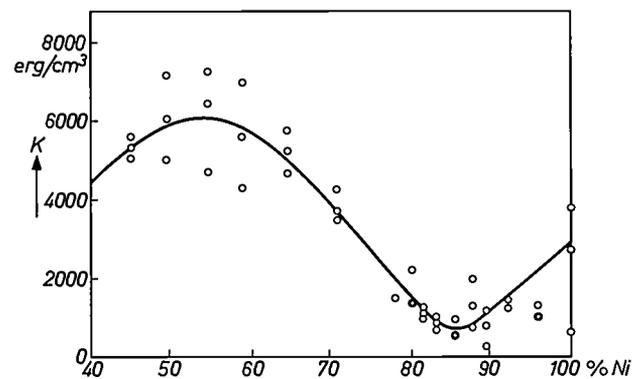


Fig. 4. Anisotropy constant K plotted against %Ni in the NiFe film. Note minimum K between 80% and 90% Ni composition.

in the position which it would occupy if that field were steady. It is clear, therefore, that on reading-out a film by turning the magnetization, or a fixed fraction of it, into the hard direction the peak output voltage induced in the pick up system is inversely proportional to the rise time. The choice of this rise time by the store designer is a rather sophisticated engineering compromise,

[4] G. Robinson, Proc. Leuven Conf. on the electric and magnetic properties of thin metallic layers, p. 140, 1961.

[5] J. H. Engelman and A. J. Hardwick, Trans. 9th Nat. Vac. Symp., p. 100, 1962.

[6] B. R. Hearn, J. Electronics and Control 16, 33, 1964.

[7] B. R. Hearn, Proc. Leuven Conf. on the electric and magnetic properties of thin metallic layers, p. 167, 1961.

[8] W. Metzdorf, Siemens and Halske Report, 1963.

but for current thin film store work it is usually in the 5-30 ns region.

With a given rise time, one would expect that the output would be proportional to the cross-sectional area of the film in the hard direction. Now the upper thickness of a storage film is limited by other parameter requirements to around 100 nm or less. Consequently one is only at liberty to vary the hard and easy direction

necessary digit field are discussed in the following sections.

The influence of film shape

The influence of the easy direction demagnetization field H_d of an unskewed film on the apparent dispersion has been dealt with by Metzdorf [8]. Imagine a film under the influence of a hard direction word field and

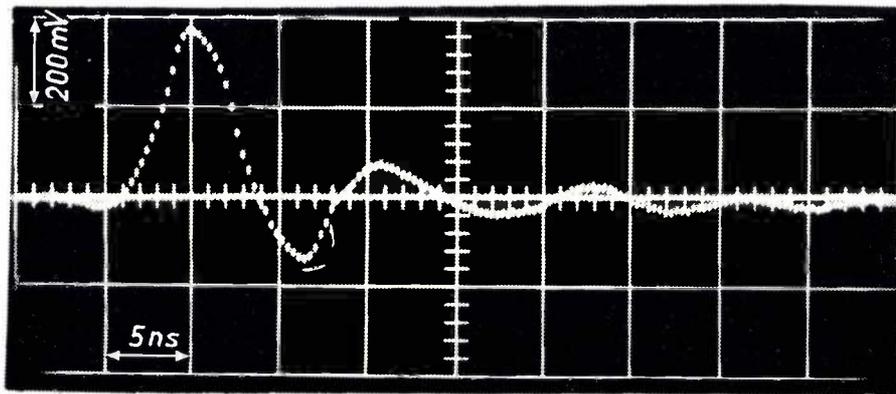


Fig. 5. Typical hard direction switching waveform of an 80/20 NiFe film. The field applied was 7 Oe and the H_K of the film was 4 Oe. The vertical scale represents the film output voltage.

lengths. Fig. 6 shows the output of quasi-elliptical films with various sizes and shapes where a fixed fraction (≈ 0.8) of the film turns to the hard direction with a word field rise time of 20 ns. As expected, the points fall roughly on a straight line through the origin.

It should be mentioned here that because of the dispersion of the magnitude of H_K which exists within a film [7] and because of the curling mechanism previously mentioned, it is usually necessary to apply word fields considerably in excess of the measured intrinsic anisotropy field of the film before the output approaches the maximum value.

The minimum digit field required to write into a film

In an ideal film one would imagine that with a word field greater or equal to the film anisotropy field, an infinitesimally small digit field would be sufficient to tilt the whole of the magnetization into the selected easy direction. In a real film this is not true. Finite fields are needed for three reasons:

- a) There is an influence of film shape and demagnetization fields.
- b) The mean easy axis may not be parallel with the geometrical axis.
- c) There is an effect due to dispersion in the easy axis within a film.

The relationships between these three factors and the

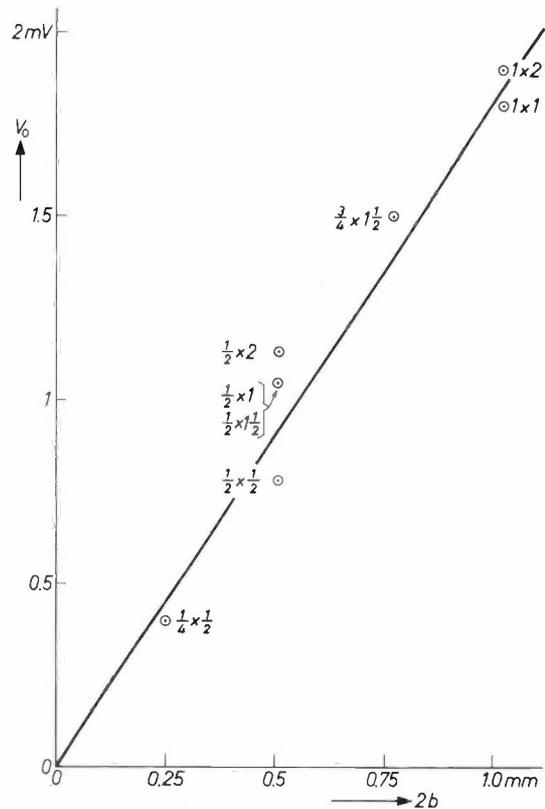


Fig. 6. Measured values of output. The output voltage V_0 when a drive pulse of 20 ns rise time is applied is plotted as a function of the hard direction width $2b$ of the film. As expected, the relation is linear.

an easy direction digit field. The resultant effective field acting on the material of the film will be the vector sum of the applied fields and the demagnetization field of the film; the switching behaviour of the film will depend upon how the amplitude and direction of this resultant field is related to the switching threshold of the intrinsic film material.

A simplified calculation [3] to derive this digit field shows that, for an unskewed film with low dispersion of the easy axis within a film, the digit field necessary to overcome this effect is proportional to $H_a^{3/2}/H_K^{1/2}$. This relationship seems to be valid within the limits of experiments.

Paradoxically this means that the *larger* a film is made in the easy direction, and therefore the *wider* the digit conductor, the lower is the current required. This might well lead one to assume that there is a real advantage in practical memories in having films which are long in the easy direction. This is not in fact true because of skew. This will be discussed in the next section.

The influence of the skew of the easy axis

If an ideal film with no angular or magnitude dispersion of its anisotropy were deposited such that its mean easy axis is skewed at an angle θ from the digit field direction, then the lowest digit field which could write information into that film would be $H_K \sin \theta$. If, in addition, there is dispersion of anisotropy magnitude or direction, the necessary field is further increased.

Skew of a film in an evaporated matrix is chiefly a function of the angle of incidence of deposition on to the surface. The degree to which this angle of incidence affects the easy direction of a film depends on the magnetostrictive constant of the particular alloy used [9]. The main effect is to induce an additional anisotropy in the plane of the film orthogonal to the incident vapour beam. Close examination of the structure of such films in these Laboratories has shown that there tends to be a "grain" within the film orthogonal to the direction of incidence [10], fig. 7.

This additionally induced anisotropy changes both the magnitude and, more important, the direction of the mean anisotropy field within a film. Fig. 8 shows the skew angles induced at the corners of a 5 cm square centred on the axis with the source 25 cm distant as a function of composition with two different substrate materials.

The influence of angular dispersion of the anisotropy axis

Even in a large (low easy direction demagnetizing field) unskewed film it would be necessary to apply a finite digit field to ensure that a large fraction (e.g. 80%) of the magnetization fell into the intended easy direc-

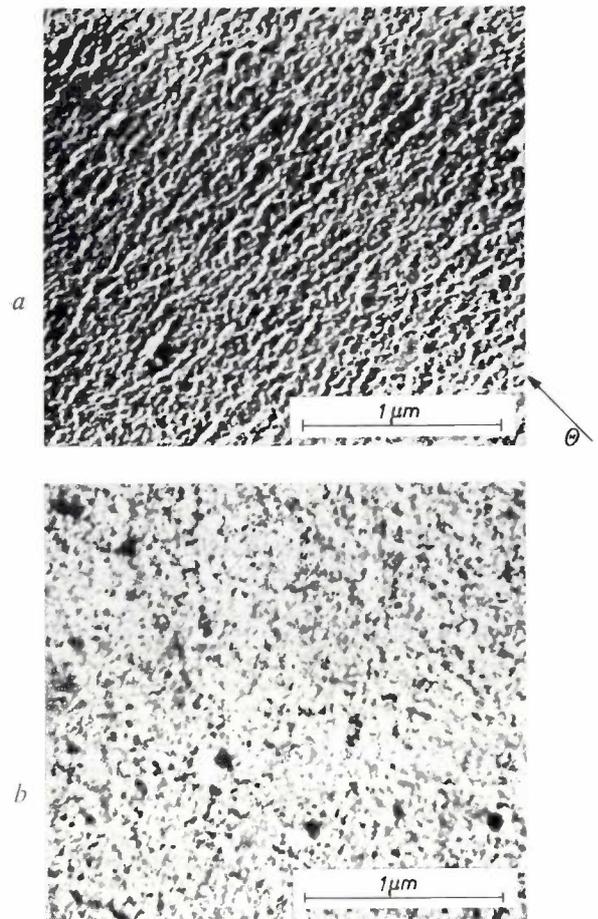


Fig. 7. Electron microscope study of Ni film structure on silica substrate, a) at 200 °C and $\Theta = 75^\circ$, b) at 150 °C and $\Theta = 0^\circ$, where Θ is the incident beam direction.

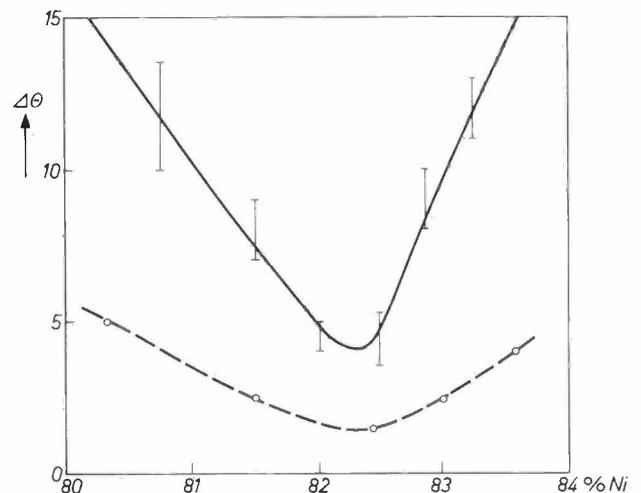


Fig. 8. Range of skew angles $\Delta\theta$ plotted against %Ni in the film. Drawn line: 0.125 mm glass substrate, with dispersions of measured values indicated by the vertical lines. Dashed line: 6 mm aluminium substrate. Note the minimum in $\Delta\theta$ at 82-83% Ni.

tion when the word field was removed. This would be because of variations of the direction of the uniaxial anisotropy within one film which would have to be overcome by the digit field.

The magnitude of this dispersion as a function of composition of NiFe films is shown in *fig. 9*. It can be seen to drop to less than 1° and have a broad minimum over the range 65 to 85% Ni. It is well-known that the precise value of this parameter is related to evaporation

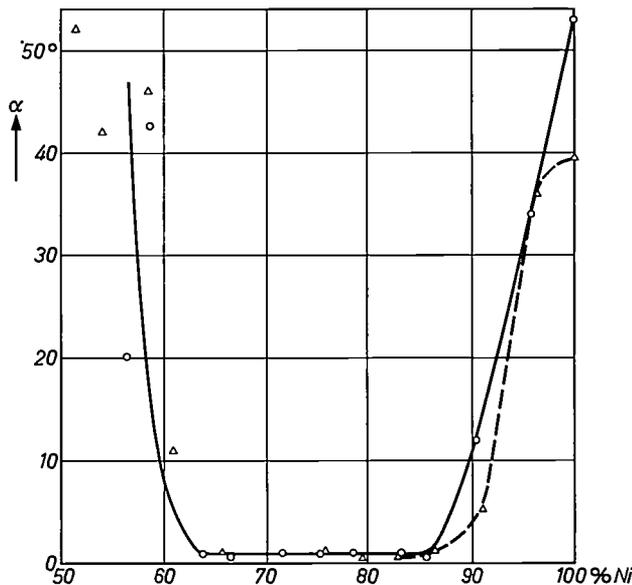


Fig. 9. Easy direction dispersion α plotted against %Ni in the film. Circles for 100 nm and triangles for 55 nm film thickness.

conditions [5] and the film structure when it starts to build up on the substrate [11]. There is evidence that at low evaporation rates (≈ 1.5 -5.0 nm/s) the angular dispersion increases with increasing substrate temperature, but that it is relatively constant, although at a higher value, for more rapid evaporations. It is also a function of grain size [12].

In spite of a considerable amount of attention no one model of the mechanism adequately describes the behaviour of films regarding angular dispersion over the whole composition range. At the extremes of composition the random uniaxial model is reasonably satis-

factory, but over the range of most interest to computer store designers the ripple model seems to be best [13].

In general, therefore, the angular dispersion obtained in a film is to a large extent both controllable and predictable in a purely empirical way but is not altogether explicable.

The upper limit imposed on the digit field

In normal word organized film stores, digit fields are applied to unselected films in the store. Clearly these must not be greater than the easy axis coercivity of the films or unselected information may be disturbed. This disturb threshold is clearly a function both of film geometry and the material coercivity. The variations of disturb threshold with film geometry which are open to the store designer are discussed in the next section.

In a real store, additional small stray pulsed hard direction fields are unintentionally applied to the films while the digit field is present. These stray fields come from small unwanted currents down unselected lines, fields from nearby selected word lines and the field of neighbouring films which are read-out. It is known that a small pulsed hard direction field greatly reduces the threshold at which the magnetization (and hence the stored information) may be disturbed by a digit current [14]. When the magnetization is disturbed in this way it changes by the movement of domain walls in a sequence of steps. This phenomenon which has been investigated in these and other laboratories is known as creep [15].

The conditions under which creep occurs can be defined in any type of film and it need not present a real problem in the design of magnetic film stores. Satisfactory operation can be readily obtained provided that the packing density of films in the store is limited to minimize stray hard direction fields from adjacent words and the word circuits are carefully designed to minimize small unwanted currents. The next two sections will consider the influence of film size and shape on both the ordinary disturb and the creep threshold, and will give some idea of the influence of film thickness.

The influence of film geometry on disturb and creep thresholds

For a given intrinsic coercivity the larger the easy direction demagnetization field the lower the external field needed to disturb the magnetization. *Fig. 10* shows the measured variation of disturb threshold as a function of easy direction length for a range of films. The experimental values are shown as dots and the calculated values as dots within circles. The calculated values were derived by subtracting the easy direction demagnetization field, assuming the films to be ellipsoids,

[9] D. O. Smith, M. S. Cohen and G. P. Weiss, Lincoln Labs. Report 53G-0037, 1960.

[10] S. G. Fleet, M.R.L. Report No. 466, March 1963.

[11] K. D. Leaver, *Nature* 196, 158, 1962.

[12] H. Hoffman, *Z. angew. Physik* 18, 499, 1965.

[13] J. A. Cundall and A. P. King, M.R.L. Report No. 565, Dec. 1965.

[14] T. H. Beeforth, M.R.L. Report No. 504, Jan. 1964.

S. Middelhoek, I.B.M. Research RC846, Dec. 1962.

P. J. Hulyer, M.R.L. Report No. 505, April 1964.

[15] T. H. Beeforth and P. J. Hulyer, *Nature* 199, 793, 1963.

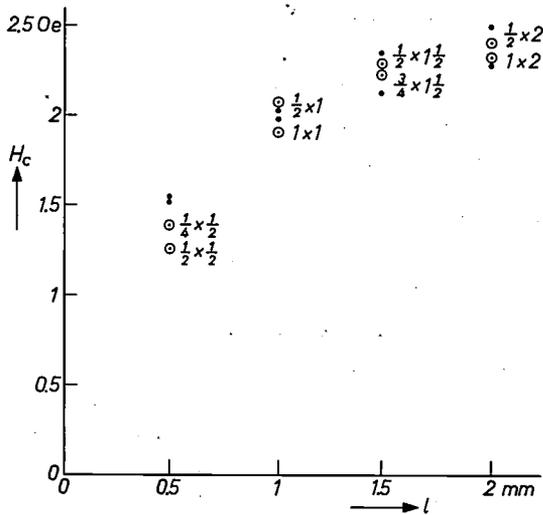


Fig. 10. Disturb threshold values H_c as a function of film length l in the easy direction. Measured values are shown as dots whilst dots within circles show calculated values.

from the intrinsic coercivity of the film (in this case 2.6 Oe). The good agreement is similar to that found by Coren [16].

If we now consider the creep threshold, the results are rather different. Let us examine the variation with film size when currents, equal to the word current, are passed down adjacent word lines to provide a hard direction pulsed field [3] (as is the case of a real store). Let the spacing between films in the hard direction be equal to one hard direction width.

In fig. 11 the median values of creep threshold are plotted as a function of the easy direction length of the film. Comparison with fig. 10 shows that this threshold is less than the disturb threshold (H_c) for films which are short in the hard direction but similar to H_c for the widest films used.

The results are simply explained. The value of creep threshold obtained in any particular case is the disturb threshold value reduced by an amount which depends on the hard direction disturbing field. For elements with short hard direction widths, the field from the adjacent word conductor is larger for two reasons. First, the actual word fields necessary are larger for small elements. Secondly, although the elements were spaced by their own dimensions in the hard direction, the separation of the word conductor from the ground plane was constant. Thus the field at the adjacent film, due to current in the word line, is larger for the narrow strip lines appropriate to films having a small hard direction width.

It is, of course, clear that with different film and line spacings the results are different but it serves to illustrate that the store designer has a more difficult

task regarding creep thresholds with small films than with large ones. The compromise between having tight digit current margins on the one hand, and large store size and long transit time delays on the other, again faces him.

Digit current margins

As mentioned previously, in designing a digit current generating circuit, the circuit must not produce a current so low that it does not write information into the worst film in the store on which it has to operate, nor must the current be so high that it disturbs the film with the lowest disturb threshold. The smaller the difference between these two quantities, the more difficult the circuit design and the more expensive the resulting

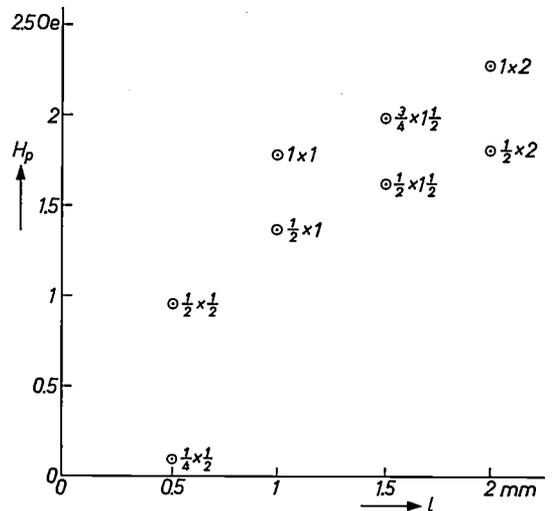


Fig. 11. Creep threshold H_p as a function of film length l . It will be seen that the creep threshold increases with film size. Comparison with fig. 10 shows that this threshold is less than the disturb threshold for films which are short in the hard direction but similar to H_c for the widest films used.

electronics. It is the difference between these two quantities which lies at the heart of the majority of compromises which have to be made.

For example, for economic reasons, the store designer likes to make very large matrix arrays at one time. However, with a given film size the larger the array the worse becomes the skew of the corner elements and the higher becomes the minimum allowable digit current, and hence the tighter the digit current margins. He would also like to reduce the easy direction length to increase the packing density, but for a given film thickness this would raise the necessary

[16] R. L. Coren, Proc. Interimag. Conf. 1965.

minimum and lower the permissible maximum current, again tightening the tolerances on the digit circuit.

It is clear that this is an area where compromises have to be made by the store designer. However, quantitative information is required to enable such decisions to be made, and the margins between the necessary minimum and the permissible maximum digit current as a function of film geometry and skew are considered below.

The influence of film size and skew

If a quality factor, Q , is defined regarding the question of digit current margins, it is logical to define the difference between the upper (permissible) and lower (necessary) limit value not in absolute terms but related to the mean current. Thus:

$$Q = \frac{\text{max. perm. dig. curr.} - \text{nec. min. dig. curr.}}{\text{mean digit current}}$$

One can then examine the influence of film shape and skew on the value of Q . This has been done [3] using disturb, rather than creep, thresholds and the results are shown in *fig. 12*. It is clear that, as expected, Q increases rapidly, giving wide circuitry margins, as the easy direction length of the element increases but this effect is less significant with highly skew films.

It is this type of information which enables the designer to calculate the most suitable compromise between circuitry margins, film size, array size and delay of signals in the stack. However, the compromises are not simple and one cannot arrive at a generally optimum film design. Consequently, store development must take place by progressive improvement of the design in an engineering manner.

This paper has attempted to outline some of the relationships which allow this evolution of design to take place.

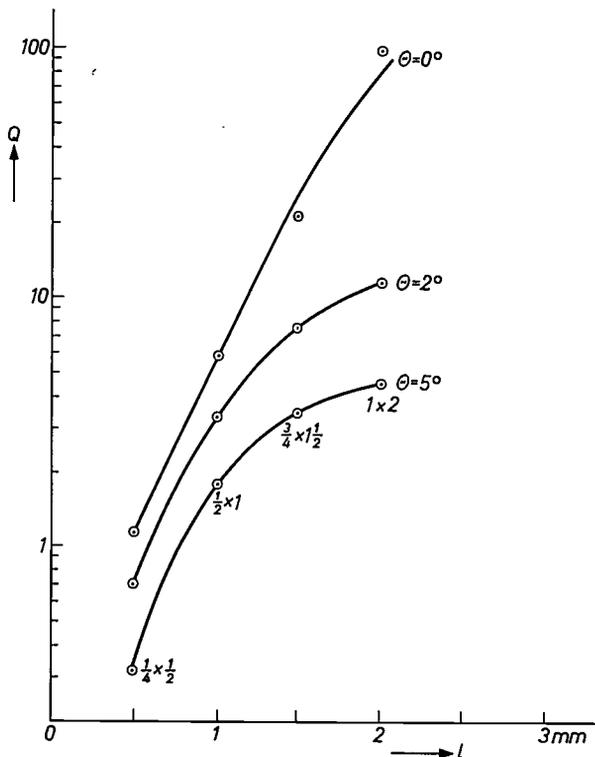
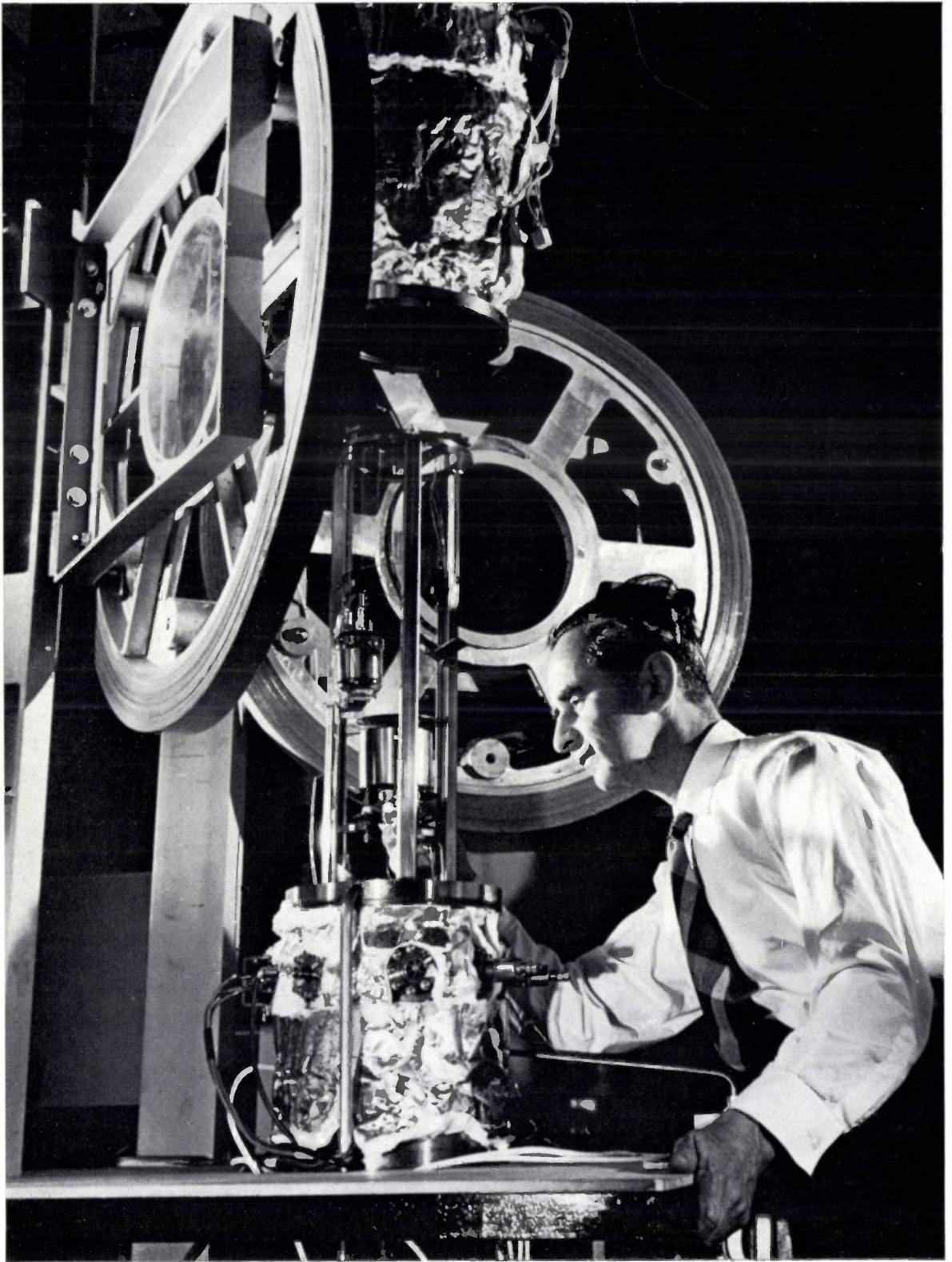


Fig. 12. Quality factor Q as a function of easy direction length l for different skew values θ . Note how Q increases rapidly as the film size increases.

Summary. The paper aims to show the influence of magnetic film parameters on the design of fast thin film stores. A number of factors are considered; just one example is the effects of film size and shape on the value of digit current and its associated margins, the output, the minimum word current and the packing density. In deciding upon values to be given to almost all film parameters, compromises are necessary and these are of a complex nature. The conclusion is reached that at present a generally optimum film design is difficult to achieve and that store development must take place by progressive improvement of the design in an engineering manner.

Magnetic thin film deposition



Photograph Walter Nürnberg

In the production of magnetic thin film storage planes it is necessary to use a magnetic field during the evaporation process to give the elements uniaxial anisotropy. The elements are formed by high vacuum evaporation of nickel-iron through a mask on to a substrate heated to about 300 °C. The photograph shows an experimental high vacuum plant for the evaporation of nickel-

iron at pressures in the range 10^{-9} torr. To outgas the chamber the bell jar of this apparatus was heated to about 300 °C. Foil covered heat insulating material was wrapped round the outside surface to reduce heat loss. The large Helmholtz coils visible in the photograph are swung out of the way to facilitate the cleaning and loading of the plant. In use they are parallel one to the other.

Optical character recognition

P. Saraga, J. A. Weaver and D. J. Woollons

Introduction

It has become increasingly apparent during the last twenty years that the digital computer is capable of far more sophisticated operations than the straightforward arithmetic calculations which were performed on the early machines. It has also become apparent that for many of these applications the usefulness of the computer is limited by the efficiency of the devices available for enabling human beings to communicate with the machine. This is particularly true in some data processing problems where the amounts of data entering and leaving the computer are large but where relatively simple operations are performed in the central processor. A significant advance would be made in many such problems if it were possible for the computer to read directly machine- or hand-printed alphanumeric characters.

There are two classes of character recognition problem. The first of these is concerned with the reading of information produced in a controlled environment. That is to say the production of the characters to be read is under the control of the organization using the reading machine. In this case the characters may be printed in one of the type founts specially designed for optical reading. The two most widely accepted founts at present are known as O.C.R. A and O.C.R. B ^[1], the former being the standard proposed by the American Standards Association and the latter being the European equivalent.

The second application for character recognition is in the reading of characters from an uncontrolled source. Perhaps the best example of this is in the automatic sorting of mail, where the characters to be read are originated by the general public. This is a much more difficult problem than the former one, and its solution will necessarily involve a more expensive reading machine.

The Mullard Research Laboratories approach

The work undertaken so far on optical character recognition at Mullard Research Laboratories has been on the application in which the origin of the characters to be recognized is uncontrolled. In particular, the re-

search has been on the recognition of machine-printed and hand-printed characters.

The main problem in formulating a character recognition process is that of making the process insensitive to the various mutilations and distortions of the character which may occur. These mutilations, in general, may take the form of breaks in the character, variations in size, position, or angular rotation of the character, smudges, creases, overprinting of various types, and general deterioration in quality due to handling of the document bearing the character.

The types of distortion which occur in characters are dependent, in part, on the method by which the character is produced. For example, hand-printed characters are produced by strokes of a pen or other writing implement and are, therefore, less subject to breaks in the character strokes than typed characters which are produced by a type hammer striking the paper.

Thus it may be advantageous to tailor the recognition method to suit the origin of the characters. A line following method, for example, is more suitable for hand-printed characters but for typed characters some sort of masking system may be more successful.

Because of the inevitable complexity of the systems needed to achieve character recognition, even to a limited degree, it is more economical to investigate the properties of such systems by computer simulation than by the construction of special purpose machines. Investigation by simulation methods also possesses the advantage that the main research, that is work on the properties of character recognition processes, is not impeded by the technological difficulties which may beset the construction of any hardware embodiment of a system. These difficulties must, however, not be ignored especially if a practical recognition machine is the ultimate aim of the research.

Special on-line equipment

In order to simulate a character recognition process on a digital computer, some method must be devised for representing the character to be recognized in a form which is suitable for the computer. This is nor-

P. Saraga, B.A., J. A. Weaver, B.A., and D. J. Woollons, Ph.D., are with Mullard Research Laboratories, Redhill, Surrey, England.

^[1] Some other forms are mentioned in W. Nijenhuis and H. van de Weg, Developments in the field of electronic computers during the last decade, Philips tech. Rev. 26, 67-80, 1965.

mally achieved by representing the input character as a matrix of points. Each point in the matrix has a value determined by the "blackness" of the matrix at that point. The number of levels of blackness may be made as large as is desired but a larger number of levels necessitates more storage space for the matrix in the computer. The simplest arrangement, and the one which requires least storage space, is to use a binary matrix, that is, to call each point either black or white. This is the method that has been adopted.

Some equipment must, therefore, be provided for converting characters printed or written on paper into a suitable matrix form. This may be achieved by using a flying spot scanner, scanning the input character, to control a paper tape punch or a card punch. Thus a representation of the character is produced on paper tape or cards. These may be entered into the computer through the standard readers.

This process however is very time consuming. A better approach is to connect the flying spot scanner directly, that is on-line, to the computer. This has been done at Mullard Research Laboratories. An example of a typical character represented in the form in which it is stored in the computer is shown in *fig. 1a*. As may be seen from the figure, the matrix containing the character consists of a sequence of ones and zeros. These are packed into the computer memory in sequence, with one matrix point occupying one binary digit (bit) of the memory. Two complete computer words are allocated to each line of the matrix. Thus if, for example, the matrix size is 60×60 points, and if the computer word length is 38 bits, each matrix line is stored in all the 38 bits of the first word allocated to it and in 22 bits of the second word. The remaining 16 bits of the second word are not used. The complete matrix, therefore, occupies 126 words of the computer memory.

Character recognition processes deal largely with operations on pictures of characters stored, as described above, in the computer store. It is, therefore, necessary to provide means whereby such pictures can be conveniently put out from the computer and displayed to the operator. In this way it is possible to put out intermediate results from the recognition process in order that the operator may be able to determine how it is progressing. Accordingly an on-line computer output cathode ray tube display has been constructed.

Character recognition systems

Most optical character recognition systems can be divided into three sections which may be called the receptor, the preprocessor and the classifier.

The receptor converts the character, printed on paper, into a suitable input form of the type previously

described. The purpose of the preprocessor is to condense and select that information in the signal from the receptor which is useful for classification. The output from the preprocessor is a processed pattern which forms the input to the classifier. The classifier then assigns the processed pattern to a class.

The receptor

The receptor must perform three functions. It must locate the character to be fed into the computer; it must set the threshold level, which is the level of decision above which a point is white and below which it is black; it must then scan the character in such a way as to extract the required information from it.

The document bearing the characters to be read is mounted on a paper handling drum located beneath the scanner. The angular position of the drum can be controlled from the computer. The computer may determine the drum position by reading a clock track on

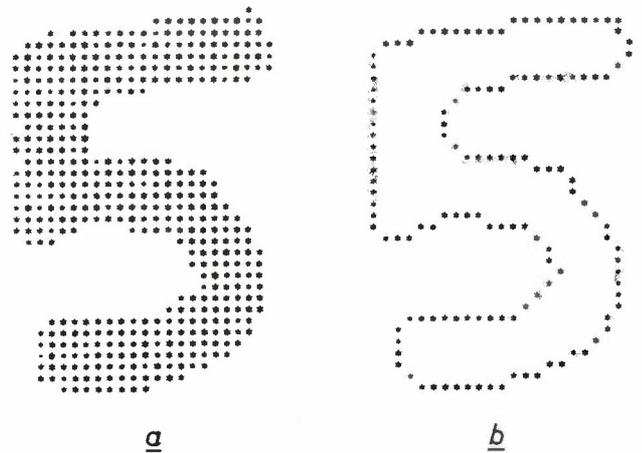


Fig. 1. *a*) A typical character as stored in the computer. *b*) The path of the tracing element round the edge of the character.

one edge which divides the drum into 256 angular increments.

A programme has been written to perform the three functions mentioned earlier in the special case of a single column of characters. At present it is not possible to search along a line of characters but this is a modification which, in principle, could be simply incorporated when it becomes necessary.

The various parameters used in the process are set automatically by reference to the size of the first character encountered by the process. Initially, the threshold level is set by hand to give a reasonable representation of the first character, as displayed on the cathode ray tube display. This is only necessary when the process is entered for the first time but must be done then to enable the process to locate the first character. Subsequently the correct threshold is automatically recalculated for each character the process meets. This

is necessary in order to compensate for variations in the required threshold setting from character to character, and to correct for drift which may occur in the threshold circuits.

On command, the programme scans the central region of the scanner's field of view from the top until it encounters the first character. It then traces round the edge of the character, in the way described below, and in so doing establishes the size of the character with which it has been presented. This size determines the area used in the threshold setting procedure and also the "coarseness" of the scanning process. The initial drum position is also noted. It can be determined empirically that the number of black points detected by a digital scanner varies for different threshold levels in the manner indicated in *fig. 2*. It has been experimentally determined that the optimum thresholding level, as determined by a human operator, is that level which corresponds to the point of inflexion on this graph.

At this point it is necessary to manually set up the maximum and minimum hardware threshold levels to correspond to the computer specification of these levels. This is done by causing the programme to scan at each of these two threshold levels whilst adjustments are made to the controls to produce a display nearly all white for minimum threshold and nearly all black for maximum threshold.

The automatic threshold setting procedure is then entered and the computer programme selects the threshold level to use by determining the point of inflexion on the graph, having executed a number of consecutive raster scans over the character area in question, one scan with each available threshold level, and having determined the number of black points corresponding to each.

Having completed the setting-up procedures just described, the operation subsequently is completely automatic. The programme scans down from the top, as before, encounters the first character again, and, since the threshold, and consequently the character size, may have changed, again determines the character size and position. It next checks that the thresholding level is correct and if so enters the preprocessing, which is described later. If the threshold needs changing this is done before entering the preprocessing. The preprocessing is only entered if the position of the character is such that there is a good chance that the whole character lies within the field of view, having regard to the size of the character being investigated which was determined at the start.

When the preprocessing is completed the search scan starts again just below the last character, and searches an area twice as wide as the last character and centred symmetrically below it. If another character is found

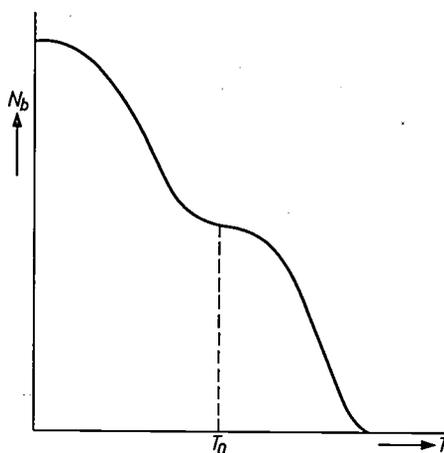


Fig. 2. The optimum threshold setting T_0 is at the point of inflexion in the graph obtained by plotting the number of black points N_b against threshold setting T .

the process above is repeated but if either the character is found too close to the bottom of the field of view, or the bottom of the field of view is so closely approached that there is insufficient room for a character to exist, then the paper carrying drum is incremented sufficiently for the point which the search scan reached to be moved to the top of the scanner's field of view. The search scan then continues from the corresponding point. Account is kept of the increments made by the drum, and the process continues automatically until one complete rotation of the drum has been made when a message to that effect is displayed and the process may be re-entered with new characters if desired.

The preprocessor

The input data to a character recognition process, that is the character printed on paper, contains a great deal of information which is not essential for the correct identification of the character. Such information relates, for example, to the position and orientation of the character, to noise present within the character field, and to irregularities of the character edge.

The purpose of the preprocessor in a character recognition system is to reduce the quantity of input information to an amount which can be accepted by the classifier. It should do this preferably by selecting from the input characters some small number of characteristics which are invariant for limited distortions of the characters, but which are adequate to form separable descriptions of the members of the character set. The research on this subject at Mullard Research Laboratories has been concentrated on forming a description of the character in terms of the direction of travel of a spot following the character edge. The actual direction of movement of the spot may be smoothed to reduce the effect of small edge irregularities. In addition the

character undergoes an averaging process before the edge-following is attempted. This also reduces small edge irregularities.

The averaging process operates as follows. A block of nine points, as shown in *fig. 3a*, is superimposed on each point of the matrix in turn. If there are five or more points present in this block the centre point is written into a new character matrix, irrespective of whether it was originally present. If there are less than five points in the block the centre point is removed in the new character matrix. This process may be repeated several times. The edge tracing process then operates on the new character matrix.

The character is detected and the tracing element, which may be a single point, or more suitably a block of nine points, is positioned on the character edge as has been described. The process then enters the edge-following mode. In this the tracing element may move (or jump) from the point at which it is situated to any one of the eight surrounding points. The jump is thus in one of the directions shown in *fig. 3b*. In this figure directions 0, 2, 4 and 6 correspond respectively to directions $+X$, $+Y$, $-X$ and $-Y$ of the matrix.

The tracing programme utilizes two directions which are called the "permanent" direction and the "present" direction. The permanent direction is the direction of the last successful movement of the tracing element, and the present direction is that direction in which the element is trying to move at the present time. If a successful jump is made, the direction in which it was made is taken as the new permanent direction. The computer programme then calculates the first choice direction for the next jump. This is the direction obtained by rotating 90° anticlockwise from the permanent direction. If the tracing element can move in this direction the sequence which has been described is repeated, and the direction is stored in a list in readiness for the subsequent recognition processes. If, however, the point to which the tracing element is trying to move is not permissible (it is outside the character, say) the first choice from the present direction is selected. This is the direction obtained by rotating 45° clockwise from the present direction. The programme then tries to move the tracing element in this direction. If repeated failures occur the programme selects the directions in a clockwise sequence from the present direction until a successful jump is achieved.

This process effectively holds the tracing element against the character edge, and thus causes it to outline this edge.

Tracing is continued until the whole character edge has been delineated (*fig. 1b*). The list of directions which have been assembled during the tracing are used as the input to the next process, that is, to the classifier.

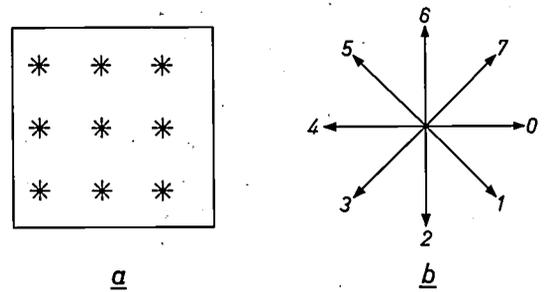


Fig. 3. a) The block of nine points which is scanned over the character matrix in the averaging process. *b)* The eight directions in which the edge tracing element may move from the point at which it is situated to one of the eight surrounding points.

The output list of jump directions is independent of the position of the character within the field of view. It is also insensitive to noise superimposed on the field of view as the largest character within the field is selected.

The classifier

Two types of classifier have been simulated. The first of these, which is intended to recognize a full alphanumeric set of characters (numerals and alphabetic characters), uses the principle of feature extraction. The second, which is at present a numeral only recognition system, uses an adaptive classification system.

1) Feature extraction

Classification of characters using feature extraction uses the approach of analysing the character edge direction to extract a number of selected features from the character.

The features which have been chosen are ends of lines, angles, curves and loops.

The average direction of the character edge at any point upon it is calculated by taking an average of the jump directions occurring at the points surrounding it. Such an average direction is calculated for each point on the character edge. The resultant average direction list, the trend list, is used in the feature extraction processes. Line ends are detected by finding a reversal of the edge directions in the trend list within a specified number of jumps, sufficient to enclose the end. As the tracing element follows the character edge in a clockwise direction, the reversal must also occur in a clockwise sense.

Angles are detected as sharp changes in the character edge direction, and curves are detected as changes in direction which are appreciable but which do not occur sharply enough for them to constitute angles.

Loops are detected by the presence of a closed inner character edge within a closed outer edge.

These features are then used to compile a codeword representing the unknown character. This is compared with a dictionary of similar codewords representing all the members of the character set which can be recognized. The dictionary word which exactly matches the code of the unknown is taken as the identification of the unknown. A facility is also included whereby together with the graphical information presented to the system, an indication may be given to the computer as to the identity of the unknown character. Thus the character has an identifying "label" attached to it. If such a labelled character is entered into the system and no exact match is found between its codeword and those in the dictionary, the new codeword will be entered into the dictionary as a standard character with the name marked on its label. It is thus possible to set up the process with a completely empty dictionary and to compile the dictionary entries by entering labelled characters into the process. This system is still under development and results are not yet available.

tern fed into the preprocessor. Each element of this new pattern is a simple number. The pattern of n elements is presented to a weighting network of $n + 1$ adjustable weights known as a threshold logic unit or T.L.U. (see fig. 4). Each element of the pattern is weighted by the appropriate parameter and the weighted elements are summed, i.e. the network gives an output $X_1w_1 + X_2w_2 + \dots + X_nw_n + w_0$ (the extra weight w_0 having a fixed input of $+1$). This output is compared with a threshold to give a final output of ± 1 .

Provided a suitable set of weights can be found, the T.L.U. can separate a set of patterns into two groups, A and B, such that patterns in class A give a $+1$ output and patterns in class B give a -1 output. The T.L.U. can be trained to find the correct weights, using a small number of patterns from each group as training patterns. This is achieved by giving each training pattern a desired output (fig. 4) which is $+1$ for all patterns in class A and -1 for all patterns in class B. Hence when a training pattern is presented to the system there is

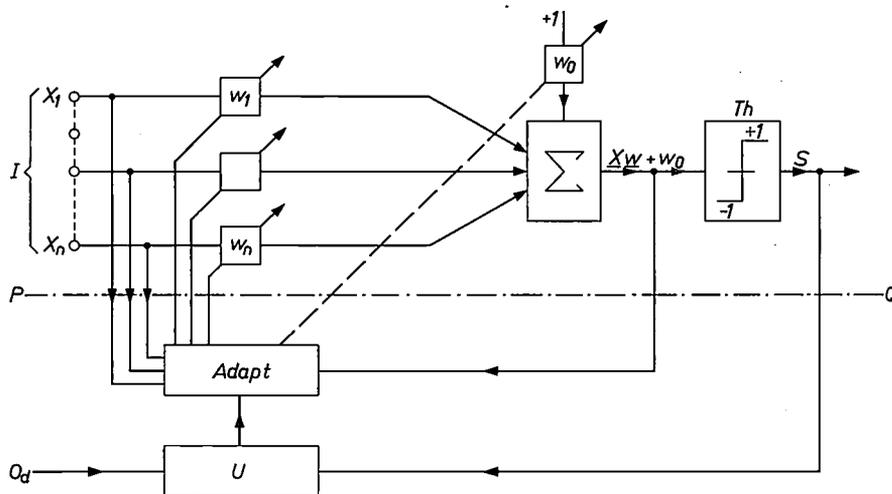


Fig. 4. The threshold logic unit (above the line P-Q) calculates the sign of $X_1w_1 + X_2w_2 + \dots + X_nw_n + w_0$. The associated training control unit below line P-Q adjusts the weights w_0, w_1, \dots, w_n until the correct separation is obtained. I input pattern, Σ summer, Th threshold, S sign of $X.w + w_0$, $Adapt$ adaptation control unit, O_d desired output, U teacher.

2) Adaptive classifier

Adaptive systems may be defined as systems that are able to use past experience in order to achieve and maintain optimal performance. This implies some kind of memory to store past experience, and the improvement in performance can be regarded as a type of learning.

Adaptive techniques can be used to design a classification system. The memory function is normally provided by a set of stored adjustable parameters or "weights".

The input to the classifier (output from the preprocessor) consists of a new pattern whose elements are derived from measurements made on the original pat-

tern both an actual output and a desired output. If the two have opposite signs the teacher tells the adaptation control to change the weights according to a specified formula.

Training patterns are presented to the unit in sequence until they are all correctly classified. This occurs when the sign of the actual output is positive for all patterns in class A and negative for all patterns in class B. Thus a single T.L.U. can be used to separate a set of patterns into two groups. To solve a multi-class problem a combination of T.L.U.'s must be used.

To classify the ten numerals 0-9 into their ten classes a minimum of four T.L.U.'s is needed. These would have to perform separations of the type 01234/56789.

There are 126 separations of this type and more than 10 000 possibilities of selecting a compatible set of four. Thus while using four T.L.U.'s has the advantage that it uses the minimum number, it has the disadvantage that it is not known in advance which separation is best to assign to each T.L.U.

If more than four T.L.U.'s are used, the number of structures, or arrangements of T.L.U.'s, which will separate 0-9 becomes very large. Whichever structure is used however, every pair of classes must be separated at some stage. For instance, a separation such as 0/123456789 is effectively composed of the "class-pair" separations 0/1, 0/2, 0/3, . . . 0/9.

For this reason, and for simplicity of training, a system using forty-five T.L.U.'s, each performing separations of the class-pair type is used (*fig. 5*). Thus for a character to be recognized as a 3 the separations 0/3, 1/3, 2/3 must give a negative output whilst 3/4, 3/5, etc. must give a positive output.

In the numeral recognition system developed at Mullard Research Laboratories, the input patterns to the T.L.U. each contain 25 elements. These are derived from the list of edge directions by an averaging process which reduces the number of jumps in the list (which is variable for variable size characters) to 25 average jumps.

The training data consisted of 400 numerals hand-printed by eight people. A test set of another 400 numerals drawn by the same people was also produced, together with 150 numerals drawn by three other people. The results obtained when the recognition system was tested using these samples are summarized in *Table I*.

Future work

These results are encouraging. If a machine with such a performance were to be used in a commercial environment it would be necessary to use a checking system to detect the reading errors. This could be done, for

Table I. Character recognition results for an edge-following system using forty-five adaptively trained T.L.U.'s. The table shows the results for 400 hand-printed training characters and 400 + 150 test characters.

Data	Number of characters	Mis-classified	Re-jected	Correctly identified
Training set (examples of numerals 0-9 incl. drawn by eight people)	400	—	—	100%
Test set (numerals drawn by the same eight people)	400	1.25%	2.75%	96%
Test set (numerals drawn by three other people)	150	7.3%	8%	84.7%

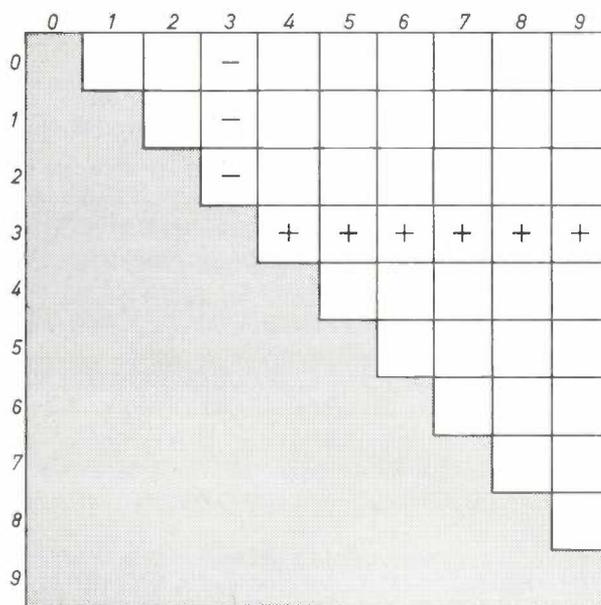


Fig. 5. Decoding for class-pair separations. Each square represents the output of one of the forty-five T.L.U.'s. For a character to be recognized as a 3 the separations 0/3, 1/3, 2/3 must give a negative output whilst 3/4, 3/5, etc. must give a positive output.

example, by incorporating a check digit, or several digits, in the numbers to be read. Such techniques are well established however.

Although the above research has been performed by a computer simulation, work is progressing on a practical embodiment of the system. This uses a commercial paper handling equipment, some special edge-following hardware, which has been constructed, and a small general purpose computer programmed to perform the recognition.

Future simulation work is aimed at the extension of the system to include the recognition of alphabetic characters, and at the completion of the feature extraction recognition system.

The authors would like to thank the United Kingdom Ministry of Technology for permission to publish this research which was supported under a normal cost sharing contract under the Ministry's Advanced Computer Techniques Project.

Summary. Two optical character recognition systems have been simulated on a digital computer. In both systems the character is observed by an on-line flying spot scanner which feeds the character into the computer in the form of a binary matrix. The automatic search for the character and setting of the threshold for the flying spot scanner are described. The direction of travel of a spot following the outside edge of the character is then obtained and used as a new description of the character. This description is used in both recognition systems, the first of which obtains predetermined features of the edge such as ends and angles. The second system uses a condensed description of the edge in an adaptively trained classifier. Both systems may be monitored by an on-line cathode ray tube display.

Optical character recognition equipment



Some of the special purpose peripheral equipment which has been constructed and connected to the computer for simulating the various processes used in the recognition of characters written, or printed, on paper. (See the article by Saraga, Weaver and Woollons in this issue, p. 197.) On the left of the photograph can be seen a digital flying spot cathode ray tube scanner which can be programmed to scan the input document in many different ways, whilst on the right is a display tube which can show the output from the scanner, as in the photograph, or indeed, other results produced by the computer. Some aspects of optical character recognition which have been investigated using this equipment are various methods of locating the character,

and of selecting the optimum threshold level to be used when scanning the character for the recognition process proper. Control of the course of the computer programme can be effected by operation of the set of key switches which can be seen under the operator's hand [*].

[*] Although this equipment was originally developed for use in the research programme on optical character recognition, many other users are now finding it of great value. The display and light-pen have been used in the design of filters for communication equipment. Some electron tube design has also been helped by the facility of being able to obtain a visual representation of electron trajectories much more rapidly than by any other means.

Low noise parametric amplifiers

C. S. Aitchison

Although a tentative approach to the subject of negative resistance amplifiers was made more than a century ago, the matter lay dormant until the late 1950's. By then, semiconductor *P-N* junctions and microwave ferrite materials had become available, and these made possible the construction of practical negative resistance amplifiers for microwave frequencies [1].

System sensitivity

All low noise amplifiers are used in receiving systems of one sort or another and it is convenient to consider the simple receiving system shown in *fig. 1*, which

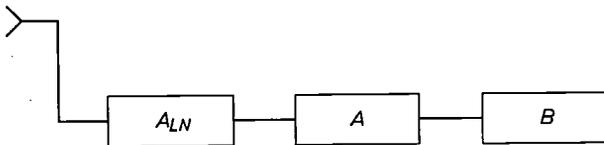


Fig. 1. A typical low noise system consisting of an aerial followed by a low noise amplifier A_{LN} of large gain and a further amplifier A of sufficient gain to operate a display device B .

consists of an aerial followed by a low noise amplifier of large gain and a further amplifier of sufficient gain to operate the display device. In such a system the noise in the display emanates from both the low noise amplifier and from noise sources in the aerial beam and side lobes.

This aerial noise can be specified in terms of an equivalent noise temperature T_{ant} which is the physical temperature of a resistor of the same impedance as the aerial which could replace the aerial in the system without affecting the system noise performance.

Fig. 2 is a graph of aerial temperature (neglecting side lobes and direct noise sources) as a function of frequency and shows a galactic noise contribution falling rapidly with increasing frequency as well as water vapour and oxygen absorption contributions. A broad minimum is seen to occur in the frequency region 2-8 GHz with a minimum noise temperature

of only a few °K. Discrete sources of noise such as the sun and radio stars are ignored in this graph and the exact sky temperature is in any case a function of the area of the sky under investigation — being higher in the galactic plane than perpendicular to it.

The low noise amplifier is likewise specified in terms of an equivalent noise temperature, T_{amp} . The signal-to-noise ratio S/N for a system is dependent on these temperatures in the following way:

$$\frac{S}{N} \propto \frac{1}{T_{ant} + T_{amp}}$$

Thus in any system where we require to maximize the signal-to-noise ratio it is desirable for T_{amp} to be very much less than T_{ant} . A microwave low noise amplifier should ideally have a noise temperature of less than 10 °K if it is operated in conjunction with an aerial of low noise temperature.

It is usual to measure the noise properties of a low noise amplifier in terms of its noise figure F , since this is defined in terms of a room temperature source. The noise temperature T_{amp} of the amplifier is related to F through the expression:

$$T_{amp} = (F - 1)290$$

and expressed in degrees Kelvin.

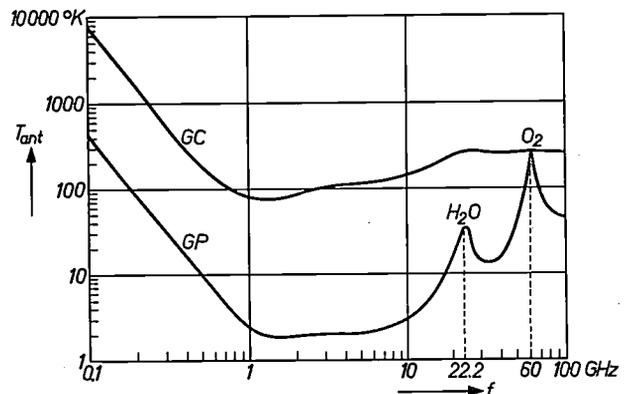


Fig. 2. A graph of aerial noise temperature T_{ant} as a function of frequency f of a perfect aerial where *GC* refers to the galactic centre, *GP* the galactic pole. The graph also shows the contributions made by water vapour and oxygen absorption.

The parametric amplifier

Introduction

A typical parametric amplifier circuit used for theoretical analysis is shown in *fig. 3*. A variable capacitance C with associated loss R_d is connected to two reactive circuits; the first consists of an inductance L_s and a circulator of impedance R_g where R_g is the signal source impedance. The circulator feeds the signal from the signal generator into the amplifier and the amplified signal returning from the amplifier into the load R_l . This first circuit is resonant at the signal frequency. A second inductance L_l resonates a second circuit at a different frequency called the idling frequency. Band-pass filters F_s and F_l are provided so that signal and idling frequency currents are confined to the appropriate circuit.

It can be shown mathematically that if we vary the capacitance C at the sum of the signal and idling frequencies, energy is transferred from the capacitance varying source to the signal and idling circuits and the current flowing in both these circuits is increased as if the loss in the circuits had been reduced by the addition of a series negative resistance. This process is known as pumping, and the energy transfer from the pump source to the signal and idling circuits occurs by virtue of the periodically varying nature of the capacitor, with consequent production of sideband energy resulting from the mixing of the signal and pump frequencies.

The current in the signal circuit is given by:

$$I = \frac{E}{(R_g + R_d)(1 - A)},$$

where A is the ratio of negative resistance introduced in the signal circuit to the total positive resistance in the signal circuit and E is the e.m.f. of the signal frequency generator.

A is given in terms of the circuit parameters by:

$$A = \frac{\gamma^2}{4\pi^2 C^2 f_s f_l (R_g + R_d)(R_l + R_d)},$$

where γ defines the amount of capacity variation due to pumping and is given by:

$$\gamma = \frac{C_{\max} - C_{\min}}{2(C_{\max} + C_{\min})}.$$

R_l is the loss in the idling circuit other than that associated with the capacitance C .

Thus it can be seen that the equivalent circuit at the signal frequency of this system is a series resonant cir-

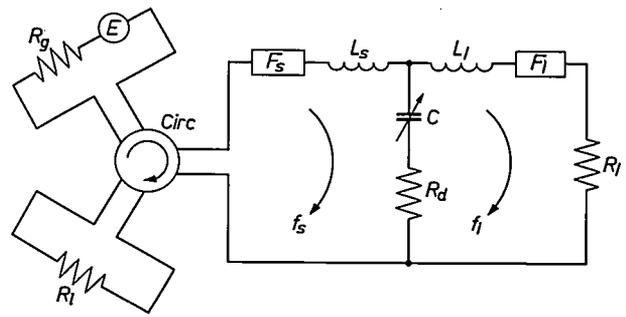


Fig. 3. Parametric amplifier circuit used for theoretical analysis. A variable capacitance C with its associated loss R_d is connected to two reactive circuits. The first consisting of inductance L_s and signal source impedance R_g is resonant at the signal frequency f_s . The second with inductance L_l is resonant at the idling frequency f_l . R_l represents the loss other than R_d . Band-pass filters F_s and F_l confine signal and idling frequency currents to the appropriate circuit. The circulator *Circ* feeds the signal from the signal generator E into the amplifier and the amplified signal returns from the amplifier into load R_l .

cuit of C and L_s fed from a source impedance R_g , with internal loss R_d and an additional series component — a negative resistance of magnitude $A(R_g + R_d)$.

Gain of the parametric amplifier

By extending the definition of reflection coefficient ρ to include negative terminating resistances the gain of this system of negative resistance amplifier and circulator can be written as:

$$G = |\rho|^2 = \left(\frac{R_g + |R_-|}{R_g - |R_-|} \right)^2,$$

where $|R_-|$ is the magnitude of the negative resistance terminating the line of characteristic impedance R_g .

Thus for a given required gain and known R_g the minimum R_d can be calculated assuming that γ has its maximum value.

Noise factor and bandwidth of the parametric amplifier

Having established expressions for the currents in terms of the generator e.m.f. and the capacitance variation, it is possible to calculate the noise figure for the amplifier using the standard noise figure definition. There are three sources of thermal noise in the circuit, viz, the source resistance R_g , the capacitor loss R_d and the loss in the idling circuit R_l . By assuming an ideal circulator (no losses) and noiseless pumping, the noise figure can be calculated as:

$$F = 1 + \frac{R_d}{R_g} \frac{T}{T_0} + A \frac{f_s}{f_l} \left(\frac{R_d + R_g}{R_g} \right) \frac{T}{T_0},$$

[1] B. Bollée and G. de Vries, Experiments in the field of parametric amplification, Philips tech. Rev. 21, 47-51, 1959/60.

where it has been assumed that R_d and R_i are at the same temperature T and T_0 is the noise figure definition standard temperature. This can be conveniently expressed in terms of amplifier noise temperature as:

$$T_{\text{amp}} = T \left\{ \frac{1}{p} + A \frac{f_s}{f_i} \left(1 + \frac{1}{p} \right) \right\},$$

where we have written the overcoupling ratio R_g/R_d as p . If we arrange that p is large, then, for high gain ($A \approx 1$) this expression can be approximated to:

$$T_{\text{amp}} = \frac{f_s}{f_i} T.$$

Thus the amplifier noise temperature can be made much less than its physical temperature by arranging for f_i to be much greater than f_s . This is the essential advantage of parametric amplification.

It is convenient to express the overcoupling ratio in terms of γ , C and R_d (the properties of the pumped capacitor) and the expression is (for $A \approx 1$):

$$\gamma f_c = \sqrt{f_s f_i (1 + p)},$$

where we have written $f_c = 1/2\pi C R_d$; f_c is known as the cut-off frequency.

Examination of the expression for noise figure suggests that, since R_g and f_i can be changed while keeping $(R_g + R_d)f_i$ constant so that the gain is unaltered, there is an optimum idling frequency corresponding to a minimum noise figure. This minimum noise figure is given by $F_{\text{min}} = 1 + (2f_s/\gamma f_c)(T/T_0)$ and occurs at the optimum pump frequency, which is γf_c (provided $\gamma f_c \gg f_s$) or in noise temperature terms:

$$T_{\text{amp min}} = \frac{2f_s T}{\gamma f_c},$$

which expresses the minimum amplifier noise temperature in terms of the γf_c value of the capacitor, the signal frequency and the physical temperature. Thus the value of γf_c is a figure of merit for the capacitor.

The bandwidth of a parametric amplifier operated in conjunction with a circulator can be expressed in terms of the unpumped signal bandwidth B_s and idling bandwidth B_i . The gain-bandwidth expression is:

$$G^{1/2} B = 2 \left(\frac{1}{B_s} + \frac{1}{B_i} \right)^{-1}.$$

This expression relates the operating gain and bandwidth to the passive parameters of the signal and idling circuits.

The varactor diode

It is known that a junction formed of P -type and N -type semiconductor material is analogous to a parallel plate capacitor since diffusion of the holes in the P -type material across the junction material and of the electrons in the N -type material results in a non-zero net charge density on either side of the junction. The potential gradient thereby set up restricts the region from which the carriers are removed and limits the so-called "depletion layer". The thickness of this depletion layer can be changed by means of applied bias and thus the capacity across the depletion region can be altered.

The capacity law is given by:

$$C = \frac{C_0}{\left(1 - \frac{V}{\Phi} \right)^n},$$

where Φ is the contact potential, V the applied bias and n the exponent which is 0.5 for abrupt junctions and 0.33 for graded junctions.

Fig. 4 shows typical capacity and current variation with bias for a P - N junction demonstrating that the

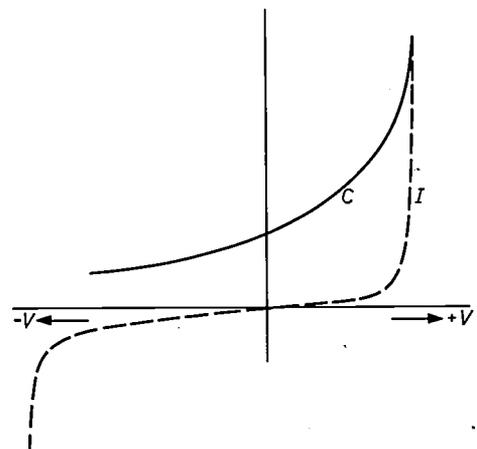


Fig. 4. Typical capacity C and current I variation with bias V for a P - N junction showing that the capacity variation is limited by forward current in the forward direction and by breakdown in the reverse direction.

capacity variation is limited by forward current in the forward direction and by breakdown in the reverse direction. The effect of this current is to decrease the cut-off frequency of the varactor as well as to provide shot noise thereby degrading the noise performance of an amplifier using such an over-driven diode. Typically 1 V reverse bias and 1 μ A forward current are acceptable limits.

lates a lumped inductance at the signal frequency and resonates with the varactor which is placed across the coaxial line. The idling current is supported in the parallel resonance of the diode and this resonance is reduced to the range of 6-8 GHz by the probe which extends into the waveguide to feed pump power to the diode. The wave-guide is cut off at the idling frequency and is matched at the pump frequency by means of a suitably placed short-circuit and screw system. A photograph of a number of these amplifiers is shown in *fig. 8*. *Table II* summarizes the results obtained with this design of amplifier at frequencies between 400 MHz and 4.6 GHz.

A design of amplifier, known as the type 3, which operates in the range 2.4-12 GHz with a pump in the range of 26-40 GHz, is shown in *fig. 9* [3]. Again, a double quarter wave transformer transforms from 50 Ω to a convenient impedance for the diode. The second half of the quarter wave transformer contains a low-pass filter which prevents propagation of the idling

and the signal inductance is a quarter wavelength long at the idling frequency. Using the CXY 10 micropill diode, the series resonance of the diode which occurs

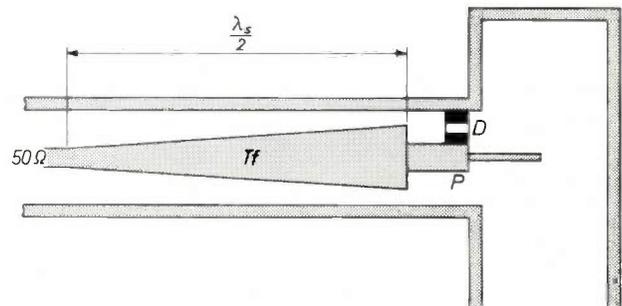


Fig. 7. Schematic diagram of type 2 parametric amplifier. This early Mullard design is capable of operation at signal frequencies in the range 400-3000 MHz using the CAY 10 diode. An impedance transformer *Tf* consisting of a linearly tapered transition of length $\lambda_s/2$ transforms from 50 Ω in coaxial line to an impedance appropriate to the diode *D* at the signal frequency. A short length of high impedance coaxial line *P* simulates a lumped inductance at the signal frequency and resonates with the varactor.

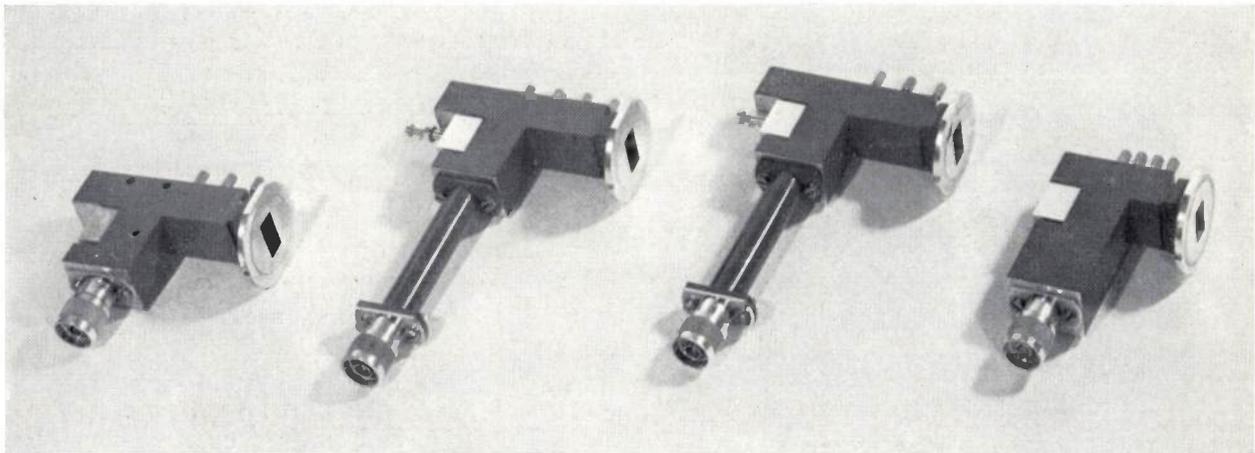


Fig. 8. A range of type 2 amplifiers covering the frequency band 400-3500 MHz. The type N connectors indicate the size of the amplifiers.

and pump frequency in the coaxial line. The diode is placed across the waveguide and a lumped inductance is formed from a short length of short-circuited coaxial line in series with the diode. Using the CAY 10 pill diode, the idling current at approximately 30 GHz is supported by the parallel resonance of the diode

Table II. Summary of the performance details of a type 2 amplifier.

Signal frequency MHz	Pump frequency GHz	Noise factor dB	Bandwidth MHz	Pump power mW	Tuning range MHz
400	7000	1.6	12	1	—
750	7500	1.9	15	1	± 50
1400	8000	2.0	20	1	± 100
3000	9500	3.0	20	10	± 250
4600	12000	4.5	25	50	—

[3] C. S. Aitchison, R. Davies and P. J. Gibson, A simple diode parametric amplifier design for use at S, C and X-band, IEEE Trans. on microwave theory and techniques MTT-15, 22-31, 1967 (No. 1).

Normally silicon or gallium arsenide is used to form the varactor junction for use in room temperature amplifiers. Gallium arsenide has the advantage of a higher cut-off frequency which is often maintained down to liquid helium temperature; silicon has the advantage of a larger γ than gallium arsenide but the disadvantage of a lower cut-off frequency and operation down to liquid nitrogen temperature only.

The equivalent circuit of a varactor within a special microwave encapsulation is shown in *fig. 5*. It consists of the junction C_j in series with the series resistance R_{ser} ; across this series combination there is an internal stray capacity C_{int} . In series with this combination is the lead inductance L with the capacity of encapsulation C_{ext} in shunt with the resultant series circuit.

It is normally more convenient to combine R_{ser} , C_j and C_{int} into an equivalent C and R_d as shown in *fig. 5* and it can be seen that this equivalent circuit has both a series and parallel resonance.

The two microwave encapsulations which are currently used are known as the "pill" and the "micropill". Both are shown in *fig. 6*.

The Mullard gallium arsenide varactor CAY 10 is manufactured in the pill encapsulation and the Mullard epitaxial gallium arsenide varactor CXY 10 is made in the micropill encapsulation. *Table I* lists the microwave properties of these varactors [2].

Table I. Typical microwave properties of the Mullard gallium arsenide varactor CAY 10 in the pill encapsulation and the Mullard epitaxial gallium arsenide varactor CXY 10 in the micropill encapsulation.

		CAY 10	CXY 10
C at 0 volts bias	(pF)	0.4	0.25
R_d at 0 volts bias	(Ω)	4	2
Series resonant frequency	(MHz)	9	35
Lead inductance	(pH)	700	100
C_{ext}	(pF)	0.15	0.25
Parallel resonant frequency	(GHz)	30	50
γf_c	(GHz)	25	40
f_c	(GHz)	150	300

The design of parametric amplifiers

Since a distributed circuit used to resonate a varactor diode has a higher Q and therefore narrower bandwidth than the corresponding lumped circuit, it is desirable to use lumped circuit systems to form the signal and idler resonances so that the gain-bandwidth product of the amplifier is not needlessly reduced. It is therefore particularly convenient to make use of the internal resonances of the varactor diodes to support

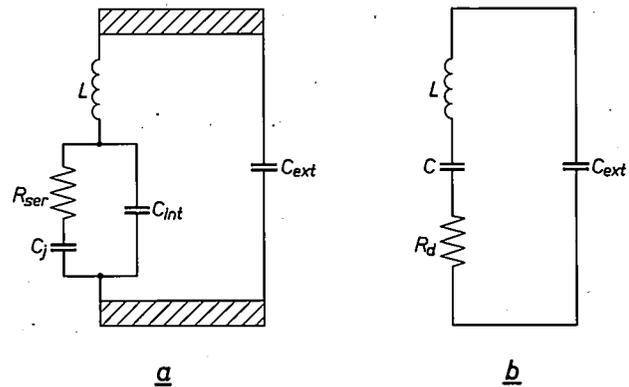


Fig. 5. a) The equivalent circuit of a varactor junction within a special microwave encapsulation. C_j junction capacity. R_{ser} series resistance. C_{int} internal stray capacity across this series combination. L lead inductance. C_{ext} capacity of encapsulation. It is convenient to combine R_{ser} , C_j and C_{int} into equivalent C and R_d as shown in (b).

the idling frequency and to use a lumped inductance to resonate the diode at the signal frequency.

An early Mullard amplifier design, known as the type 2 amplifier, which can operate at signal frequencies in the range 400-3000 MHz and is pumped in the frequency range of 7-12 GHz, uses the CAY 10 diode and is shown in *fig. 7*. A double quarter-wave transformer transforms from 50 Ω in coaxial line to a signal frequency impedance appropriate for the diode. A short length of high impedance coaxial line simu-

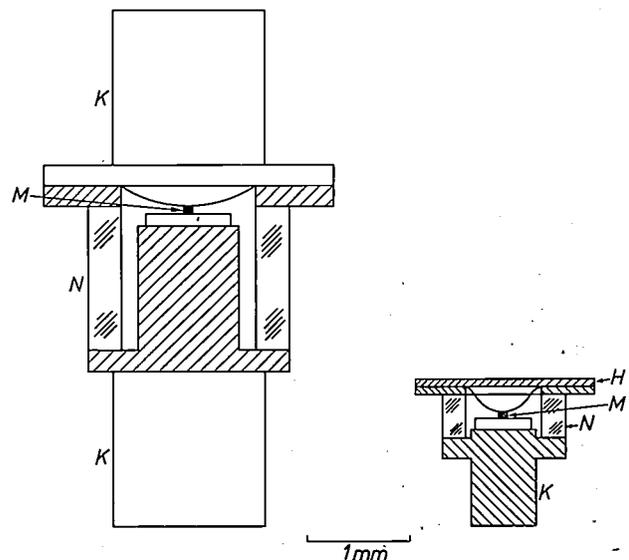


Fig. 6. The two microwave varactor encapsulations currently used are the "pill" (left) and the "micropill" (right). K kovar stud. M diffused mesa. N ceramic insulator. H top cap.

[2] These varactors were developed by Associated Semiconductor Manufacturers Ltd., Wembley, England — an associated company.

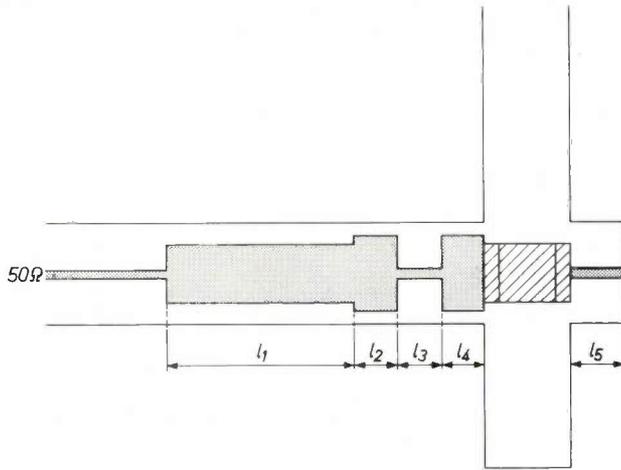


Fig. 9. Schematic diagram of type 3 amplifiers which operate in the range 2.4-12 GHz with a pump in the range 30-40 GHz. Lengths l_1 , l_2 , l_3 and l_4 form a double quarter-wave transformer at the signal frequency and l_2 , l_3 and l_4 are quarter-wave sections at the idling frequency. The diode is shaded. The section l_5 resonates the signal circuit and is a quarter-wavelength at the idling frequency.

at approximately 30 GHz supports the idling current and the coaxial line is a half wavelength at the idling frequency. Fig. 10 shows a photograph of three amplifiers of this design which operate in the frequency bands 3 GHz, 5.5 GHz and 9 GHz and Table III shows the performance of this type of amplifier.

Bandwidth optimization

Theory

Theoretical considerations suggest that for a parametric amplifier with simple signal and idling circuits the maximum operating bandwidth for a given gain will be achieved if the diode series resonance is used to support the idling resonance. It is then possible to calculate both the signal and idling unpumped bandwidths in terms of the overcoupling ratio and diode parameters. Differentiation shows that there is an optimum

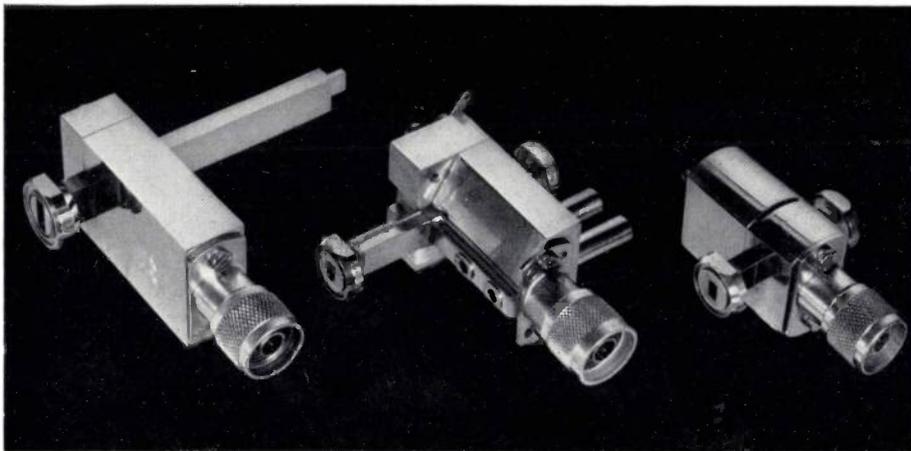


Fig. 10. Three type 3 amplifiers which operate in the frequency bands 3 GHz (right), 5.5 GHz (centre) and 9 GHz (left).

Table III. Summary of the performance details of type 3 amplifier.

Signal frequency GHz	Pump frequency GHz	Noise temperature °K	3 dB bandwidth at 20 dB gain MHz	Pump power mW	Tuning range MHz	Varactor type
3.0	33	95	25	20	150	CAY 10
5.6	39	140	50	50	250	Selected CAY 10
9.0	33	190	50	20	500	CXY 10

idling frequency for maximum root-gain-bandwidth product given by:

$$f_{i \text{ opt}} = \left(\frac{2\gamma^2 f_c^2 f_s}{1 + C_{\text{ext}}/C} \right)^{1/3}$$

and that the corresponding maximum root-gain-bandwidth product is:

$$G^{1/2} B_{\text{max}} = \frac{2}{3f_c} \left(\frac{2\gamma^2 f_c^2 f_s}{1 + C_{\text{ext}}/C} \right)^{2/3}$$

Thus the gain-frequency response is calculable in terms of the diode parameters and the signal frequency.

The gain-frequency response can be increased further by the provision of reactance compensation elements in either the signal or the idling circuit. Theoretical calculations show that whereas for the simple uncompensated amplifier the gain-bandwidth product is of the form $G^{1/2}B = \text{constant}$, with one degree of reactance compensation $G^{1/4}B = \text{constant}$; a further reactance compensation gives $G^{1/6}B = \text{constant}$.

Practical bandwidth optimization

The most convenient method of supporting the idling current in the series resonance of the varactor is to use two varactors each of which completes the other's idling resonance. This can be achieved by placing two micropill varactors half a wavelength apart in a microstrip line. If the line is of sufficiently low characteristic impedance the bandwidth of this system corresponds to that of one varactor. A parametric amplifier (type 5) using this system has been designed with the usual coaxial signal circuit and the idling structure placed within the waveguide. Fig. 11 shows a photograph of such an amplifier which at a signal frequency of 4 GHz gave a root-gain-bandwidth product of 1 GHz when pumped at 33 GHz.

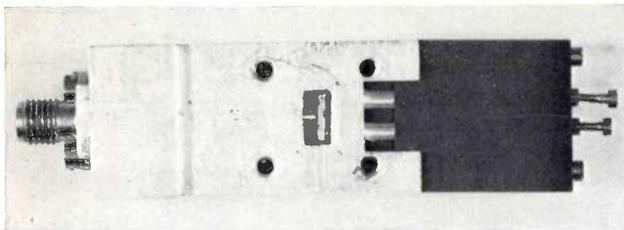


Fig. 11. Type 5 amplifier using two micropill diodes which at a signal frequency of 4 GHz gave a root-gain-bandwidth product of 1 GHz when pumped at 33 GHz.

The application of reactance compensation to this amplifier increases the bandwidth at 20 dB gain to 250 MHz. Two such amplifiers in cascade have produced the gain-frequency response shown in fig. 12, viz. 500 MHz bandwidth at 26 dB gain at a signal frequency of approximately 3.8 GHz.

Noise temperature improvement

As mentioned previously, the noise temperature for a parametric amplifier is proportional to its physical temperature. Provided the varactor figure of merit γf_c is maintained, a reduction of amplifier noise temperature should occur on cooling a parametric amplifier to temperatures such as 80 °K (liquid nitrogen)

or 4 °K (liquid helium). A type 5 amplifier with a signal frequency of 4 GHz has been cooled to a temperature of 4 °K at these Laboratories and has given a noise temperature of 3 °K.

The future

The combination of wideband techniques and physical cooling will give an amplifier system with a noise temperature of less than 10 °K and a bandwidth of 500 MHz at 26 dB with a signal frequency of 4 GHz. There is a programme at Mullard Research Laboratories to produce such an amplifier for the first stage of a receiver of a satellite communication system.

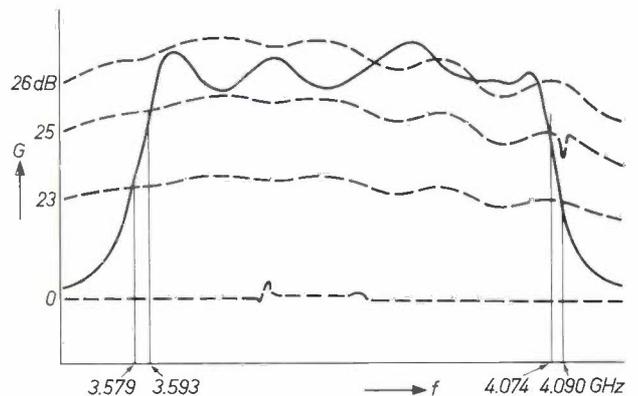


Fig. 12. A pen recording of the gain G versus frequency f response of two cascaded type 5 parametric amplifiers. The dotted lines are the 26, 25, 23 and 0 dB reference levels and the full line is the response of the cascaded amplifier system.

Physical cooling to low temperatures is, however, only a temporary requirement since the continued improvement in varactor figure of merit will eventually obviate the need to cool the amplifier.

At present parametric amplifiers are expensive to build and are therefore used only for professional applications. Work on a programme to produce evaporated lumped microwave components has already begun at these Laboratories and on its completion, cheap parametric amplifiers, pumped with cheap solid-state microwave sources, will be available and could result in their use in mass production applications.

Summary. Parametric amplifiers are very low noise devices which are employed in microwave receiving systems. The reason for their inclusion in such systems is indicated in the article and the theory of operation of varactor diode parametric amplifiers outlined. The critical diode parameters are indicated, and it is demonstrated that the amplifier can be optimized in terms of noise figure or bandwidth and that the noise performance can be improved by cooling the varactor diode. General design principles are given and the practical realization of a number of amplifiers covering the frequency range 400 MHz to 12 GHz described. Performance details are given for each type of amplifier.

A survey of coaxial and strip-line microwave components

S. J. Robinson and P. T. Saaler

Introduction

The universal demand for components having a wide bandwidth and small size in the microwave field has led to a move away from waveguide to coaxial and strip transmission line, and so although the work described is based on specialized wideband requirements, it has followed recent passive microwave component development. In particular, considerable recourse has been made to balanced circuits incorporating hybrid junctions and to the use of double-comb filters (usually called interdigital filters), and in these areas a number of innovations have been introduced. In surveying the work carried out over the last few years in these Laboratories no attempt is made to list the components studied but rather a few items which typify the general trends are described.

Due to close tolerance requirements and the existence of enclosed conductors, the constructional techniques play an important part in the design of transmission line components. These aspects will, therefore, be considered briefly, following the discussion of the electrical aspects.

At the end of World War II, waveguide was by far the most universally used form of transmission line for conveying signals at microwave frequencies. The frequency bands used were usually around 1 GHz, 3 GHz and 10 GHz and these are still popular frequencies. The chief application for microwaves in those days was radar, and a simple radar system worked in general at a spot frequency. There was therefore little requirement for a transmission line which would propagate a wide frequency range. In fact, rectangular waveguide is capable of working over a fractional band of approximately 40% of the centre frequency without higher modes being set up and waveguides are still used extensively in systems requiring a fairly wide band, e.g. communication links. Nevertheless, waveguides have some severe disadvantages. The working bandwidth is limited by the phenomenon of cut-off at the lower frequency and overmoding (i.e. the setting up of higher modes) at the upper. Waveguide is large and heavy, especially at low microwave frequencies; a guide capable of propagating at 1 GHz (wavelength 30 cm) has to be approximately 20 cm by 10 cm. For these reasons alternative forms of transmission line were investigated and three basic types became popular; these were coaxial line, strip-

line (or slab-line) and microstrip. Strip-line consists of two flat ground planes with a rectangular section inner conductor (fig. 1a), in slab-line the inner conductor has a circular cross-section (fig. 1b), and microstrip consists of one ground plane and an exposed "hot" line (fig. 1c). These lines, when operated in their dominant mode, have no cut-off frequency although the upper working frequency is limited by overmoding. They are capable of working over much greater bandwidths than waveguides and are considerably smaller and lighter. The disadvantages are their high attenuation compared with that of a waveguide and the difficulty of making

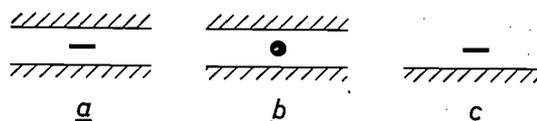


Fig. 1. The three types of transmission line used for making microwave components.

- Strip-line consists of two flat ground planes with a rectangular section inner conductor.
- Slab-line — the inner conductor has a circular cross-section.
- Microstrip consists of one ground plane and an exposed "hot" line.

connections between one component and the next. In addition, microstrip has a drawback which is not present in coaxial or strip-line. Since it is not entirely enclosed, microstrip tends to radiate unless the line is supported on a material having a high dielectric constant. This causes power loss and interference with adjacent circuits. For this reason microstrip has not been studied extensively at the Mullard Research Laboratories and no further mention will be made of it.

Coaxial line makes an excellent transmission medium for frequencies at least up to 10 GHz, although flexible cable has a high loss at this frequency. Air-dielectric coaxial line has been extensively used for building microwave components, in particular the hybrid ring and devices based upon it, e.g. mixers and phase-discriminators. Work has also been done on microwave filters which use coaxial line; mainly on low-pass, band-pass and band-rejection types.

Strip transmission line did not become popular until Bates^[1] in 1956 resolved the difficulty of calculating

S. J. Robinson, M.A., and P. T. Saaler, Ph.D., B.Sc.(Eng.), are with Mullard Research Laboratories, Redhill, Surrey, England.

[1] R. H. T. Bates, The characteristic impedance of the shielded slab line, IRE Trans. on microwave theory and techniques MTT-4, 28-33, 1956.

the characteristic impedance of the line from its dimensions. Strip-line in itself has few advantages over coaxial line but it is much more convenient for making a class of microwave components which depends on electromagnetic coupling for its operation. These devices require the close proximity of two or more transmission-line inner conductors so that the electromagnetic field set up on the input line couples across to the nearby conductors. Using this technique a variety of directional couplers (hybrids) and microwave filters can be designed. Coaxial line, being totally enclosed, does not lend itself to this arrangement whereas a strip-line configuration allows a number of inner conductors to be placed close together between a pair of common ground planes. Strip-line has one further important advantage over waveguide or coaxial line. This lies in the facility to print the inner conductor as a thin copper strip on a sheet of solid dielectric. A second sheet of the same thickness is added, both the sheets being clad with copper on their outer faces to form the ground planes. This technique, which is often simpler to apply than its "thick-line", air-dielectric, counterpart, is particularly useful in the production of complicated structures such as filters and phase-discriminators, especially where a large number of the same device is required. In the past few years a number of microwave components using the above techniques have been built at Mullard Research Laboratories and a few typical examples will be described in more detail.

Hybrid junctions and couplers

A simple hybrid junction is essentially a four port device in which the power input at one port splits (usually equally) between two ports, the fourth port being isolated from the input. The addition of a reversed-polarity pair of microwave diodes to the isolated ports of such a device forms a balanced mixer.

The simple hybrid junction in coaxial line or in waveguide consists of a closed ring with four ports spaced at quarter-wavelength and three-quarter-wavelength intervals. A hybrid with a broader bandwidth is obtained if a phase reversal section is inserted between two ports, the four ports being then symmetrically placed around the ring (*fig. 2*). A signal input at port P_1 divides equally between ports P_3 and P_4 , the outputs being in antiphase. This device has been used as the basis of a range of microwave mixers marketed by the M.E.L. Equipment Company Ltd. A mixer in the band 2.5-4.1 GHz is shown in *fig. 2*.

A more recent mixer design employs the electromagnetic directional coupler, a solid-line version of which is shown in *fig. 3*. This type of coupler has the property that a signal at port P_1 divides equally (for 3 dB coupling) between ports P_3 and P_4 , port P_2 being

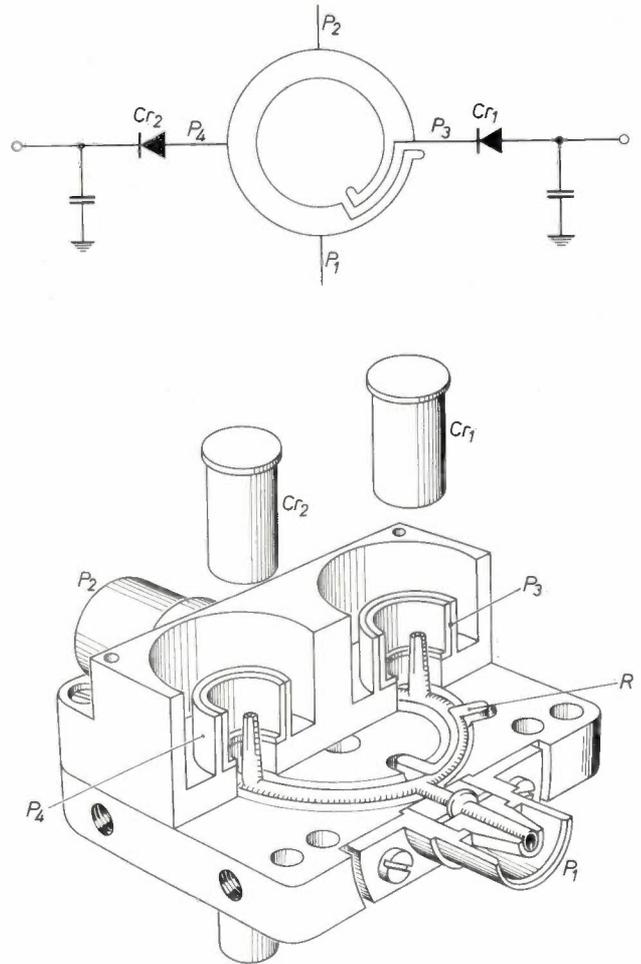


Fig. 2. Coaxial phase reversal ring balanced mixer. Band 2.5-4.1 GHz, isolation 15 dB. The signal input at port P_1 divides equally between ports P_3 and P_4 , the outputs being in antiphase. Cr_1 and Cr_2 are crystal diodes. R phase reversal hybrid ring.

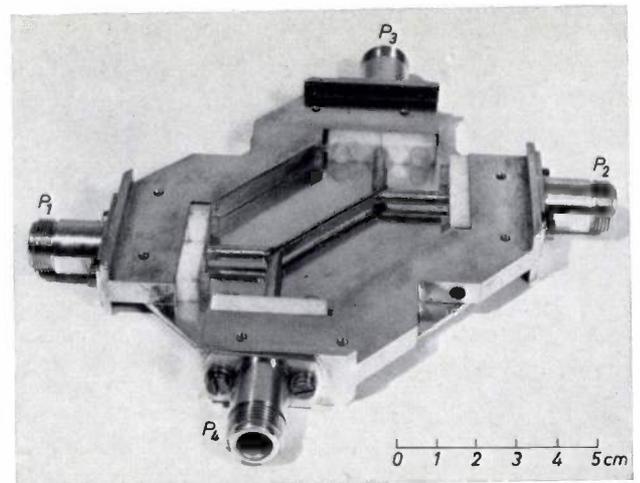


Fig. 3. Solid-line ninety-degree coupler (1.0-2.6 GHz). This coupler divides a signal at P_1 equally (for 3 dB coupling) between P_3 and P_4 , P_2 being theoretically isolated. The unwanted signal in P_2 is typically 25 dB down on the input signal. The output signals at P_3 and P_4 are in phase quadrature at all frequencies — hence ninety-degree coupler.

theoretically isolated, although the unwanted signal in port P_2 is typically 25 dB down on the input signal. In respect of isolation or directivity, coaxial and strip-line hybrids are inferior to their waveguide counterparts. The output signals at ports P_3 and P_4 (fig. 3) are in phase quadrature at all frequencies and this device is therefore known as a ninety-degree coupler. It has a broader working bandwidth than the phase reversal ring and can easily cover a 3 : 1 or 4 : 1 frequency range. The application of one microwave diode to such a coupler yields a single-ended mixer. Such a device employing a printed coupler for use in the band 7-11.5 GHz is shown in fig. 4. A balanced mixer is obtained by fitting a 3 dB coupler with two diodes.

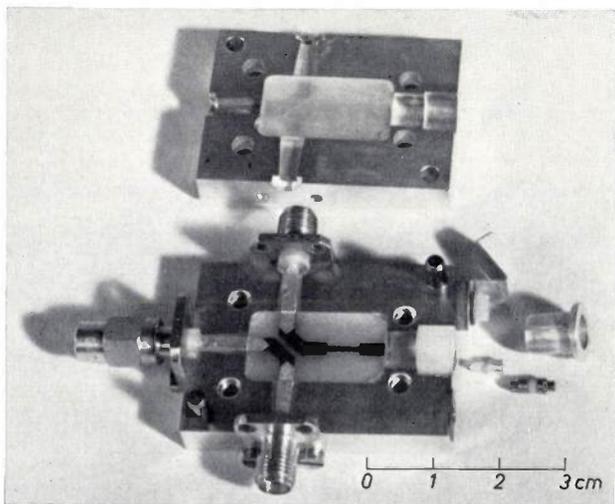


Fig. 4. Printed single-ended mixer within the band 7-11.5 GHz. At 9.6 GHz the local oscillator signal isolation is 20 dB and the voltage standing wave ratio is 1.3.

A variation of the solid-line ninety-degree coupler is shown in fig. 5. The two inner conductors pass through a metallic block forming a re-entrant section. This configuration was devised by Cohn in America [2] and has been developed in printed form at Mullard Research Laboratories. In this case the surrounding metallic block is replaced by two printed copper flat sheets which are not earthed. The advantage of the arrangement is that the small coupling gap between the lines which is otherwise required to achieve 3 dB coupling is avoided, hence the mechanical tolerances are less stringent. The solid-line re-entrant coupler uses slab-line for the feed arms, and a manufacturing method which takes advantage of this is shown by the power splitter in fig. 6. Here the slab-line inner conductors are made from ordinary copper wire and are sealed into the dielectric by electric current heating.

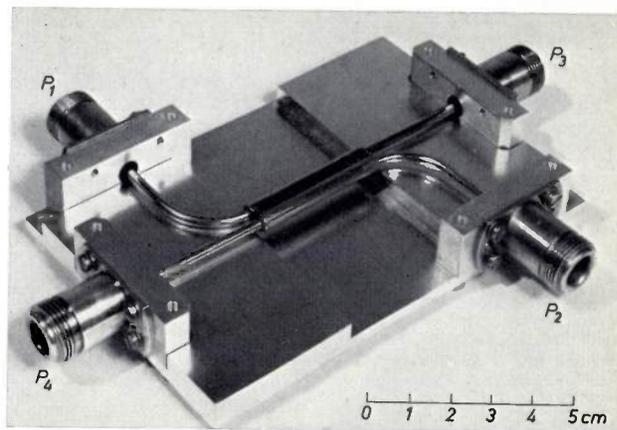


Fig. 5. Slab-line re-entrant ninety-degree coupler (1.0-2.6 GHz). Coupling 3 dB, isolation 23 dB. The two inner conductors pass through a metallic block forming a re-entrant section.

Microwave bridge circuits as frequency and phase-discriminators

The 3 dB ninety-degree coupler or the phase reversal hybrid can be used as a phase comparator. Interconnecting five couplers or hybrids with transmission lines produces a twin-bridge circuit having four outputs from one input. The relative amplitudes of the outputs depend on the line lengths and it can therefore be arranged that the outputs yield the input signal frequency without ambiguity. Such bridge circuits find applications for frequency and phase measurement in frequency metering and interferometer systems. High incremental accuracy ($< 1\%$) can be obtained over

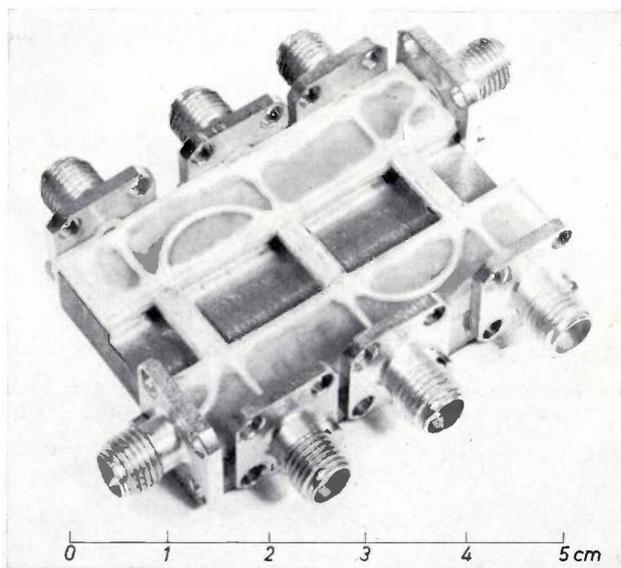


Fig. 6. Power splitter consisting of three re-entrant couplers. Here the slab-line inner conductors are sealed into the dielectric by electric current heating.

[2] S. B. Cohn, The re-entrant cross section and wide-band 3-dB hybrid couplers, IEEE Trans. on microwave theory and techniques MTT-11, 254-258, 1963.

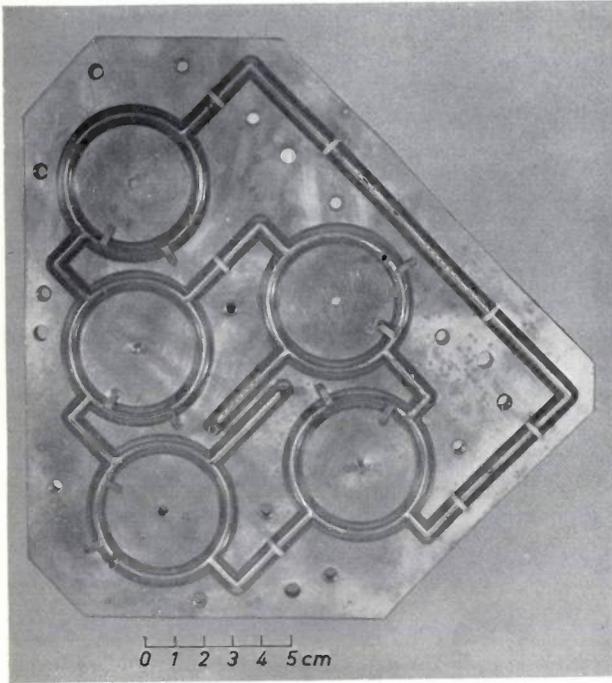


Fig. 7. Coaxial phase-discriminator (1.7-2.6 GHz) using solid-line, coaxial phase reversal ring hybrids.

wide frequency ranges. Fig. 7 shows a discriminator for the 2 GHz band which uses solid-line, coaxial phase reversal ring hybrids. Such a discriminator is extremely difficult and expensive to produce and in recent years printed techniques and ninety-degree couplers have therefore been employed. Fig. 8 shows a printed discriminator (within the band 2.5-4.1 GHz) with the ground plane removed. The inner conductors of the re-entrant couplers can be seen, as the auxiliary planes have also been removed.

Modulators, switches and limiters

The association of diodes with directional couplers leads to a number of other wideband balanced circuits. Among these are carrier suppressed modulators, microwave limiters and switches. A three-element passive limiter can be made by fitting diodes to six ports of the power splitter arrangement shown in fig. 6. The diodes are heavily reflecting at low signal levels but become matched to the incident signal power at a level of about 1 mW. Thus the insertion loss through the device is low at low level but will rise to about 20 dB at high level. This type of component may be used to protect a sensitive receiver from overload by large signals.

A *P-I-N* diode does not rectify microwave signals but acts as an electrically variable impedance. Circuits incorporating a pair of such diodes and a 3 dB coupler may be designed to operate as switches, attenuators and modulators. Such circuits have been constructed and

have achieved a rejection of 20 dB with good input match over the frequency range 2-6 GHz.

Microwave filters

Very broad band filters are difficult to construct in waveguide, whereas strip and coaxial line lend themselves well to filter design. A low-pass filter can be formed as a cascade of transmission-line sections which have alternatively high and low characteristic impedance (stepped-impedance filter).

Band-pass filters are formed by coupling a number of microwave cavities or resonant sections together. In these components strip and slab-line find their most powerful applications. The most popular forms of transmission-line filters are:

- 1) Half-wave resonant sections, quarter-wave coupled.
- 2) Double-comb (interdigital) transmission line.
- 3) Comb transmission line.

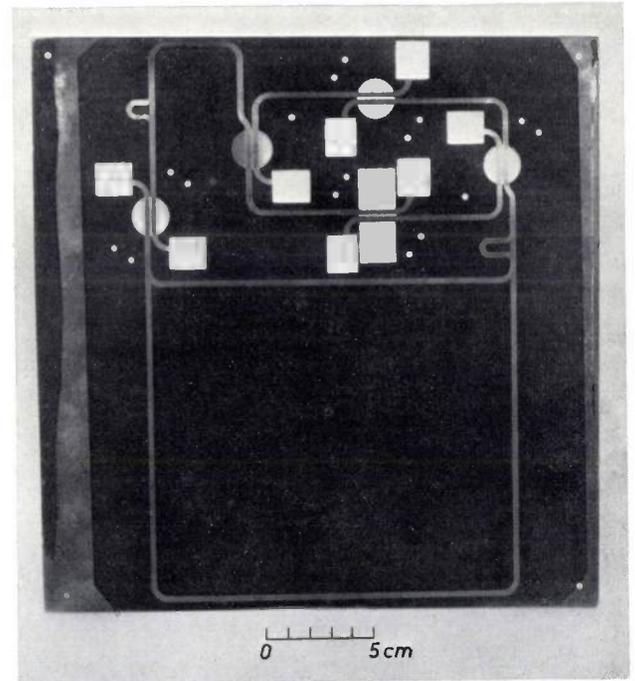


Fig. 8. Printed phase-discriminator (2.5-4.1 GHz), shown with the ground plane removed. The inner conductors of the re-entrant couplers can be seen, as the auxiliary planes have also been removed.

The double-comb (interdigital) filter was derived from the half-wave filter by Matthaei^[3] and is now extensively used. It consists of two sets of resonant fingers (combs), each finger being capacitively coupled to its neighbour and ideally short-circuited at one end and open-circuited at the other (fig. 9a). This type of filter can be made

[3] G. L. Matthaei, Interdigital band-pass filters, IRE Trans. on microwave theory and techniques MTT-10, 479-491, 1962.

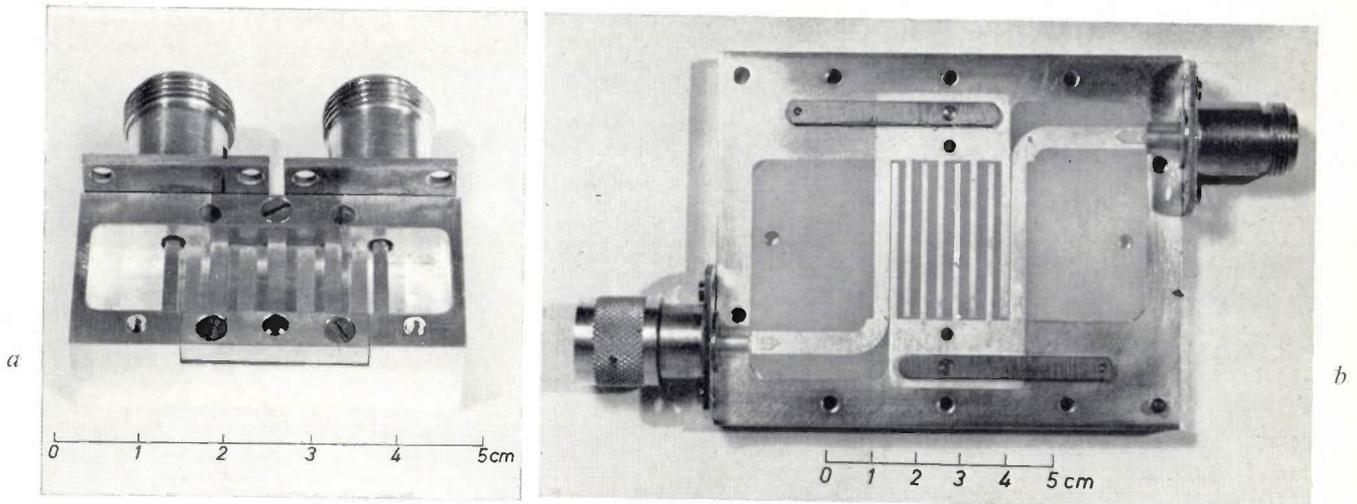


Fig. 9. *a*) Solid-line double-comb (interdigital) filter. Pass-band 4-8 GHz. Insertion loss 0.5 dB. Voltage standing wave ratio 2.0. It consists of two sets of resonant fingers (combs), each finger being capacitively coupled to its neighbour and ideally short-circuited at one end and open-circuited at the other.
b) Typical printed double-comb filter.

broadband (3 : 1 frequency range) or narrowband (less than 10% of centre frequency). The limits are imposed by the narrowness of the coupling gaps between the fingers and the width of the fingers themselves. The length of the fingers is approximately a quarter-wavelength at midband and this at high frequencies becomes very short. Since for good operation the length-to-width ratio of the fingers should be high (3 or more) the ground plane spacing has to be made small, which scales down all the dimensions except the fingerlength. Thus it is difficult to make double-comb filters work satisfactorily much above 10 GHz. The situation is improved by printing on high dielectric-constant material since for the same finger impedance the width is reduced more than the length (fig. 9*b*). Matthaei's design equations are approximate and it has been found useful to make these filters (when printed) as separate combs of beryllium-copper foil which are mounted between sheets of unclad dielectric. It is then fairly simple to move one comb relative to the other to provide some experimental adjustment of the coupling gaps. The foreshortening of the fingers at their open-circuit ends has to be found experimentally; for narrowband filters, especially above 3 GHz, this is very critical.

Also due to Matthaei and closely allied to the double-comb filter is the comb filter. This has only one comb, all the fingers being short-circuited at the same end. The other end of each finger must be loaded by capacitance in order to produce a pass-band so that comb filters are really only suitable for narrow bandwidths. They have the advantage over double-comb filters that they can be tuned over a wide frequency range by varying the capacitance. One comb filter can, for

example, be made to tune to any one of twenty UHF television channels. The comb filter is particularly well suited to the UHF range since the fingers can be made quite short, usually an eighth-wavelength at midband.

Other forms of coaxial and strip-line filter have been studied. A line with shunt open- or short-circuit stubs at quarter-wavelength intervals along its length forms a band-pass filter which is particularly useful for broadband operation. For narrowband performance the impedance values required for the stubs may be too high to be mechanically practicable. A printed band-pass filter (2.5-4.1 GHz) of this form is shown in fig. 10, together with a printed low-pass stepped-impedance filter of the type mentioned at the beginning of this section.

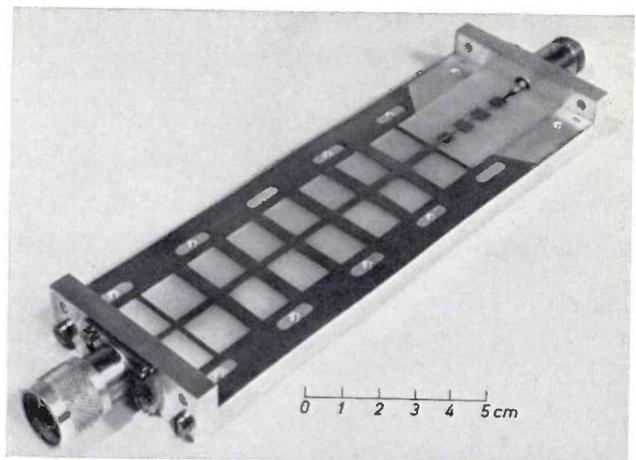


Fig. 10. Printed stubs and line band-pass filter (left) and printed stepped-impedance low-pass filter (right). Pass-band 2.5-4.1 GHz. Insertion loss 1 dB.

Construction techniques

High dimensional tolerances can lead to constructional difficulties with small microwave components, and careful mechanical design is necessary if they are to be made using normal machine shop practice.

As far as printing is concerned, the main problem is that the dimensional tolerances are a smaller fraction of the pattern size than is usual in printed wiring. The optical systems in the printing process and dielectric material stability impose undesirable limitations on what can be easily achieved.

Three techniques for printing strip-line microwave components have been useful. In the first method, most conventional but least accurate, a photographic negative is made of a drawing and the print then produced from the negative on copper-clad dielectric which has been coated with photoresist material.

In the second method, which is used for the most accurate work, a negative is cut by a precision machine in a thick acrylic sheet which has been coated with a photo-opaque material. An accuracy of a few microns can be achieved but the master negative may suffer dimensional changes with time and temperature.

The third method is very simple; a special laminated material is used in which the negative is prepared by cutting the top layer which is opaque and peeling it off. The accuracy obtained is of the order of 100 μm .

For all methods, considerable experience in etching the copper pattern is required for good and consistent results. It has been found that the copper patterns for double-comb filters can be reproduced to an accuracy better than 25 μm where the combs are made as a free foil and not on clad dielectric. In this case beryllium-copper is used to give the necessary mechanical strength.

The dielectrics are low-loss irradiated cross-linked polystyrene or polyolefin. These materials have dissipation factors of the order of 0.0001 in the microwave region. For narrowband filters the polyolefins are preferred as, apart from their somewhat lower loss, they are softer, a feature which is useful if small tuning screws have to be inserted.

Future objectives

Good performance much above 10 GHz is difficult to obtain in coaxial and strip-line devices. The reason

for this lies in the manufacturing techniques since the size (diameter or ground-plane spacing) has to be made small to prevent overmoding and the reduction in size increases the power loss in the device. In addition, any small discontinuities present have a greater effect at high frequencies than at low since they are not usually scaled down in proportion. The situation is aggravated by the poor length-to-width ratios of the line which are inevitable in components in which the line is resonant, i.e. when it is a quarter-wavelength long. One further difficulty, which is inherent in the use of conventional strip and coaxial line, is the comparatively poor performance of coaxial to strip-line transitions and coaxial connectors which have to be used. Even subminiature connectors become highly reflective at frequencies around 20 GHz. Nevertheless, some work at these frequencies is now being carried out using conventional printing techniques.

Some of these basic difficulties are now being tackled by building miniaturized components on ceramic material. Alumina materials having very low losses and dielectric-constants of about 10 are now available and on these strip-line can be laid down by evaporation. Using this technique the hybrid for a mixer within the band 7-11.5 GHz measures approx. 2.5 mm². It is intended to build arrays of integrated components in this way, i.e. mixers, filters etc. laid down on one piece of dielectric material without the need for separate connectors.

It is a logical further step to adopt a high-resistivity semiconductor substrate as the dielectric and to add devices such as diodes and transistors to the circuit by standard integrated circuit methods. Work is now going on, in co-operation with the microwave device groups of Associated Semiconductor Manufacturers Limited, to exploit this approach.

Summary. A short survey is presented of work on coaxial and strip-line passive microwave components. Most interest has been centred on the design of hybrid junctions and directional couplers, microwave mixers, filters and phase-discriminators. Some work has also been carried out on microwave modulators, switches and limiters. General considerations for good design technique are presented and reference is made to methods of manufacture. Results show that with modern printing techniques it is possible to build smaller and cheaper strip-line components comparable in performance with their coaxial counterparts.

An opto-acoustic cross-correlator in radar signal detection

J. S. Palfreeman

Introduction

The problem of obtaining a high distance resolution at the same time as a long range is familiar to all concerned with pulse radar systems design.

A long range requires a long period between transmitted signals and, in order to maintain an adequate signal-to-noise ratio, a high transmitted mean power; a high resolution needs a brief signal after reception. In short range, high resolution radars, this brief received signal is occasioned by a similarly brief transmission.

The peak transmitted power usually has some finite limit, due for example, to dielectric voltage breakdown in the aerial system and this limitation restricts the minimum ratio of the transmitted pulse length to the period between transmitted pulses for a given mean power. Thus in a simple pulse radar, the minimum pulse length, and hence the maximum resolution for a given range, is limited by the peak power capabilities of the system.

Pulse compression systems [1] are one way of avoiding this limitation. The emitted signal occupies a large bandwidth and is transmitted for a long time, for example as a swept frequency. This signal after reception is time compressed into a pulse of short duration $\Delta T = 1/\Delta F$, where ΔF is the signal bandwidth. Such systems require an accurate specification of the transmitted pulse and of the pulse compression network.

Correlation detection is an alternative method of obtaining this pulse compression which does not depend upon the exact nature of the transmitted signal, but merely upon the "goodness of fit" between a delayed replica of the transmitted signal and the returned echo from the target.

This article discusses correlation detection of signals in radar, shows the reason for the choice of an opto-acoustic system to perform the correlation, and discusses a novel, two delay line correlator in which one signal is time reversed. The opto-acoustic design considerations for a suitable system are briefly considered, the practical design discussed and its potential performance described.

Correlation detection

The cross-correlation function $C(T_r)$ [2] is a measure of the degree of similarity of two functions of time,

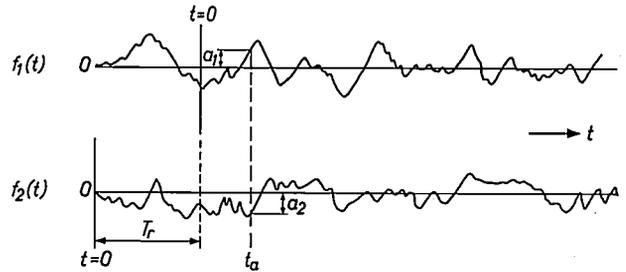


Fig. 1. Cross-correlation. The figure shows two functions of time $f_1(t)$ and $f_2(t)$, the former delayed by a time T_r . The cross-correlation coefficient is the long term average of all the products of coincident sample pairs in the two waveforms. The product at the sample t_a is $(a_1)(-a_2)$.

$f_1(t)$ and $f_2(t)$, and is the "long term average" of the product of all the coincident pairs of samples of the two functions one of which is delayed by the time T_r (T_r is in this case the time interval between the transmission of the radar signal and its reception) (fig. 1). This may be written for an aperiodic function as:

$$C(T_r) = \int_{-\infty}^{+\infty} f_1(t) f_2(t + T_r) dt.$$

In the case of radar in which a pulse of finite duration T is transmitted, the signal may be taken as zero outside the time T and hence the integral outside of the pulse is also zero so that we may then write the integral:

$$C(T_r) = \frac{1}{2T} \int_{-T}^{+T} f_1(t) f_2(t + T_r) dt. \quad \dots \dots (1)$$

In the proposed system the two waveforms have the same source, i.e. the transmitted pulse, but the received echo from the target will be diluted with noise and it would probably be more correct to replace $f_2(t)$ with $f_1(t - T_r) + f_N(t)$ where $f_N(t)$ is the unwanted noise. Thus the integral becomes:

$$C(T_r) = \frac{1}{2T} \int_{-T}^{+T} f_1(t) [f_1(t) + f_N(t + T_r)] dt.$$

[1] H. O. Ramp and E. R. Wingrove, Principles of pulse compression, IRE Trans. on military electronics MIL-5, 109, 1961.

[2] P. M. Woodward, Probability and information theory, Pergamon Press, London 1957.

This process is a powerful means of recognizing a signal submerged [3] in noise when one has a knowledge of the exact nature of the signal. It is thus of use in radar and sonar where these conditions frequently exist, and has noise suppression properties similar to those of the matched filter.

One simple application of this type of detection to a radar signal is shown in *fig. 2*. The output from the transmitter is fed to the aerial and to a variable delay line of delay T_d . An echo received from the target after a time T_r is correlated in the multiplier with this delayed signal, and the integration is achieved by the low pass filter. A maximum correlation occurs when $T_r = T_d$.

Fig. 3 illustrates a multi-channel correlation detector in which the signal delay is achieved by a tapped delay line, the number of taps corresponding to the number of discrete range elements required and the total delay to the maximum range, a signal at a given range resulting in an output from a particular element. In a typical radar with a range of 150 km and a desired resolution of 150 m the delay line will require 1000 taps and a total delay of 1 ms with a bandwidth of at least 1 MHz. This performance cannot at the moment be obtained in a single unit but it is conceivable that in the future an integrated circuit approach may prove practical.

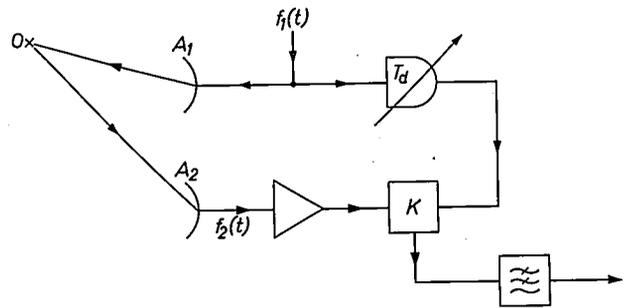


Fig. 2. Variable delay correlator for radar signal detection. The diagram shows the input $f_1(t)$ from the transmitter being fed to both the radar transmitting aerial A_1 and variable time delay T_d . The signal $f_2(t)$ from the target O is fed from the receiving aerial A_2 , after amplification, to a multiplier K , the other input to which is the output from the variable delay T_d . The multiplier output is fed through a low pass filter. The multiplier and low pass filter form the correlator. The whole range is scanned by varying the delay time.

interval the stress wave has taken to reach a given point.

In the opto-acoustic cross-correlator [5] the received signal modulates a light source which illuminates the photo-elastic delay line with the function $f_2(t)$. The light intensity is modulated a second time by the function $f_1(t)$ by the photo-elastic delay line, the time delay T_d corresponding to a distance along the delay

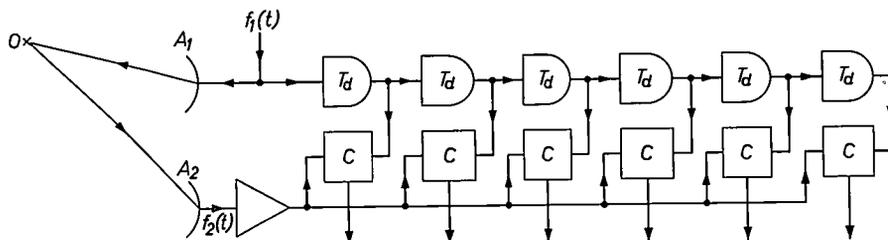


Fig. 3. Tapped delay correlator. This figure shows a multiple tapped delay, the output from each tap feeding a correlator C .

Single-delay line correlator

Fig. 4 shows one method of achieving this type of performance without the use of a multi-tapped line and individual multipliers. In this system a photo-elastic delay line replaces the tapped delay line.

In a photo-elastic delay line, incident light is modulated by the stress in the delay medium so that at any instant the complete stress distribution in the line is visible [4]. (The physical mechanism used in this process will be described later but for the present we will assume that this is a linear process.) The photo-elastic delay line forms an infinitely tapped delay line in which the distance along the bar represents the time

line. The multiplication process is this double modulation of the light and the integration is achieved by viewing over a period. The presence of a target at a given delay range will be indicated by a high mean light intensity at the point in the bar corresponding to that delay.

The distribution along the bar of the total quantity of light transmitted in any one correlation period is the function $C(s)$ in which the variable s of position along the bar is equivalent to T_r in equation (1). In the ultrasonic case when both signals $f_1(t)$ and $f_2(t)$ are modulated on a carrier of frequency f_0 , the light intensity fluctuates at this frequency.

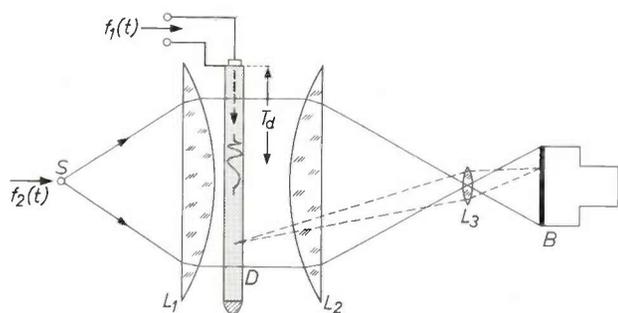


Fig. 4. Single delay line "spatial" correlator. A light source S modulated by signal $f_2(t)$ illuminates an ultrasonic delay line D . The light is collimated by the lens L_1 upon D . The input $f_1(t)$ to the delay line again modulates the transmitted light. Lens L_2 images the source S upon a third lens L_3 which forms an image of the bar on the storage tube B .

The spatial resolution (the reciprocal of the light spot length along the bar) is proportional to the signal bandwidth, and the time over which this correlation from a single target is formed is the time duration of the transmitted signal.

This is inconvenient for system application since a large number of integrating photo-detectors distributed along the bar would be required to identify the range of a target. If a "one lead" output were required, with the time of a signal corresponding to the range of a target, some sampling mechanism would be required to turn this parallel access into a time sequential system.

This conversion from parallel-to-sequential may be overcome by using electronic integration achieved by imaging the photo-elastic delay line on a storage tube and reading the stored signals by the scanning electron beam. This overcomes the photo-detector array problem.

In practice both of the signals modulating the light would be present on a carrier frequency and it would be necessary on the storage tube to resolve the equivalent spatial frequency corresponding to the acoustic wavelength of this carrier frequency. This may be several times the necessary resolution determined strictly from a bandwidth criterion and it is again not an ideal system, suffering in particular from storage tube non-linearity and high noise levels.

Two delay line correlator

The method under consideration at Mullard Research Laboratories [6] involves the use of two photo-elastic

lines and is shown in fig. 5.

The transmitted signal is launched in one delay line which modulates the light with the function $f_1(t - T_1)$. The received signal launched in the second line in the reverse direction modulates the light with $f_2(t - T_2)$ and the product of these two light modulations is again a correlation, spatially distributed along the delay line, with the difference that the fluctuations in light intensity occur at a carrier frequency of $2f_0$. The figure shows the modulation as a swept frequency, and it will be seen that only at one point on the delay lines do identical parts of the waveform correlate and that the correlation exists for the total signal duration: at all

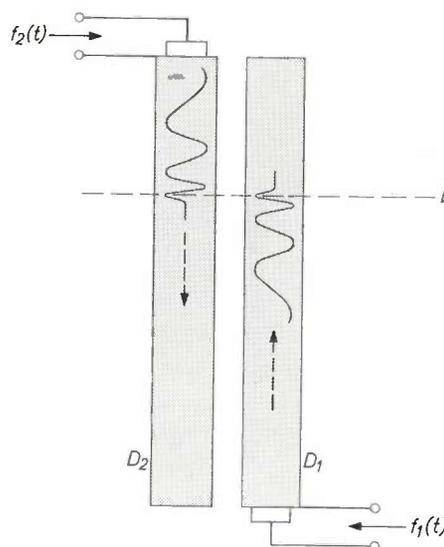


Fig. 5. Two delay line correlator. The two swept frequency signals $f_1(t)$ and $f_2(t)$ shown in the bars D_1 and D_2 travelling in opposite directions correlate as they pass the point b . The correlation exists at this point for the whole time of transit, i.e. a complete signal length.

other points on the line the correlation will be small. This situation is again not ideal for the reasons outlined in the previous section.

However, consider the case of one signal in the bar reversed in time, fig. 6, which again shows two swept frequency signals propagating in opposite directions, one in each photo-elastic bar. One signal has the lowest frequency at the beginning of the pulse, the other, time reversed signal has the highest frequency at the beginning. The correlation now is a maximum at the instant when both signals are completely superimposed; at all other times the correlation is small. On the other hand the correlation at this instant will exist along the whole length of the pulse.

[3] M. I. Skolnik, Introduction to radar systems, chapter 9, McGraw-Hill, New York 1962.

[4] C. F. Brockelsby and J. S. Palfreeman, Ultrasonic delay lines and their applications to television, Philips tech. Rev. 25, 234-252, 1963/64.

[5] R. W. Wilmotte, Instantaneous cross correlation, final report R.A.D.C. Contract No. AF 30, AD 215485, Sept. 1957.

[6] A. Browne, An optical correlation technique, IEE Conference on delay devices for pulse compression, 48-57, 1966.

The situation is the reverse of the one previously described in that we have interchanged the "time" domain for the "space" domain. The integration now occurs as the result of summing the transmitted light at all points along the bar and the correlation coefficient now is a direct function of time, $C(T_r)$. The correlation exists for a time which is short compared with the pulse length, approximately the reciprocal of the bandwidth of the signals being correlated.

This process has achieved directly the desired "one lead" output. The correlation is now detected by imaging the whole of the light passing through the bar on a photo-detector and the correlated signal will be

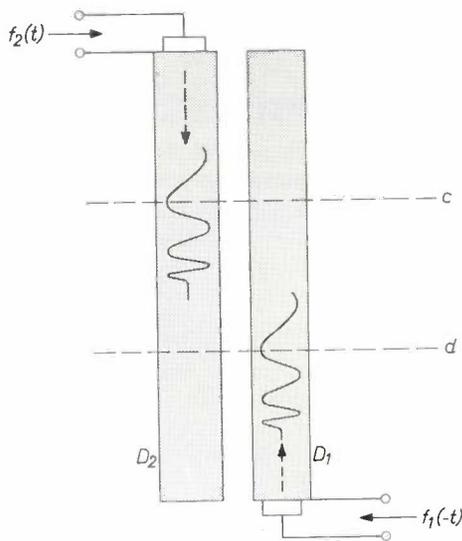


Fig. 6. Two delay line correlator. One of the two signals, $f_1(-t)$, is now reversed in time and the signals correlate when they are completely superimposed. The correlation occurs only for a short time duration but occupies the entire pulse length between c and d .

present on a carrier frequency which is twice the carrier frequency used to modulate the light.

This system is a type of pulse compression system and may be used in any conventional pulse compression application. It furthermore has the advantage that, assuming the time reversal of one of the signals can be achieved, the exact detail of the transmitted pulse is of no importance (on a pulse compression system using linear FM a high degree of sweep linearity is required) either within the pulse or from pulse to pulse, since a given radar return is correlated only with the transmitted signal which occasioned it.

The photo-elastic process

When an isotropic transparent medium is stressed it becomes birefringent [7] and the effective refraction

index becomes dependent upon the direction of the stress and the polarization of the light. In the case of an extension the strains and hence refraction index changes lie along and normal to the direction of the originating stress. Hence a light ray polarized along one of these vectors will be advanced in phase relative to a light ray polarized along the other.

A shear strain may be resolved into an extension and a compression along axes at 45° to the direction of shear, a similar situation with the maximum and minimum phase differences occurring along these extension and compression axes (see fig. 7).

In acoustic delay lines the stress wave may be

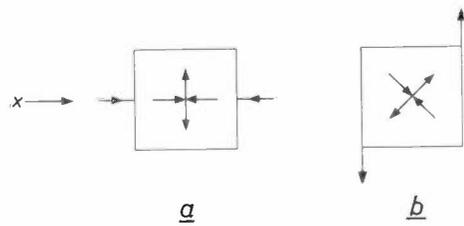


Fig. 7. Stress axes for longitudinal (a) and shear (b) stress waves propagating in the direction x .

propagated in the medium in both the longitudinal (extension) and the shear mode of vibration. The medium must be transparent and of a low acoustic loss with preferably a high stress optical coefficient and a high refractive index. The material most suitable is undoubtedly fused quartz although both glass and water are possible under some circumstances. The preferred mode of propagation is shear; this is due to the lower shear velocity which increases the delay obtained in a given length by about 50% and also to the freedom of the shear mode of propagation from mode conversion at the delay line boundaries [8] which would cause spurious signals.

Fig. 8 shows the birefringent process. Plane polarized light obtained by passing the incident light through a polarizer may be resolved into two equal orthogonal components in phase. Transmission of these through the birefringent material produces a phase difference so that the light is now elliptically polarized. Analysing this light by an analyser at right angles to the polarizer gives rise to a single component whose intensity is:

$$I = \frac{1}{2} I_0 (1 - \cos \Theta) = I_0 \sin^2 \Theta/2, \dots (2)$$

where Θ is proportional to the length of the opto-acoustic interaction and to the stress σ , which in the ultrasonic delay line is in turn proportional to the transducer drive voltage E . This characteristic is shown

in fig. 9. When the polarizer and analyser axes are orthogonal and the birefringent material is unstressed, no light is transmitted. Small stresses have a roughly square law stress/transmitted intensity characteristic so that a sinusoidal applied stress is frequency doubled (see σ_1 and I_1 in fig. 9).

A more linear mode of operation may be obtained by a static bias to the centre of this characteristic which corresponds to a phase difference of the two orthogonal

components of $\pi/2$ (σ_2 and I_2 in fig. 9). This bias may be obtained by a static stress in the bar or, more commonly, by the use of a "quarter wave plate" placed between the polarizer and analyser with its axes at $\pi/4$ to the direction of polarization. In this case the intensity is:

$$I = \frac{1}{2} I_0 (1 + \sin \Theta)$$

and is linear to 5% for a range of intensity of $\pm \frac{1}{4} I_0$.

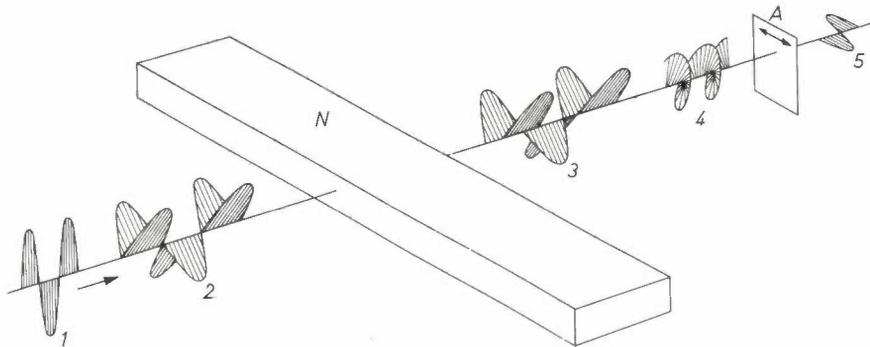


Fig. 8. Detection of birefringence. Plane polarized light at 1 may be resolved into two orthogonal components, shown at 2. After transmission through the birefringent material N one component is phase delayed relative to the other, at 3. These two components combine to give elliptical polarization shown at 4. Transmission through an analyser A results in plane polarized light at 5 whose intensity is dependent upon the birefringence.

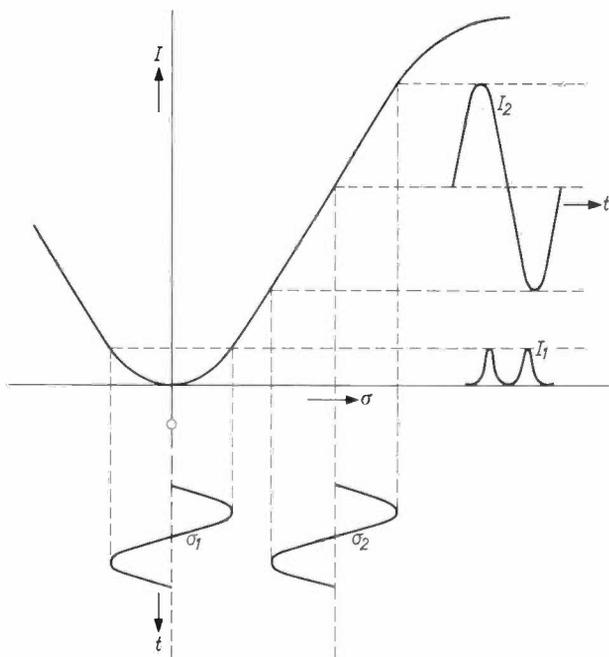


Fig. 9. The light intensity I transmitted by the birefringent medium is shown as a function of stress σ . An alternating stress σ_1 for the case when the polarizer and analyser axes are orthogonal and no quarter wave plate is included results in an output intensity I_1 which is frequency doubled. The addition of a quarter wave plate biases the characteristic to the centre point resulting in a stress σ_2 giving rise to a linear intensity modulation I_2 .

Ultrasonic transducers

The ultrasonic transducer most frequently used in the construction of delay lines is a thin slice of piezoelectric quartz crystal fundamentally resonant in its thickness mode at the carrier frequency [8]. This transducer is bonded to the fused quartz medium by a conducting layer, which forms the front electrode of the crystal, and a rear electrode is deposited upon the back face of the crystal. The drive voltage is applied between these electrodes. Either shear or longitudinal modes of vibration may be generated depending upon the direction of the crystal axes in the slice.

The frequency response of such a transducer/delay medium combination is an arithmetic series of passbands at the odd harmonics of the crystal resonance, each passband having a fractional bandwidth of rather over 50% of the fundamental carrier frequency, so that for a passband of 10 MHz a band centre of about 20 MHz is required. Recently piezoelectric ceramics such as lead zirconate titanate (the Philips PXE range of materials) have found increasing use as delay line

[7] M. Born and E. Wolfe, Principles of optics, Pergamon Press, London 1959.

[8] C. F. Brockelsby, J. S. Palfreeman and R. W. Gibson, Ultrasonic delay lines, Iliffe, London 1963.

transducers at frequencies of up to 10 or 15 MHz^[9] [10]. The advantage of these materials is in the very high electromechanical coupling coefficient k which for crystalline quartz is about 0.14 and for PXE3 is about 0.7. Since the power conversion efficiency is proportional to k^2 this forms a very significant improvement although the very high dielectric constant of this material (about 1500 compared with 4.5 for quartz) is in this case an embarrassment since this creates a very high transducer capacitance.

At fundamental frequencies above 15 MHz the grain size in the PXE material becomes comparable to the thickness and this material is no longer of use. A new ceramic, potassium sodium niobate (P.S.N.), has, however, a very much finer grain size, partially due to the hot pressing technique used in its manufacture, and also has a considerably lower dielectric constant of about 450 and a coupling coefficient of about 0.6^[11].

It is also possible that a recently investigated single crystal ferroelectric, lithium niobate, large crystals of which have been successfully grown at Mullard Research Laboratories, will prove useful as a piezoelectric transducer. This material has a coupling coefficient of about 0.54 and a relative dielectric constant of about 80. Problems exist in bonding this material satisfactorily to the fused quartz delay medium.

In order to reduce the transducer capacitance which with these high dielectric constant materials may be as high as 10 000 pF, and thus raise the transducer impedance to a suitable level for convenient electrical drive, the transducer is sub-divided and the sections are connected in series as shown in *fig. 10*. This reduces the capacitance by a factor r^2 where r is the number of sections.

Ultrasonic path cross-dimensions

The acoustic wavefront in the "near field" or "Fresnel" region is plane^[8] and here no significant beam spread occurs. For a frequency of about 20 MHz and a transducer cross-dimension of 1 cm, the near field extends to approximately 7 cm. Beyond this region (in the "far field") the wavefront gradually assumes a spherical form.

In order to avoid excessive beam spread in this far field which in a narrow bar would result in wavefront errors due to reflection from side walls and in a wide bar to excessive curvature of wavefront and loss of sound intensity, it is desirable to make the transducer

many wavelengths wide. A minimum cross-dimension of 50 wavelengths (about 9 mm in a fused quartz medium at a frequency of 20 MHz) is typical.

Light modulation efficiency

For a constant depth of modulation the product of the drive power and the acoustic beam width must be constant.

Optical considerations, however, limit the width of the interaction since any light ray contributing fully to the effect must lie within two adjacent acoustic wavefronts. An oblique ray which crosses two or more sound wavefronts will have the modulation partially cancelled since it will pass through regions of positive and negative birefringence. Thus for a given light beam divergence φ and a given sound wavelength λ_s the maximum interaction width is limited to λ_s/φ .

In general φ can only be decreased at the expense of light flux. The requirements of the final signal-to-noise ratio and the maximum intensity of available light sources affect this consideration.

A typical interaction width for the system with a 20 MHz carrier and 10 MHz bandwidth would be about 2 cm for a maximum light divergence of 5 milliradians.

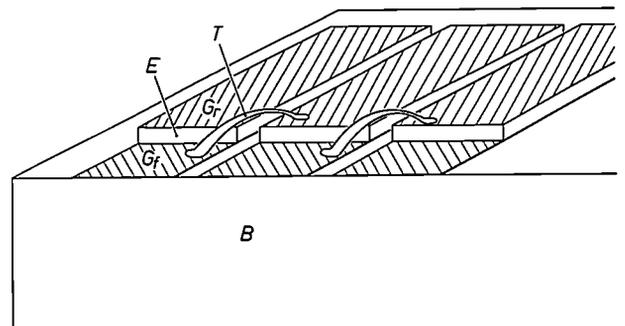


Fig. 10. The transducer E bonded to the fused quartz medium B is subdivided into a number of sections in order to reduce the transducer capacitance. The front electrodes G_f are connected to the rear electrodes G_r with tape interconnecting leads T .

Correlation system

The correlation system is shown in *fig. 11*. A light source which is a 1 kW compact mercury arc is imaged upon a source slit by the condenser lens L . Mirror M_1 collimates the light upon the photo-elastic delay lines D_1 and D_2 . The polarizer P polarizes the light either along or normal to the bar in the case of shear waves, and the analysers A_1 and A_2 either parallel or

[9] R. W. Gibson, Solid ultrasonic delay lines, *Ultrasonics* 3, 49-61, 1965.

[10] C. M. van der Burgt, Transducer materials and isopaustic glasses for delay lines, *Ultrasonics*, 1967 (to be published).

[11] Transducers using this material have been made fundamentally resonant at frequencies as high as 75 MHz.

orthogonal to this direction. The first analyser forms the polarizer for the second system.

The quarter wave plates Q_1 , Q_2 must be correctly orientated as mentioned previously and must lie between the appropriate analyser and polarizer, although the order in which the light passes through the fused quartz bar and the quarter wave plate is not important. The light, after passing through the system, is imaged upon a photo-multiplier by M_2 .

the signal-to-noise improvement for weak signals below noise level, as would be expected with a 1000 : 1 pulse compression, is 30 dB.

Time reversal

For this system to succeed, as mentioned previously, it is necessary to reverse one signal in time. It appears preferable to select the transmitted signal for this operation since its amplitude is less variable than an

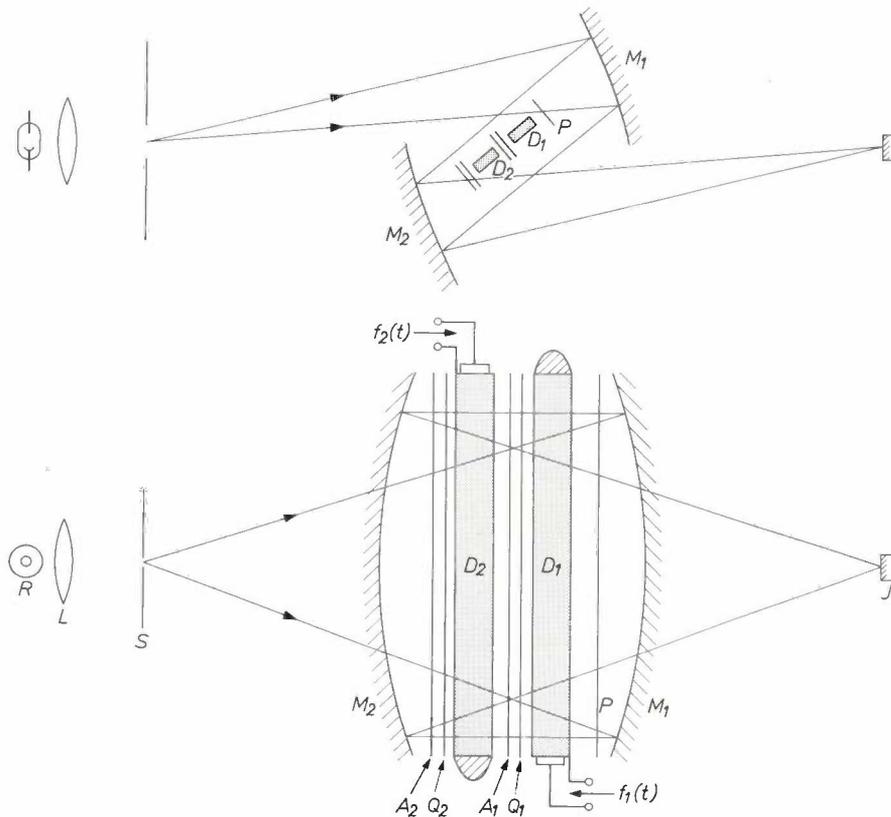


Fig. 11. Two delay line correlator (viewed in plan and elevation). The light source R is imaged by the condenser lens L upon the source slit S . The light is collimated upon the photo-elastic delay lines D_1 and D_2 by mirror M_1 . Light passes in succession through the polarizer P , the photo-elastic bar D_1 , the quarter wave plate Q_1 , analyser A_1 , second photo-elastic bar D_2 , quarter wave plate Q_2 and analyser A_2 . This light is then imaged upon the photo-detector J by the mirror M_2 .

The delay lines must be twice the length necessary to accommodate the signal to be correlated. Thus if the signal for correlation is of $100 \mu\text{s}$ duration, the lines must be $200 \mu\text{s}$ long, which for shear waves in fused quartz is a length of about 73 cm. In systems of this length lenses are impractical and are replaced by mirrors.

Pulse compression ratios of 1000 are possible with systems of this type. The final signal-to-noise ratio after correlation for noise-free radar returns is restricted to about 50 dB by the practical limits of the system and

echo which may well have a total range of amplitudes of 60 dB.

Two methods have been proposed for this reversal which requires storage of the signal for a time equal to its duration, the first using a similar opto-acoustic system and the second a storage tube.

Fig. 12 illustrates the first system using lenses as the optical components although mirrors may equally well be used. The signal to be reversed is fed into a photo-elastic line which modulates the light intensity in a linear fashion as in the correlation system. The trans-

mitted light intensity through this system is then a replica of the stress pattern in the photo-elastic delay line and this light pattern is moving at the sonic velocity in the line.

The optical system forms a minified and inverted image of this light pattern upon the second delay line. The minification of this optical system is exactly 2 : 1 so that this optical image occupies half the length of the original pattern and moves at half the acoustic velocity. The second delay line is operated in the cut-off mode (fig. 9) and a single stress pulse whose total time duration is short in comparison with the reciprocal of

screen area, the writing and reading spot will be defocused in a radial direction by the use of a rotating quadrupole magnetic deflection (a means of producing radial astigmatism).

This latter system appears to be capable of producing an adequate signal-to-noise performance and resolution, and will probably be preferred to the former.

Range improvement

The system so far described will only correlate over a total time given by the difference between the total photo-elastic delay line length and pulse length. Thus

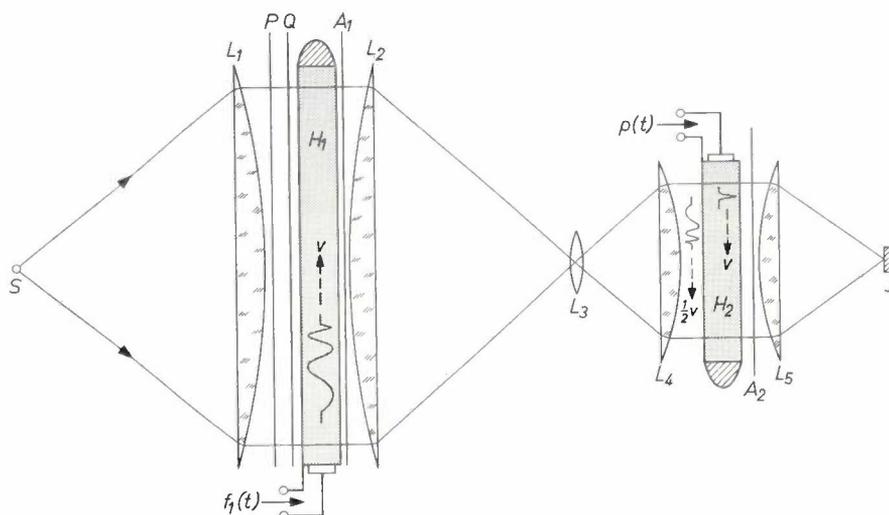


Fig. 12. Optical system for time reversal. Light from the source slit S is collimated on the first acoustic delay line H_1 . This delay line contains the signal to be reversed $f_1(t)$, which travels down the bar with the sound velocity v . Using the polarizer P , quarter wave plate Q and analyser A_1 , the light is modulated with this function. Lens L_2 images this modulated light upon the lens L_3 which, in turn, forms a minified image of H_1 upon the second delay line H_2 . Lens L_4 collimates the light upon this line. The light incident upon H_2 contains a 2 : 1 minified image of the function $f_1(t)$ which travels at a velocity $v/2$, a sampling pulse $p(t)$ travels down the line H_2 at the velocity v . This pulse overtakes the optical image and "interrogates" it in a time exactly equal to the duration of the original pulse. Analyser A_2 detects the transmitted signal and the light is imaged upon the photo-detector J by lens L_5 .

the bandwidth of the signal to be inverted is fed into it. This single stress pulse, moving at the full acoustic velocity, "interrogates" the moving light pattern, taking exactly the original pulse duration in which to accomplish this. The resulting total light intensity transmitted through this second photo-elastic line is a time reversed replica of the original input signal.

In the second system, the transmitted signal is recorded by writing it on a storage tube^[12], and reversed by reading it backwards. In order to utilize a large area of the storage surface and hence enhance the signal-to-noise ratio of the system, it is proposed to write the signal in a spiral scan around the outer third of the storage screen. In order to use an even greater

for a pulse length of 100 μs and a line length of 200 μs the correlation time will be 100 μs . Ideally the input signal will be delayed by a pulse duration prior to propagation in the photo-elastic line.

A second correlator can be time shared with the first to obtain a complete 200 μs correlation period.

In order to extend the range still further the transmitted, reversed, signal can be read from the storage tube a number of times until the complete radar range has been covered, at which time it is erased. This period must, of course, be less than the repetition period of the radar.

[12] M. Knoll and B. Kazon, Storage tubes, Wiley, New York 1952.

System limitations

The system limitations arise from two main sources, the practical restrictions of the acoustic system and the restrictions due to the quantum nature of light.

The former restrict the length of the photo-elastic delay line to about 200 μs (about 75 cm for shear waves in fused quartz), so that pulses of about 100 μs duration may be correlated. The acoustic loss down such a line restricts the maximum frequency to typically 60 MHz, unless very low loss quartz is selected and the delay line is operated at an elevated temperature when perhaps 80 or even 100 MHz may be possible.

Assuming that an acoustic fractional bandwidth of 50% can be maintained, the typical figure leads to possible pulse compressions of 3000 : 1 or a signal-to-noise improvement of 35 dB.

The limitations of the system due to the quantum nature of light arise from the practical considerations of the total useful light flux emitted by the light source and the sensitivity and maximum light detection capability of the photo-detector.

Since the photo-detector must resolve twice the ultrasonic carrier frequency, it must have an acceptable frequency response at around 60 MHz. This fact and the required high sensitivity and low noise requirements suggest a photo-multiplier as the only presently useful photo-detector.

The total useful light flux and the photo-detector sensitivity together form the lower boundary for a detected signal. The upper boundary is formed by the combination of this total light flux and the maximum light detectable by the photo-detector.

Photo-cathodes typically are limited to maximum photo-electron currents of about 1 $\mu\text{A}/\text{cm}^2$, corresponding to 10^5 electrons per cycle of carrier at the maximum frequency $2(f_0 + \Delta f)$ of 60 MHz, which approximates to a noise figure of 50 dB. If the minimum detectable signal from the equipment is set at 0 dB signal-to-noise level, the maximum dynamic range will be the full 50 dB.

If the equipment has a pulse compression ratio of 1000 : 1 (e.g. 10 MHz bandwidth and 100 μs pulse duration) the input signal-to-noise ratio to achieve this maximum sensitivity of 0 dB output level will be

—30 dB and saturation will occur for input signals of greater than 20 dB above noise level.

The light source used for this equipment is a 1 kW compact source mercury arc and the light is filtered to pass only the 0.405 μm and 0.436 μm lines. A gas laser would form a more suitable light source from most points of view, but at present the need for high light output at the peak of the photo-cathode spectral response makes such a laser very expensive, and an adequate performance can be obtained from the mercury arc.

A significant improvement could perhaps be made with the use of a semiconductor photo-detector which has a much higher quantum efficiency than a photo-cathode, but problems of internally generated noise in the semiconductor and the lack of a suitable internal gain mechanism at present prevent this.

Conclusion

The correlation system described has several advantages over other pulse compression systems in that accurate pulse compressors and expanders are not required and that the radar output need not be accurately specified.

Some improvements in the components, e.g. light source, photo-detector, and the success of present work on ultrasonic transducer development could produce an attractive device. Its major limitation will then be in the physical size of the equipment, making its use possible only in situations where space is not at a premium.

Summary. Correlation detection has long been known to form an ideal means of extracting a reflected radar signal from noise. Another property of the correlation detector is that a long, broadband signal can be time-compressed to provide a short collapsed pulse. Hitherto the realizations of this type of detector have suffered from a great complexity. The proposed system, although bulky, using large optical components, is essentially simple. It uses two photo-elastic delay lines to carry the transmitted and received signals. The correlation process requires the multiplication of coincident samples of the two signals and this product is formed by the double modulation of light passing sequentially through these two delay lines. In the general case of this system, one signal requires reversing in time and means for achieving this are described. System bandwidth of 10 MHz and correlation times of about 100 μs are possible giving a signal-to-noise improvement of 30 dB and a pulse compression ratio of 1000 : 1.

A digital direction finder

R. N. Alcock

Introduction

Interferometer techniques have been applied to radio frequency direction finding for some years. The direction of a radio frequency transmitter may be determined by receiving the signal at two points with known separation and by comparing the phases of the received signals. Such an interferometer will either provide an accurate indication of direction over a narrow field of view or an inaccurate indication over a wide field of view, depending on the separation of the two receiving points. A novel interferometer system has been built which works at microwave frequencies and gives an accurate indication of direction over a wide field of view and presents the information in digital form. This is achieved by receiving signals at more than two points and by exploiting digital processing techniques.

The system has several applications. It may be incorporated in aircraft and helicopters as an aid to blind landing. The position of pulsed microwave beacons on the ground can be accurately determined in digital form and the information used to display a perspective view of the ground beacon pattern. Alternatively the system could be placed at the end of a runway to receive the weather radar signals from an aircraft and use these to monitor accurately its glide path. The data would be transmitted to the aircraft or used to calibrate other landing aids.

The system also provides a navigational aid for ships in rivers or estuaries, the position of ships being determined accurately from bearing measurements on the ship's radar (or special beacon) at two land stations, or at one station in conjunction with secondary radar.

Principle of operation

Fig. 1 shows an interferometer consisting of two aerials with spacing d feeding a phase discriminator. A plane wave from the transmitter arrives at an angle θ with respect to a line through the aerials. The measured phase φ is given by:

$$\varphi = \frac{2\pi d}{\lambda} \sin \theta.$$

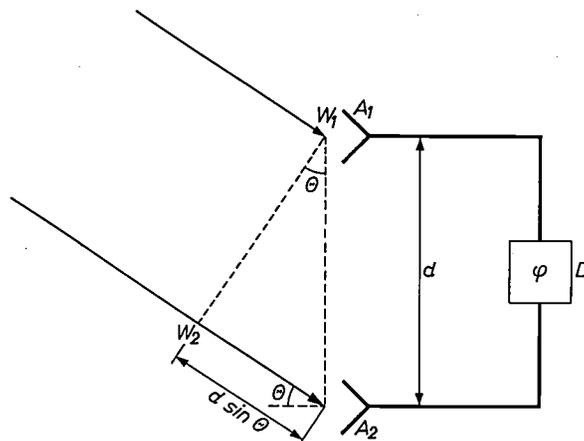


Fig. 1. Bearing measurement with an interferometer. A plane wave W_1-W_2 passes through two spaced aerials A_1 and A_2 which feed a phase discriminator D . The measured phase φ depends on the direction θ of arrival of the wave.

Fig. 2 shows the variation of phase with bearing for two interferometers with spacings d and $4d$. For clarity θ is assumed to be small so that $\varphi = 2\pi d\theta/\lambda$. It is seen that the phase measured at the larger interferometer varies four times more rapidly with bearing than that at the smaller interferometer. The larger interferometer thus provides a more accurate meas-

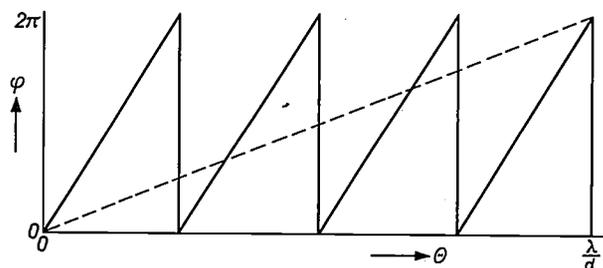


Fig. 2. Variation of phase with bearing. The variation of the measured phase φ with signal bearing θ is shown for an interferometer with spacing d (dotted line) and an interferometer with spacing $4d$ (unbroken line). The phase varies four times more rapidly in the larger interferometer but jumps from 2π back to 0 giving an ambiguous indication of bearing in the range $0 < \theta < \lambda/4d$.

ure of bearing in the presence of phase measuring error. However, multiples of 2π cannot be distinguished by the discriminator so that signals at four different bearings give rise to the same phase difference and the larger interferometer gives an ambiguous measurement when used over the same field of view as the smaller interferometer ($0 \leq \theta \leq \lambda/d$). It is clear that, by the use of two interferometers and appropriate processing of two phases, the ambiguity of the larger interferometer can be resolved by reference to the smaller one, en-

terminated by the smallest. The system is now considered in more detail.

Digital phase measurement

The receiver operates over a bandwidth of 700 MHz around 9.5 GHz. Three pairs of microwave aerials in the form of horns feed three discriminators, one of which is shown in *fig. 4*.

Each pulsed microwave signal is split into two paths by magic tee hybrid junctions *K* and *L*. The signals re-

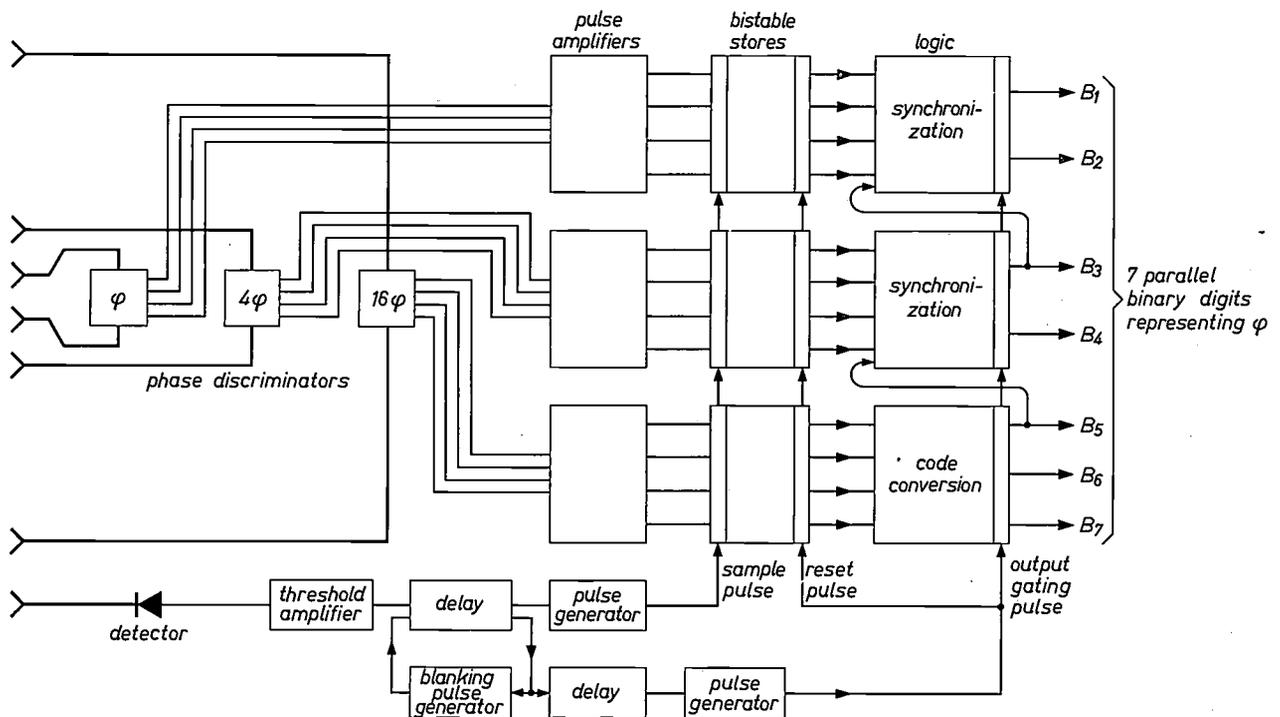
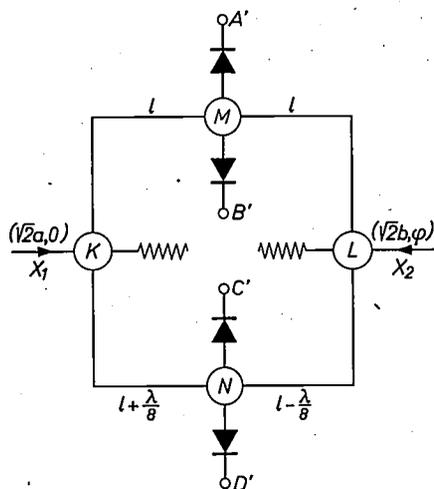


Fig. 3. Digital direction finder. Pulsed signals are received in three interferometers with spacings in the ratio 1 : 4 : 16. Detected pulses from three phase discriminators indicate phase digitally and pass in parallel through amplifiers to trigger bistable stores. The three digital phases are processed in logic to obtain binary indication φ which is related to signal bearing θ .

abling the accurate measurement to be extended to a wider field of view.

This principle has been extended to three interferometers in the system which has been built (*fig. 3*). Three interferometers with spacings in the ratio 1 : 4 : 16 provide three phase measurements which are converted directly into digital form, amplified and processed by a logic unit which combines the digits to provide a measure of bearing in binary code. The accuracy is that of the largest interferometer with a field of view de-

combine in phase comparison magic tee junctions *M* and *N*; the phase differences at these junctions differ by $\pi/2$ due to unequal feed paths. This phase difference, although frequency dependent, does not vary significantly over the receiver bandwidth. Square law detectors are attached to the free ports of each junction *M* and *N*; signals at these ports represent the vector sum and difference of the inputs to the junction. The pulse outputs *A'*, *B'*, *C'*, *D'* are subtracted to obtain outputs *A*, *B*, *C*, *D*, with amplitudes proportional to



$$\begin{aligned}
 A' &= a^2 + b^2 + 2ab \cos \varphi \\
 B' &= a^2 + b^2 - 2ab \cos \varphi \\
 C' &= a^2 + b^2 + 2ab \cos \left(\varphi + \frac{\pi}{2} \right) \\
 D' &= a^2 + b^2 - 2ab \cos \left(\varphi + \frac{\pi}{2} \right)
 \end{aligned}$$

$$\begin{aligned}
 \frac{A'}{B'} &\rightarrow A \rightarrow 4ab \cos \varphi \\
 \frac{A'}{D'} &\rightarrow B \rightarrow \frac{4}{\sqrt{2}} ab \cos \left(\varphi + \frac{\pi}{4} \right) \\
 \frac{D'}{C'} &\rightarrow C \rightarrow 4ab \sin \varphi \\
 \frac{A'}{C'} &\rightarrow D \rightarrow \frac{4}{\sqrt{2}} ab \sin \left(\varphi + \frac{\pi}{4} \right)
 \end{aligned}$$

Fig. 4. Phase discriminator. Two microwave signals X_1 and X_2 pass into magic tee junctions K and L where they are each split equally into two paths. Magic tee junctions M and N perform the phase comparison and signals A', B', C', D' are derived from square law detectors. These signals are subtracted to provide the required signals A, B, C, D .

$\sin \varphi, \sin (\varphi + \pi/4), \cos \varphi$ and $\cos (\varphi + \pi/4)$.

Fig. 5a shows how the four outputs vary with phase in the range $0 \leq \varphi \leq 2\pi$. It is seen that the four amplitudes pass through zero at phases which are multiples of $\pi/4$. The polarities of the four signals form a digital code which places the phase in one of eight $\pi/4$ intervals. The coding of the intervals, shown in fig. 5b, is obtained by designating positive amplitudes as "0" and negative amplitudes as "1". The signals are then amplified having regard only to pulse polarity; the amplifiers may saturate and need not be particularly linear or gain stable. The negative pulses from these amplifiers trigger bistable multivibrators thereby establishing the digital code in standard levels. These digits pass in parallel to the logic, along with two sets of digits from the other discriminators which represent 4φ and 16φ .

Logic

Let us suppose that each of the phases $\varphi, 4\varphi, 16\varphi$ were digitized into four $\pi/2$ intervals described by the binary coding 00, 01, 10, 11. The digits representing the three phases would appear as follows:

φ (radians)	φ	4φ	16φ
0	00	00	00
	00	00	01
	00	00	10
	00	00	11
$\pi/8$	00	01	00

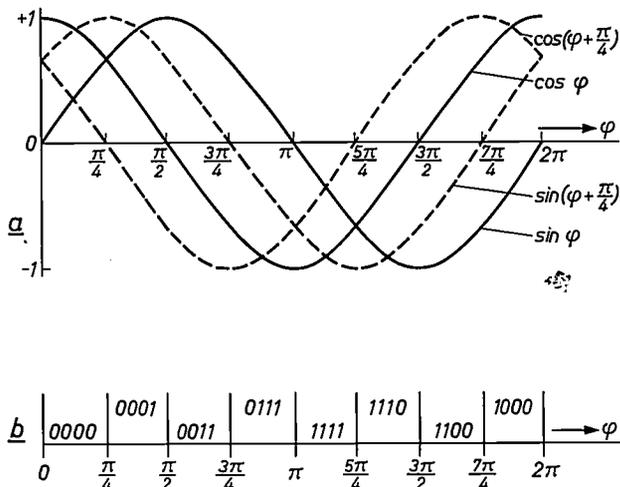


Fig. 5. Digitization of phase. a) shows how the four discriminator outputs vary with phase. The polarities of the four signals form a digital code which places the phase in one of eight $\pi/4$ intervals. b) shows the digitization intervals with their coding. The polarities of $\sin \varphi, \sin (\varphi + \pi/4), \cos \varphi, \cos (\varphi + \pi/4)$ form the code in which positive amplitudes are designated "0" and negative amplitudes "1". For example, in range $\pi < \varphi < 5\pi/4$ all the amplitudes are negative and the code is 1111.

It is clear by examination of the "4 φ " and "16 φ " digits that the two separate pairs of digits form a composite binary code; moreover the three pairs of digits form a six bit binary code in which numbers are proportional to φ . Thus the choice of 4 : 1 for the interferometer spacing ratio enables the measured phases to be in a form which may be readily converted to a binary number.

However, each discriminator has an independent phase measuring error with the result that the three pairs of digits will be slightly displaced with respect to

each other. In the above table, for example, 4φ will change from 00 to 01 at a slightly different bearing to the change of 16φ from 11 to 00 and an error will result. There is therefore a need to synchronize digital changes in the φ and 4φ codes with those in the 16φ code in order to provide an error free binary output.

The operation of the logic is as follows: the four digits from the "16 φ " discriminator are converted directly to the three least significant binary digits B_5, B_6, B_7 which act as reference digits. Table I shows the code from the discriminator alongside the binary output code. A simple logic may be designed from this table to perform the code conversion.

The most significant of the three reference digits, B_5 , is combined with the four digits from the "4 φ " discriminator to provide the next pair of binary digits, B_3 and B_4 which are synchronized to the reference digits. Table II shows the synchronization table from which the logic is designed. Briefly, the reference digit B_5 has fourfold ambiguity for a 2π change in 4φ . This

is resolved by reference to the 4φ discriminator code giving eight unique combinations which are labelled with the binary digits B_3 and B_4 (thus changes in B_3 and B_4 are synchronized to changes of B_5). In order to simplify the synchronization a phase shift of $\pi/8$ is introduced in one of the feed paths of the "4 φ " discriminator (omitted for clarity from fig. 3). Dashes in the 4φ code indicate regions of uncertainty near digital changes in the 4φ code. A maximum uncertainty of $\pm 22\frac{1}{2}^\circ$ is permitted in each discriminator before B_3 and B_4 become erroneous. In a similar fashion B_3 combines with the four φ digits to provide the two synchronized digits B_1 and B_2 .

Threshold

Operation of the system is dependent on a threshold channel in order to avoid errors on small signals which cannot trigger the bistable stores (fig. 3). An auxiliary horn feeds a detector and pulse amplifier and pulses of adequate amplitude trigger a set of pulse generators. These have four functions:

- a) The twelve amplified pulses from the discriminators are sampled and gated to the bistable stores for a short period (50 ns) when they have reached full amplitude. This eliminates front and rear edge spikes arising in the subtraction at the outputs of the discriminators.
- b) Gates on the seven outputs from the logic are opened when the logic has had time to operate.
- c) The bistable stores are reset when the output gate pulse closes.
- d) The sampling pulse is blanked for a few microseconds after a measurement to ensure that only one signal is processed at a time. This also eliminates multipath reflections.

Table I. Code conversion table. The discriminator code is derived from the outputs of the discriminator on the largest interferometer. The required binary outputs B_5, B_6, B_7 are placed beside the discriminator code. Logic to perform the code conversion may be designed from this table.

16 φ	Input: discriminator code				Output: binary code		
	$\sin 16\varphi$	$\sin\left(16\varphi + \frac{\pi}{4}\right)$	$\cos 16\varphi$	$\cos\left(16\varphi + \frac{\pi}{4}\right)$	B_5	B_6	B_7
0	0	0	0	0	0	0	0
	0	0	0	1	0	0	1
	0	0	1	0	0	1	0
	0	1	1	1	0	1	1
π	1	1	1	1	1	0	0
	1	1	1	0	1	0	1
	1	1	0	0	1	1	0
	1	0	0	0	1	1	1
2π							

Table II. Synchronization table. The phase shifted digital code from the discriminator on the middle interferometer resolves ambiguity in the binary reference digit B_5 . The required binary outputs B_3 and B_4 synchronize with B_5 . Dashes indicate uncertainty regions of $\pm 22\frac{1}{2}^\circ$ in the discriminator code. Synchronization logic may be designed from this table.

4 φ	Input discriminator code				binary reference B_5	Output synchron- ized binary code	
	$\sin\left(4\varphi + \frac{\pi}{8}\right)$	$\sin\left(4\varphi + \frac{3\pi}{8}\right)$	$\cos\left(4\varphi + \frac{\pi}{8}\right)$	$\cos\left(4\varphi + \frac{3\pi}{8}\right)$		B_3	B_4
0	—	0	0	0	0	0	0
	0	0	0	—	1	0	0
	0	0	—	1	0	0	1
	0	—	1	1	1	0	1
π	—	1	1	1	0	1	0
	1	1	1	—	1	1	0
	1	1	—	0	0	1	1
	1	—	0	0	1	1	1
2π							

Performance

The experimental receiver has a sensitivity of $6 \mu\text{W}$ and accuracy of $10'$ over 20° field of view. The per-

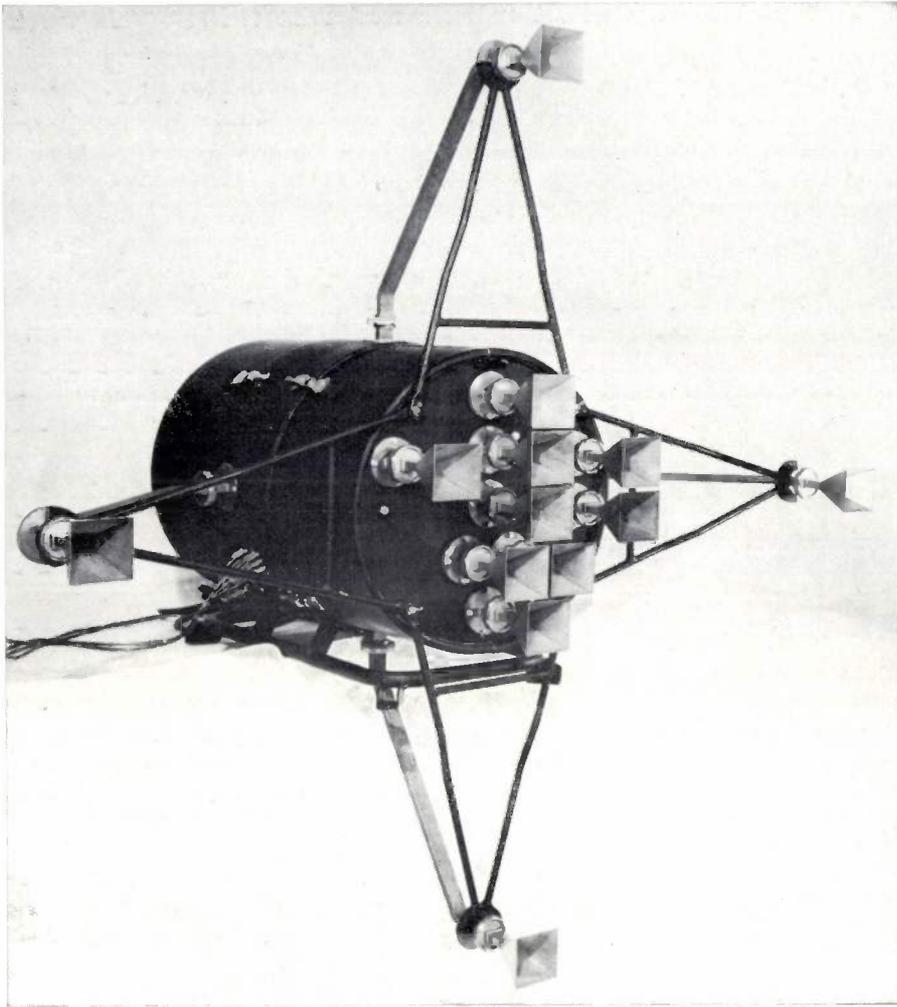


Fig. 6. Experimental model of direction finder. Two orthogonal interferometer arrays are integrated with the processing circuits in the cylindrical unit.

mitted phase measuring uncertainty of $\pm 22\frac{1}{2}^\circ$ covers discriminator errors and wavefront distortions resulting from reflections. A model has been built with two orthogonal interferometers which measure azimuth and elevation for blind landing applications (*fig. 6*). The spacing of the largest interferometer is 132 cm and the rest of the system, including horns, is contained in a cylinder of diameter 46 cm and length 77 cm. A complete system operating at 15 GHz with 20' accuracy is expected to measure 35 cm \times 35 cm \times 12 cm.

The system may be extended to give greater accuracy or field of view by increasing the number of interferometers while retaining the 4 : 1 spacing ratio. By reducing the spacing ratio to 2 : 1 and changing the logic, the permitted phase uncertainty may be increased

to $\pm 67\frac{1}{2}^\circ$ in order to cope with propagation errors in poor sites. The system which has been described is one of a family of systems where the spacing ratio may be any integer or fraction. Spacing ratios of the form 2^K where K is an integer lead to simple logic and binary coded outputs^[1]. The system sensitivity might be increased by the use of superheterodyne techniques.

Summary. An accurate microwave direction-finding receiver with digital output is described. The receiver has application in the blind landing of aircraft and in navigation systems in rivers and estuaries. The direction of a microwave transmitter may be measured with high accuracy over a wide field of view by the use of several interferometers with different aerial spacings. Bearing ambiguity in a large interferometer may be resolved by reference to a smaller interferometer. A simple, fast digital method of ambiguity resolution is described in which phase differences are derived digitally and are processed to provide a measure of signal bearing in binary code. An experimental system which operates on single pulsed signals at 9.5 GHz using three pairs of receiving aerials with a 4 : 1 spacing ratio, digitizes bearing to 10' over a 20° field of view. The field of view and resolution may be extended using more interferometers.

[1] R. N. Alcock, P. W. East and S. J. Robinson, A digital method of ambiguity resolution in multiple interferometers, to be published.

Pyrolytic graphite

W. F. Knippenberg, B. Lersmacher, H. Lydtin and A. W. Moore

Now that the initial phase of investigations on pyrolytic graphite appears to be over, it seems to us appropriate to devote an article to this material, describing the present state of methods of preparation and applications. Pyrolytic graphite has various interesting properties, and it has already found useful application in a wide variety of fields.

Introduction

The formation of carbon deposits on substrates heated in carbon-containing gases has long been known. The present demand for refractory protective coatings on materials has however speeded up the investigations of carbon deposits, so that a great deal more knowledge has been gained about them, particularly in recent years.

Carbon layers deposited from a carbon-containing gas (e.g. a hydrocarbon) at temperatures above 2000 °K and at pressures below a few tens of torr, show under X-ray analysis a well-crystallized structure, which has much in common with that of a graphite crystal. Such carbon layers are known for this reason as *pyrolytic graphite*. Going towards lower deposition temperatures the long-range order decreases. The deposit can then be regarded as a stacked structure of small graphitic crystals, the mode of stacking being affected by the temperature. The properties of layers produced at about 1000 °C can best be explained if one assumes the presence of linear carbon polymers in addition to graphitic areas containing two-dimensional ring systems of carbon atoms. In common parlance a distinction is made between low-temperature *carbon* layers and high-temperature *graphite* layers. In actual fact, however, there is a continuous transition in crystallite size.

The properties of graphite are highly anisotropic.

The extent to which this anisotropy is apparent depends on the structure of the pyrolytic carbon layers. It is found most strongly in pyrolytic *graphite*. Particularly good use can be made of the strong anisotropy of the thermal and electrical conductivities. Pyrolytic graphite layers also show very great chemical inertia and heat resistance. Pyrolytic graphite is useful not only in high temperature applications but also, owing to the high degree of purity it can be given, in semiconductor technology.

Unlike pyrolytic graphite, normal graphite, also known as electrographite, is in general isotropic, polycrystalline, porous and therefore much more reactive (*fig. 1*). It is also much more difficult to obtain in a

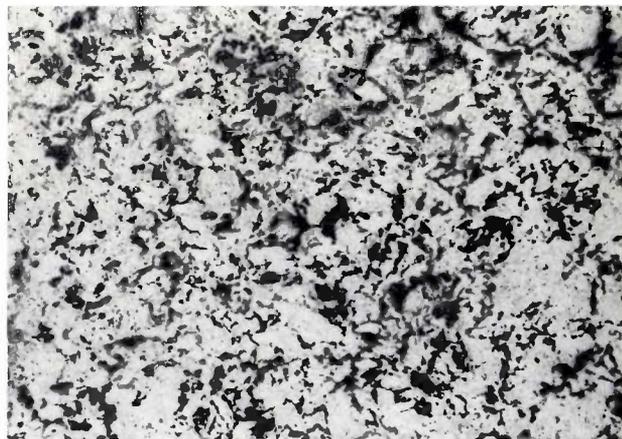


Fig. 1. Photomicrograph of a polished section through electrographite. The porous polycrystalline structure is clearly visible. Magnification 120 ×.

Dr. W. F. Knippenberg is with Philips Research Laboratories, Eindhoven; Dr. B. Lersmacher and Dr. H. Lydtin are with the Aachen laboratory of Philips Zentrallaboratorium GmbH; Dr. A. W. Moore was formerly with Philips Research Laboratories, Eindhoven.

highly pure form. The basic materials for graphite of this kind (e.g. coal or charcoal) are subjected to the successive processes of milling, mixing, forming, pressing, baking and finally to electrographitization, a process of resistance heating at about 2500 °C for several days. For many applications (e.g. for crucibles) it is most advantageous to coat an object made of electrographite with a layer of pyrolytic graphite.

The long-range order of pyrolytic graphite can be increased by subjecting a piece of the material deposited at high temperature to a further heat treatment under mechanical pressure. With the aid of this method it has proved possible to obtain a piece of pyrolytic graphite, several cubic centimetres in volume, in a form in which its properties are practically the same as those of a graphite monocrystal.

The methods of preparing pyrolytic graphite are generally developed empirically. An initial step towards a theoretical approach is the thermodynamical treatment of the temperature-dependence of the "solubility" of carbon in a gas of specific composition. These calculations have now been completed for the principal gas systems containing carbon [1]. The present article will give the results of these calculations for some of these systems. In the thermodynamic treatment an attempt will also be made to provide a systematic basis for the empirical preparation methods. Many aspects of the deposition process are however still insufficiently known. There is a need for further study of the way in which the carbon is deposited from the gas phase and also of the nucleation in the gas and on the substrate, and the growth of these nuclei. The strong pressure and temperature dependence of these processes is responsible for the marked differences mentioned above between the properties of carbon layers formed at various pressures and temperatures.

In view of the close resemblance between pyrolytic graphite layers and a graphite crystal, it will be useful to examine some of the properties of the ideal graphite crystal before discussing pyrolytic graphite and its preparation.

The ideal graphite crystal

The ideal graphite crystal has a layered structure; this structure can clearly be seen in *fig. 2*. Strong bonding exists between the carbon atoms in each layer. The bonding between the successive layers is much weaker. Because of this the layers are easily displaced relative to one another and the graphite crystal can easily be cleaved parallel to these layers. It can also be seen from *fig. 2* that in a graphite crystal each third layer is identical with the first one.

An important consequence of the layered structure of graphite is the pronounced anisotropy of the phys-

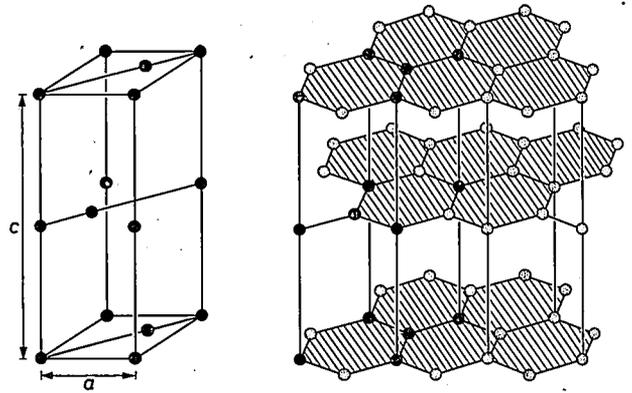


Fig. 2. The graphite lattice. On the left, the unit cell; $a = 0.246$ nm, $c = 0.669$ nm. On the right, part of a graphite lattice arranged so that the layer structure can clearly be seen. The carbon atoms are at the nodes of the hexagonal networks.

ical properties which we encountered earlier. This applies in particular to the electrical and thermal conductivities. Using a subscript a for conductivity parallel to the layers and a subscript c for conductivity perpendicular to the layers (parallel to the crystallographic c -axis) then for the thermal conductivity:

$$\lambda_c/\lambda_a = 0.01,$$

and for the electrical conductivity:

$$\sigma_c/\sigma_a = 0.001.$$

It should be added that the graphite crystal is a relatively good electrical conductor even in the direction of the c -axis ($\rho \approx 0.01$ - $1 \Omega\text{cm}$), while it can be described as a good thermal insulator in this direction. The thermal conductivity along the c -axis is about a factor of 10 lower than that of Al_2O_3 . The anisotropy is also clearly to be seen in the thermal expansion. Linear thermal expansion takes place in the c -direction, whereas in the a -direction there is a slight contraction up to about 400 °C. Above this temperature the ratio of the thermal expansion coefficients (L_c and L_a) [2] is:

$$L_c/L_a \approx 25 - 30.$$

The anisotropy is also clearly seen when the crystal is attacked by chemicals, and in evaporation: the faces perpendicular to the c -axis are attacked more slowly than those perpendicular to the a -axis directions. Similar behaviour is found when carbon is sputtered by bombardment with ions or electrons.

[1] B. Lersmacher, H. Lydtin, W. F. Knippenberg and A. W. Moore, to be published shortly in *Carbon*.

[2] A. R. Ubbelohde and F. A. Lewis, *Graphite and its crystal compounds*, Clarendon Press, Oxford 1960.

[3] M. Pirani and W. Fehse, *Z. Elektrochemie* 29, 168, 1923.

[4] J. Gibson, M. Holohan and H. L. Riley, *J. Chem. Soc.* 1946, p. 456.

Pyrolytic graphite

It is obvious that a material with properties like those of a graphite crystal can have important applications. Large ideal crystals of graphite are seldom found in nature, however. It was therefore a very important discovery that in pyrolytic graphite the graphitic crystallites are mainly oriented with their c -axes at right angles to the surface of the substrate, i.e. with the layers of rings of six carbon atoms lying parallel to the surface of the substrate [3]. A carbon deposit of this kind looks like a metal.

The difference compared with an ideal graphite crystal is illustrated in *fig. 3*. The networks represent a diagrammatic projection in the direction of the c -axis of the ideal graphite lattice (*fig. 3a*) and the pyrolytic graphite lattice (*fig. 3b*). As a result of the identical nature of stacking layers in ideal graphite which we mentioned earlier, all that can be seen in *fig. 3a* is a lattice plane n and the adjacent planes $n + 1$ or $n - 1$.

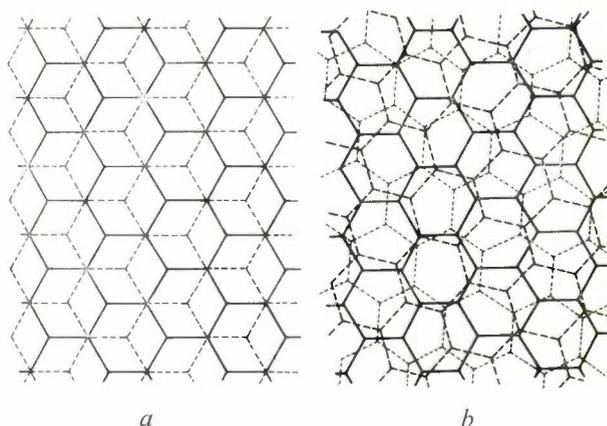


Fig. 3. a) Projection drawing of a few layers of an ideal graphite lattice. The carbon atoms are situated at the nodes of the hexagonal networks. *b)* Projection drawing of the atomic lattice of pyrolytic graphite as produced by deposition. This lattice can be changed to an ideal graphite lattice by combined thermal and mechanical after-treatment.



Fig. 4. Photomicrograph of a polished section through a pyrolytic graphite layer parallel to the direction of growth. The substrate is parallel to the lower edge of the photograph. Magnification $80\times$.

The projection of the pyrolytic graphite lattice demonstrates in this case that the successive layers are statistically twisted relative to one another. A slight spread is also usually to be found in the directions of the c -axes of adjacent crystalline areas. They lie mainly, however, in the direction normal to the substrate [4]. As a result of this divergence of c -axes and the consequent divergence of the direction of growth, a layer of pyrolytic graphite consists of whiskery cones, as shown in *fig. 4*.

All details of the photographs lie in the same plane. The light and dark areas are due to differences in the reflection of the polarized light from the pyrolytic graphite cones, which have been cut at different angles in the polishing procedure. *Fig. 5a* shows a similar polished section through a pyrolytic graphite layer with a fine-whiskered cone structure.

The layers that can be seen are examples of deposits in which the nucleation has mainly taken place on the

surface of the substrate. The coarse or fine-whiskered structure is a direct result of the structure of the substrate surface^[5]. Nucleation can however also take place at a later stage. An example is shown in fig. 5*b*. The conical growth is due to a perturbation, which may for example be the incorporation of a particle of soot, giving rise to further nucleation.

Preparation of pyrolytic graphite

In order to deposit a homogeneous layer of pyrolytic graphite upon a substrate, which may be an object of arbitrary shape, it is necessary to ensure that the tem-

perature is uniform over the entire surface and is higher than 2000 °C. The flow rate of the carbon-containing gas should be such that the concentration of carbon in the gas can be regarded as uniform above the surface. A low gas pressure is usually employed so that no graphite forms in the gas; "soot" formation can give rise to irregularities in the layer, as was shown in fig. 5*b*.

There are two different methods of heating the substrate. In one method the substrate (and the deposit already formed on it) is heated by an electric current passed through it or by induction heating, with the substrate set up freely radiating in a stream of carbon-containing gas (the "cold wall" process). In the other method the substrate is entirely surrounded by a source of heat, so that the surface is heated ("hot wall" process).

One of the first systematic experimental investigations of the formation of pyrolytic graphite was carried out by Pirani and Fehse in 1923^[3]. They studied the dissociation of carbon compounds on thin carbon filaments that were heated by current conduction up to about 2000 °C. Thirty years later, Brown, Hall and Watt used a similar method to obtain thicker deposits on larger substrates^[6]. They used dissociation

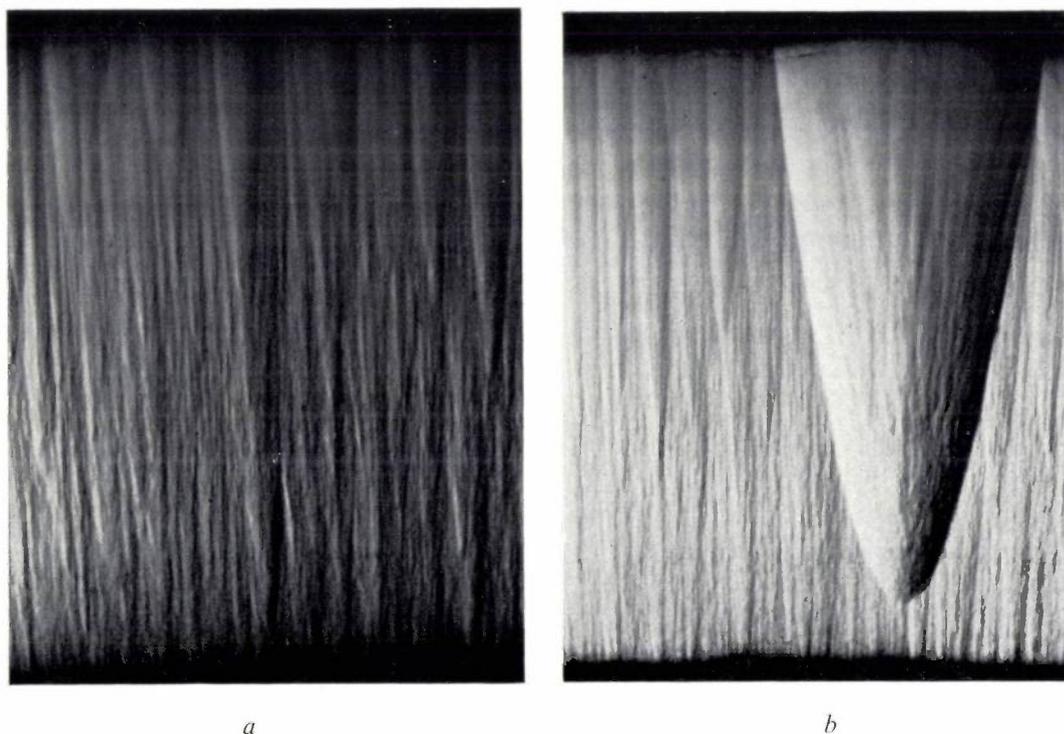


Fig. 5. *a*) Photomicrograph of a polished section through a piece of pyrolytic graphite of fine whiskery structure, parallel to the direction of growth. The substrate is parallel to the lower edge of the photograph. Magnification 110 × .
b) Polished section similar to that in (*a*); the inclusion of an impurity has given rise to fresh nucleation. Magnification 110 × .

perature is uniform over the entire surface and is higher than 2000 °C. The flow rate of the carbon-containing gas should be such that the concentration of carbon in the gas can be regarded as uniform above the surface. A low gas pressure is usually employed so that no graphite forms in the gas; "soot" formation can give rise to irregularities in the layer, as was shown in fig. 5*b*.

There are two different methods of heating the substrate. In one method the substrate (and the deposit already formed on it) is heated by an electric current passed through it or by induction heating, with the substrate set up freely radiating in a stream of carbon-

of methane and propane on graphite objects at a pressure of 10-15 torr. The substrates were resistively heated to a surface temperature between 1600 and 2100 °C in a water-cooled vacuum vessel.

Subsequent investigations indicated that, owing to the low thermal conductivity of the deposit in the growth direction (along the *c*-axis) and the heat losses due to surface radiation, very high substrate temperatures are needed in this method to keep the outer surface of the deposit at the appropriate temperature. Because of this, the first deposited layers already show strong recrystallization^[7]. This process is not suitable

for producing well-defined homogeneous material. Moreover the maximum thickness obtainable in this way is only about 5-6 mm since, owing to the high temperature which the substrate must have, the layers first deposited evaporate at this thickness. Thermal gradients can be avoided by high-frequency induction heating. High power is however required even for quite small objects, and this therefore limits the use of this type of heating. Large, thick and homogeneous layers of pyrolytic graphite are produced by the second method mentioned, in which the substrates are kept at the required temperature by radiation from an external source. This is the method used in the equipment described in this article.

Recrystallization of pyrolytic graphite

For fundamental research on graphite and its compounds it is desirable to have fairly large graphite crystals to work on. As we have seen, these are seldom found in nature. Nowadays, however, the thermal after-treatment of pyrolytic graphite can yield material which is practically equivalent to a graphite crystal, and which serves in measurements as a good substitute for the ideal crystal. A difficulty in this recrystallization process is that the anisotropic thermal expansion tends to prevent the corresponding layers in the adjacent crystallites from moving into the same plane. Various stages can be distinguished in the recrystallization process [8]. Upon heating above 3000 °C the number of layers whose position is twisted relative to that of the ideal graphite crystal declines considerably, and the average crystallite thickness increases by a factor of 100. At temperatures above 3500 °C the deposits recrystallize to form almost ideal graphite crystals [9].

Combined tensile stress and heating procedures are more effective for graphitization than heating alone. If a tensile stress is applied during the heat treatment, it is easier to bring the corresponding layers from neighbouring crystallites into one plane. Quite often, however, this results in fracturing of the layers and formation of pores [10]. The obvious alternative is to heat the pyrolytic graphite while at the same time applying pressure in the direction of the *c*-axis. Using this method, by heating a piece of pyrolytic graphite at temperatures between 2800 and 3000 °C while applying a unilateral pressure of 300-500 bar, pieces of graphite could be produced which were more than 1 cm thick in the direction of the *c*-axis, and had a density of 99.95% of the theoretical value [7]. Pyrolytic graphite of this kind, after hot-pressing, can very easily be cleaved along the basal planes, which then exhibit a very high metallic reflection due to the almost perfect orientation of the crystallites (*fig. 6*). Further tempering of such material between 3400 and 3500 °C,

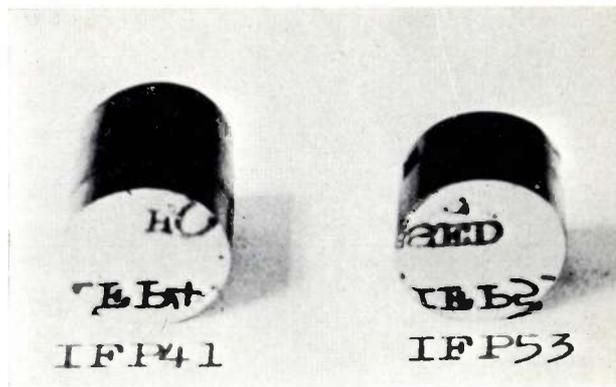


Fig. 6. Pieces of a pyrolytic graphite cylinder cleaved along the basal planes, after hot pressing.

under slight pressure, gives practically ideal graphite.

Before we consider the preparation technique for pyrolytic graphite, we shall first deal in more detail with the gas used in the preparation. Not all carbon-containing gases are equally suitable for this process. A thermodynamic treatment of the deposition of carbon from different gases can provide a general picture of the usefulness of these gases, and also give a better understanding of the empirical methods of preparing pyrolytic graphite.

Carbon deposition from a gas

When a carbon layer begins to grow upon a substrate heated in a carbon-containing gas, it shows that at a higher temperature the gas cannot contain as much carbon as it does at a lower temperature. The solubility of carbon in the gas can be said to have been reduced by the increase of temperature. The form in which the carbon occurs in the gas at different temperatures is irrelevant here. It may occur as atomic carbon, as carbon molecules (C_2 , C_3) or as an element in other molecules, for example CO , C_2H_2 , etc. As a measure of the solubility the ratio of the number of carbon atoms in the gas to the number of all other atoms present in the gas is therefore used.

In practice hydrocarbons are generally employed for the gas system. Among the reasons given for this choice are ease of manipulation, the availability of hydrocarbons on an industrial scale, cheapness, etc. Other

- [5] E. R. Stover, G. E. C. Res. Rep. No. 62-RL-2991 M, 1962.
 [6] A. R. G. Brown, A. R. Hall and W. Watt, *Nature* **172**, 1145, 1953.
 [7] A. W. Moore, A. R. Ubbelohde and D. A. Young, *Proc. Roy. Soc. A* **280**, 153, 1964.
 [8] E. R. Stover, G. E. C. Res. Rep. No. 60-RL-2564 M, 1960.
 [9] C. A. Klein, W. D. Straub and R. J. Diefendorf, *Phys. Rev.* **125**, 468, 1962.
 [10] H. E. Martens and W. V. Kotlensky, Jet Propulsion Laboratory, California Institute of Technology, Tech. Rep. No. 32-71, 10th March 1961; *Proc. 5th Carbon Conf.*, Vol. 2, 625, 1963.

systems have been virtually ignored. It takes a great deal of time to investigate experimentally the suitability of a system for carbon deposition, and moreover the number of systems that would have to be considered for such investigation is rather considerable. It certainly saves time if the fundamental suitability of a system is verified by thermodynamic calculations before starting an empirical investigation, although the calculations involved may be quite extensive. Of course, only thermodynamic possibilities can be indicated, and a thermodynamic possibility may sometimes not be a practical one because the establishment of chemical equilibrium is kinetically blocked.

Table I shows the carbon-containing gas systems which may be of interest. The first one is the unary system C. This is of interest when considering the preparation of carbon deposits by the evaporation of carbon at high temperatures and condensation at lower temperatures. It is followed in the table by a few binary systems, with the carbon compounds occurring in the gas. The combination of two binary systems then leads to the ternary systems that follow. One can continue in this way composing higher systems from combinations of lower ones.

In order to make an exact calculation of the gas composition it is necessary to know all the types of molecule contained in the gas, together with their thermodynamic properties. The calculations are simplified to some extent by the fact that the temperatures of interest for carbon deposition are all above 1000 °C. It would be a practical impossibility to make an exact calculation of gas compositions below this temperature, because of the large number of organic compounds that would have to be included in the calculation. At temperatures above 1000 °C most of these compounds are however no longer thermodynamically stable, so that the calculation is considerably simplified by reduction of the number of types of molecule contained in the gas.

We shall first deal with the unary system C. The evaporation and deposition of carbon in a vacuum can be regarded as a border-line case of the deposition of carbon from a gas. Using the carbon vapour pressures given by Kuthe [11] and the evaporation coefficient ($\alpha = 0.15$) determined by Thorn and Winslow for graphite [12], we can calculate the total evaporation flow rate (i_{tot}) of carbon atoms as a function of temperature. At 2500 °K it is 7.82×10^{16} C atoms/cm²s and at 3000 °K it is 8.1×10^{18} C atoms/cm²s. The rates of growth of a layer obtained by condensation from carbon evaporated at these temperatures are 24.8 μ m/hour and 2570 μ m/hour respectively. These are growth rates similar to those encountered when preparing pyrolytic graphite with the aid of hydrocarbons. The

Table I. Systems relevant to carbon transport.

	System	Molecular types in the gas phase
Unary	C	C ₁ ; C ₂ ; C ₃ ; ... (C _{β})
Binary	C-H	C _{β} ; CH; CH ₂ ; CH ₃ ; C ₂ H; ... aliphatic, cyclic, aromatic hydrocarbons
	C-N	C _{β} ; CN; (CN) ₂ ; CN ₄ ; ...
	C-O	C _{β} ; CO; CO ₂ ; C ₃ O ₂ ; ...
	C-S	C _{β} ; CS; CS ₂ ; ...
	C-F	C _{β} ; CF ₄ ; ...
	C-Cl	C _{β} ; CCl ₄ ; C ₂ Cl ₂ ; C ₂ Cl ₄ ; C ₂ Cl ₆ ; ...
	C-Br	C _{β} ; CBr ₄ ; ...
Ternary	C-H-N	molecular types of the relevant binary systems, but also amines, etc.
	C-O-H	molecular types of the relevant binary systems, but also ethers, aldehydes, ketones, etc.
	C-S-H	molecular types of the relevant binary systems, but also thio-compounds, etc.
	C-O-S	molecular types of the relevant binary systems and COS

evaporation of carbon in a vacuum involves more experimental difficulties than the use of a gas system such as a hydrocarbon. Moreover, this method can only be used for covering substrates of simple shape with carbon. Its practical value is therefore limited.

The gas phase of the system C-O contains mainly CO, CO₂, O, O₂ and carbon in the molecular types C₁, C₂ and C₃. It follows from the phase rule that the system has two degrees of freedom, that is to say the composition of the gas phase is uniquely defined at a given temperature and pressure. The procedure for calculating the gas composition is given in Table II (see p. 238). An account of the thermodynamic quantities employed is given in an article published elsewhere [1].

A result of this calculation is presented in fig. 7, where the partial pressures of the respective molecular species are shown as a function of absolute temperature for a total pressure of 1 bar. At low temperatures the gas phase contains principally CO₂ molecules. At temperatures above 1000 °K it is virtually only CO that is stable. The dissociation of CO into C and O does not begin until the temperature has reached 3000 °K and above. The partial pressures of atomic and molecular oxygen at 1 bar are negligible in the temperature interval concerned. In fig. 8 the "solubility" of carbon in the gas phase is plotted against temperature for various total pressures. The solubility of the carbon in the system C-O is found to be strongly temperature-dependent between 700 and 1200 °K. In this temperature range a gas which is saturated with carbon at high temperature (i.e. in equilibrium with solid carbon)

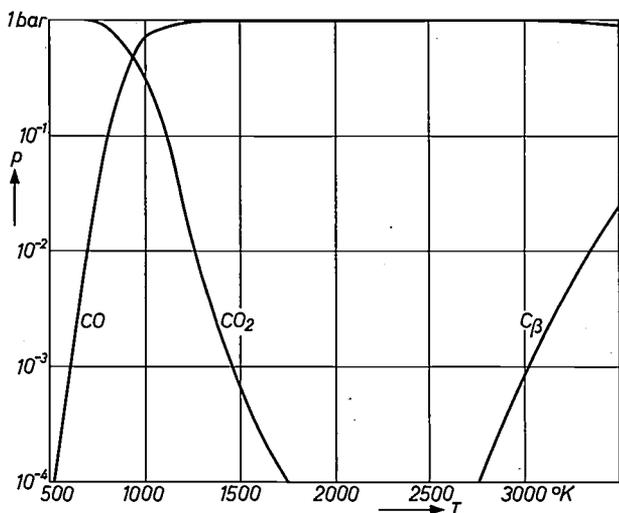


Fig. 7. Gas composition in the system C-O as a function of temperature for a total pressure of 1 bar.

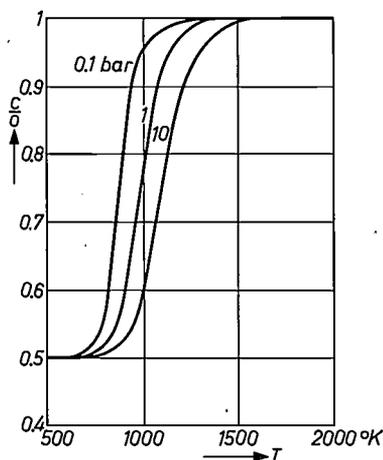


Fig. 8. The solubility of carbon in the system C-O, expressed by the ratio C/O, as a function of temperature for various total pressures.

should be possible via this gas phase from a "hot" to a "cold" substrate. This temperature region is the one in which, as already mentioned, pyrolytic graphite is formed.

To illustrate the method of calculating the effect of a third component we can take the system C-O-S. It was assumed here that the gas contained the molecular types C₁, C₂, C₃, CO, CO₂, S₁, S₂, CS, CS₂, COS, SO, SO₂, SO₃, O₁ and O₂ [1]. Fig. 11 shows the calculated partial pressures as a function of temperature for a total pressure of 1 bar, and for a ratio S/O = 1. The oxygen present in the system is almost completely bound in CO molecules. The contribution of the CO₂,

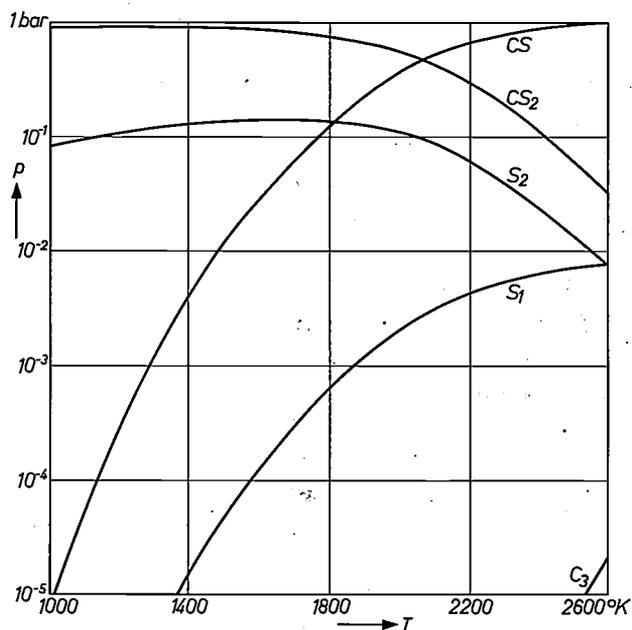


Fig. 9. Gas composition in the system C-S as a function of temperature for a total pressure of 1 bar.

would give up carbon at a lower temperature. There would therefore have to be a flow of carbon from "hot" to "cold". In reality, however, this does not occur. The conversion of CO is therefore blocked. No such blocking is found in the comparable system C-S, and therefore the results of the calculations for this system prove to be of practical importance.

For the system C-S it was assumed that the molecular types C₁, C₂, C₃, S₁, S₂, CS and CS₂ (fig. 9) [1] are present in the gas phase. The results of the calculation are presented in fig. 10, showing the solubility of carbon in the gas phase as a function of temperature for various pressures. According to the calculation, at temperatures in the region of 2000 °C carbon transport

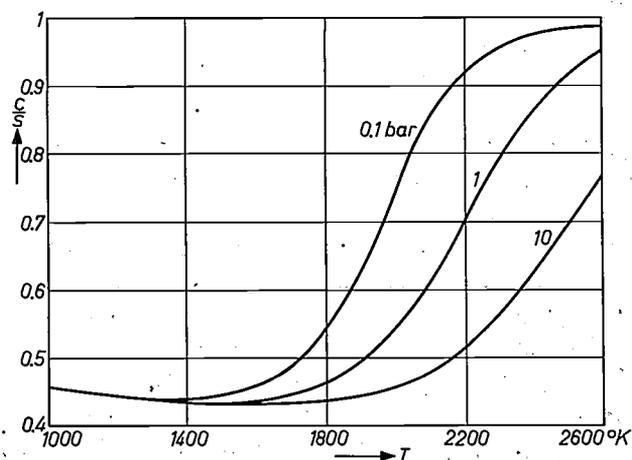


Fig. 10. The solubility of carbon in the system C-S, expressed by the ratio C/S, as a function of temperature for various total pressures.

[1] R. Kuthe, Interne Mitteilung des Instituts für Chemische Technologie der Technischen Hochschule Braunschweig.

[2] R. J. Thorn and G. H. Winslow, J. chem. Phys. 26, 186, 1957.

Table II. Procedure for calculating the equilibrium composition in the system C-O over solid carbon.

a) Components (<i>n</i>):	$C_\beta, O_1, O_2, CO, CO_2$ $p_{C_\beta} = p_{C_1} + p_{C_2} + p_{C_3}$
b) Phases (<i>q</i>):	solid; gas
c) Reactions (linearly independent) (<i>r</i>):	$C \rightleftharpoons C_\beta; \quad K_1 = p_{C_\beta}$ $2O_1 \rightleftharpoons O_2; \quad K_2 = p_{O_2}/p_{O_1}^2$ $C + \frac{1}{2}O_2 \rightleftharpoons CO; \quad K_3 = p_{CO}/p_{O_2}^{1/2}$ $C + O_2 \rightleftharpoons CO_2; \quad K_4 = p_{CO_2}/p_{O_2}$
d) Degrees of freedom (<i>f</i>):	$f = n - q + 2 - r = 2$
e) Equilibrium data:	$\ln K_p = -\frac{\Delta G^0}{RT}$ $\log_{10} K_1 = +7.334 - 31.228 \frac{10^3}{T}$ $\log_{10} K_2 = -2.135 + 25.734 \frac{10^3}{T} - 1.420 \log_{10} T$ $\quad \quad \quad + 0.111 \cdot 10^3 T - 0.008 \frac{10^6}{T^2}$ $\log_{10} K_3 = +6.293 + 5.548 \frac{10^3}{T} - 0.448 \log_{10} T$ $\quad \quad \quad - 0.059 \cdot 10^3 T + 0.024 \frac{10^6}{T^2}$ $\log_{10} K_4 = +1.245 + 20.483 \frac{10^3}{T} - 0.357 \log_{10} T$ $\quad \quad \quad + 0.016 \cdot 10^3 T + 0.005 \frac{10^6}{T^2}$
f) Balance equation:	$p_{tot} = p_{C_\beta} + p_O + p_{O_2} + p_{CO} + p_{CO_2}$ $= K_1 + p_{O_2}^{1/2} K_2^{-1/2} + p_{O_2} + p_{O_2}^{1/2} K_3 + p_{O_2} K_4$ $= K_1 + p_{O_2}^{1/2} (K_2^{-1/2} + K_3) + p_{O_2} (1 + K_4)$

COS, SO, SO₂ and SO₃ molecules to the oxygen content is negligible. At lower temperatures the sulphur occurs mainly in the CS₂ molecules and to a lesser extent in S₂. With rising temperature the CS content of the gas increases. Above 2200 °K sulphur is found almost only in CS molecules. Fig. 12 shows the solubility of carbon in the system as a function of temperature at a total pressure of 1 bar for various S/O ratios. It is found that with increasing oxygen content in the gas the carbon solubility becomes increasingly less temperature-dependent. The addition of oxygen to the system must therefore have an unfavourable effect on carbon transport.

As a final example we shall consider the system C-H. Although there was no longer any question about the practical usefulness of this system, the thermodynamic treatment was nevertheless found to reveal entirely new and interesting aspects. It was assumed that the system C-H contained the molecular species C₁, C₂, C₃, H₁,

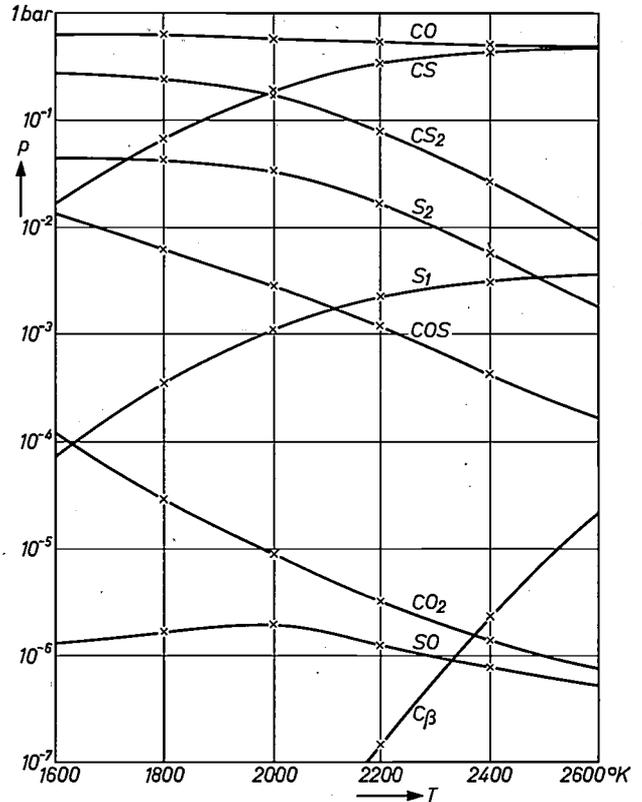


Fig. 11. Gas composition in the system C-O-S as a function of temperature for a total pressure of 1 bar and a ratio S/O = 1.

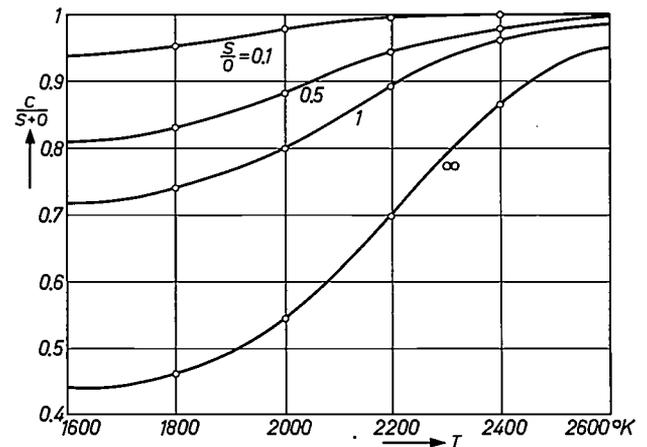


Fig. 12. The solubility of carbon in the system C-O-S expressed by the ratio $C/(S + O)$, as a function of temperature for a total pressure of 1 bar.

H₂, CH, CH₂, CH₃, CH₄, C₂H, C₂H₂, C₂H₄, H⁺, C⁺ and electrons in the gas. Fig. 13, which is due to Kuthe [11], shows the partial pressures of the molecular types under consideration as a function of temperature at a total pressure of 1 bar. As can be seen from this figure, the main carbon-carrying molecule up to 1500 °K is CH₄. Above 2000 °K the formation of C₂H₂ becomes particularly important. The contribution of the

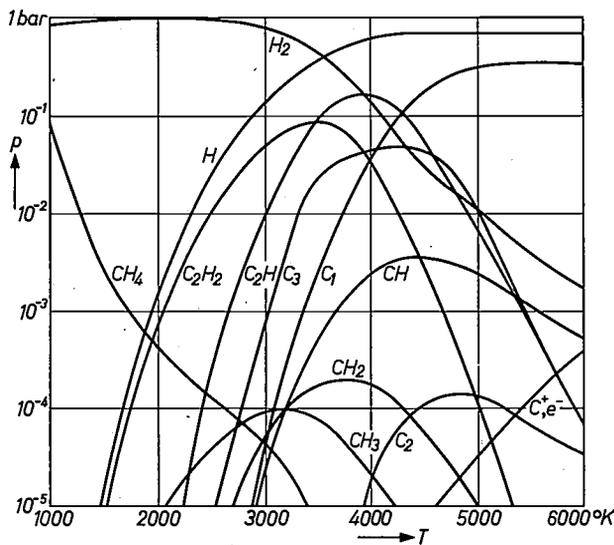


Fig. 13. Gas composition in the system C-H as a function of temperature for a total pressure of 1 bar.

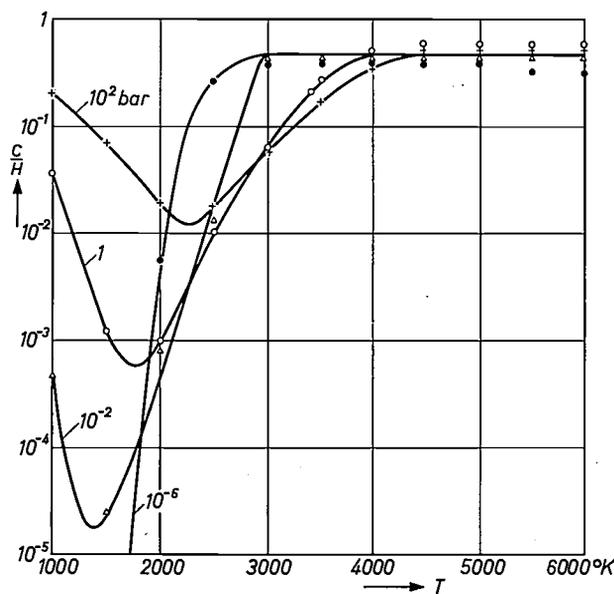


Fig. 14. The solubility of carbon in the system C-H, expressed by the ratio C/H , as a function of temperature for various total pressures and a ratio $C/H = \frac{1}{2}$ in the homogeneous region.

C_2H molecule and of atomic and molecular carbon to the carbon content of the gas phase only becomes noticeable above 3000 °K. The solubility of carbon in the gas is shown in fig. 14 as a function of temperature for various total pressures. The solubility curves are seen to have a pronounced minimum which shifts towards higher temperatures as the total pressures increase. If we now assume that all the carbon coming from the gas is deposited on the substrate, then at a given pressure there is a maximum in the curve of deposition rate

as a function of temperature, and this maximum corresponds to the minimum in the solubility curve. Carbon deposition can be obtained both by conducting the gas from "cold" to "hot" (to the left of the minimum) and from "hot" to "cold" (to the right of the minimum). Pyrolysis can be characterized by the substrate-to-gas temperature gradient. One can use the term "hot-gas pyrolysis" if the gas has a higher temperature than the substrate, and the term "cold-gas pyrolysis" when the gas has a lower temperature than the substrate. Both possibilities have now been realized.

The usefulness of calculating the effect of a third component appeared from a calculation of the effect of the addition of sulphur to the system C-H. The addition of sulphur gives this gas system a greater capability for carbon transport.

In the foregoing we have already attempted to draw a conclusion about the rate of carbon deposition from a gas. This deposition rate, which is a function of the total gas pressure and of the temperature, is however also affected by the specific properties of the carbon-carrying gas molecules. The problems connected with this have not yet been very widely studied, either experimentally or theoretically. Theoretical studies have been carried out only for the very low pressures region. At high pressures, apart from growth on the substrate, there is also some nucleation and growth of carbon particles in the gas (soot formation). When making carbon layers it is usual to work in a pressure range in which (visible) soot formation is just avoided. In growing layers on the substrate, factors which have to be taken into account include the nucleation on the substrate, the incorporation in the layer of nuclei that have formed in the gas phase or the dissolving of these nuclei, and the recrystallization of the grown layers. These effects are particularly important in the initial stage of growth, which means that small variations, for example in the surface properties of the substrate, can give rise to considerable differences in deposition. After this initial stage, in which the growth may be very irregular, the thickness of the layer increases almost linearly with time.

Preparation technique

Fig. 15 shows a cross-section of a furnace used for producing pyrolytic graphite. A high temperature has to be generated in this furnace in a volume large enough to contain the whole substrate to be coated with a carbon layer. This is done by means of a graphite heater element in the form of a cylinder H which encompasses the whole working space. The heat is produced by passing an electric current through the element by means of watercooled electrodes E at the base of the vacuum vessel. The electrodes are water-cooled (C).

The heater element has a cross-section of about 20 cm and is about 50 cm long. Inside this element there is a graphite tube *I* which serves to prevent carbon deposits on the element. The substrate *S* is attached to the lid *L* by means of a rod *R*. The outer tube *O* is a further shield for the heater element. Heat insulation is provided by several layers of graphite felt (shown grey in the figure). The carbon-containing gas is admitted through a water-cooled tube *G*. The residual gases are exhausted from the system through the pump aperture *P*. The temperature of the object to be coated is measured by optical pyrometry through a slit in the heater element and the window *W* in the water-cooled wall of the furnace. Fig. 16 shows the layout of the furnace.

To prepare pyrolytic graphite the furnace is evacuated and brought to a temperature of 2200 °C. A carbon-containing gas, e.g. propane, is then admitted until the pressure in the furnace is about 5 torr. This pressure is maintained by adjusting the pumping speed appropriately (1.5 l/min at s.t.p.). The furnace described can operate continuously for longer than 50 hours without any risk of stoppages. The rate of deposition in these conditions is about 0.02 cm/hour, making it possible to produce layers more than 1 cm thick in a single experimental run. A power of 35 to 40 kW is needed to maintain the high temperature. It takes about 18 hours to coat the crucibles illustrated in fig. 17.

Particularly with thick deposits, differences between the expansion coefficients of the layer and of the substrate cause the pyrolytic graphite to break away from the substrate when it cools from the deposition temperature to room temperature. On the other hand,

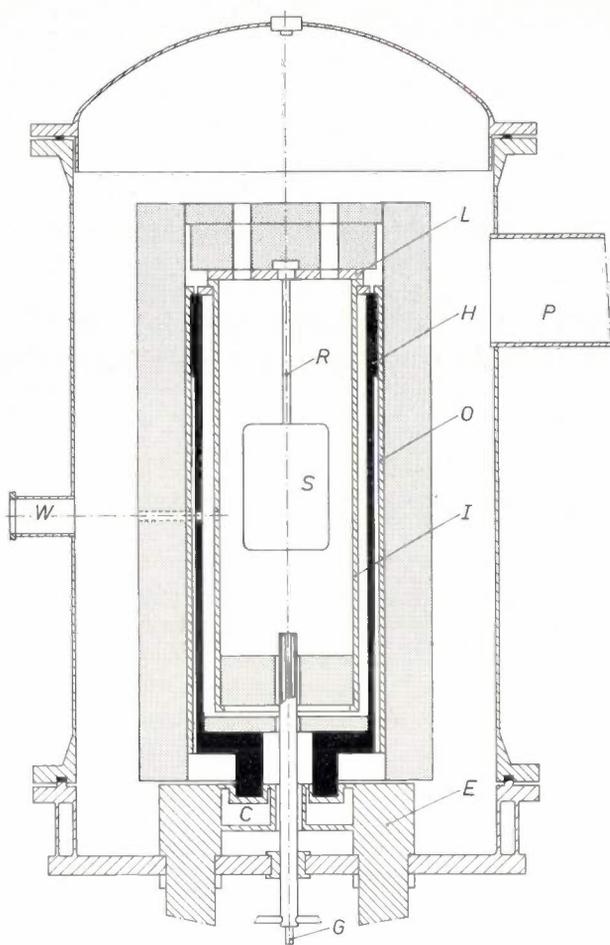


Fig. 15. Cross-section of a furnace for the preparation of pyrolytic graphite layers on large objects of arbitrary shape. *H* heater element. *E* electrodes with channels *C* for cooling water. *I* and *O* inner and outer protective tubes. *L* lid to which the substrate *S* is attached by means of a rod *R*. *G* inlet for the carbon-containing gas. *P* pump aperture. *W* window for temperature measurements by optical pyrometry. The graphite felt for thermal insulation is shown grey.

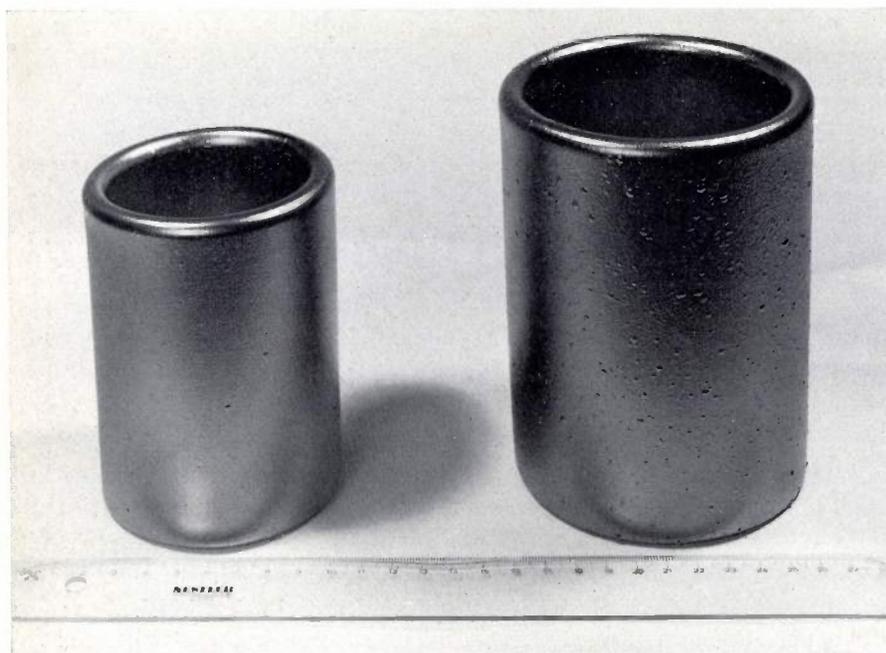


Fig. 17. Carbon crucibles coated inside and outside with a homogeneous pyrolytic graphite layer. The total wall thickness of the pyrolytic graphite is about 1 cm.

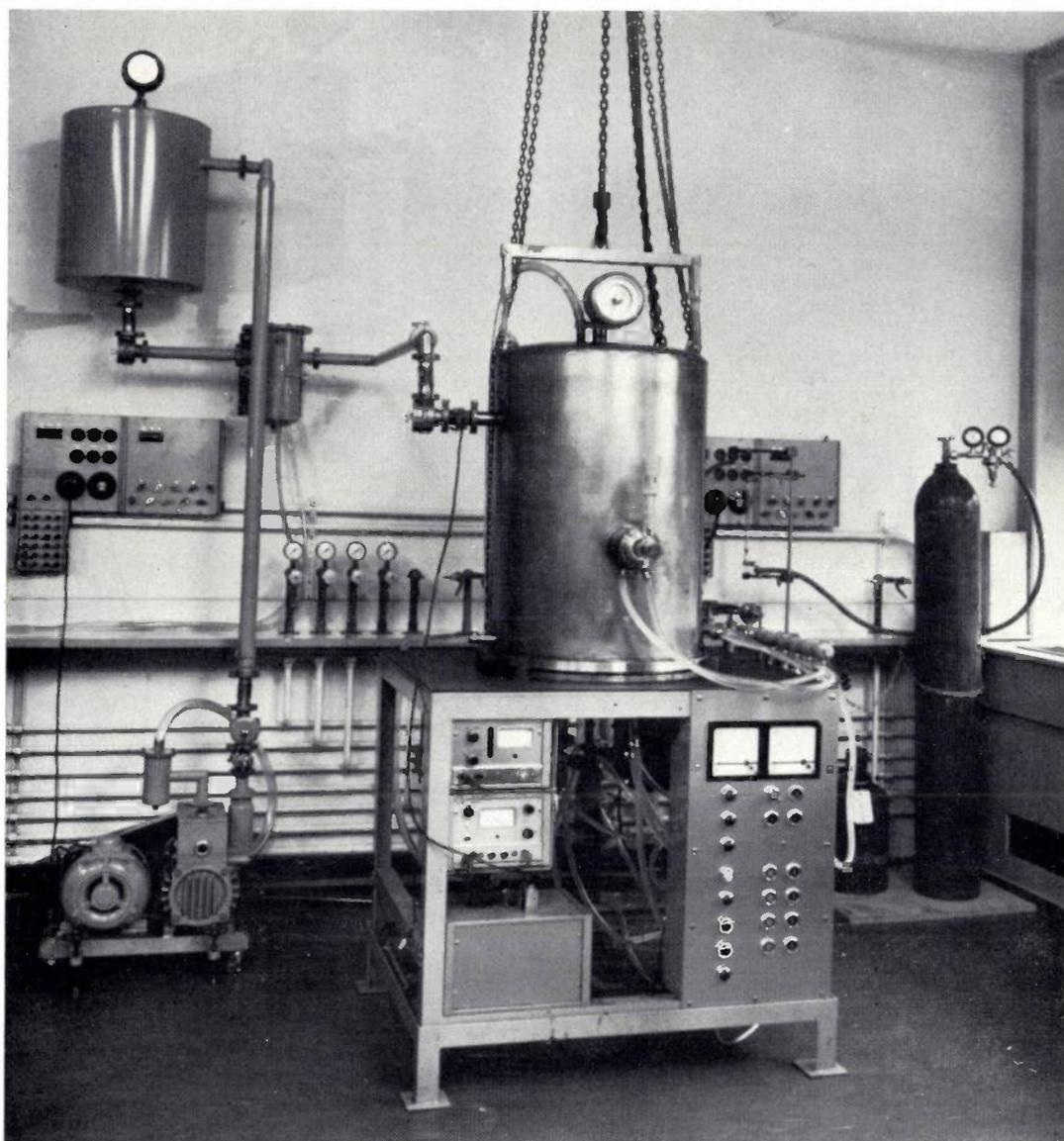


Fig. 16. Layout of the furnace in fig. 15. On the right, a cylinder of the carbon-containing gas which is admitted into the furnace; on the left, the pump for regulating the gas pressure in the furnace. The window through which the temperature is measured can be seen at the front of the furnace. The control desk under the furnace contains the electrical circuits for the Penning gauges and for the furnace heating.

this provides a means of making objects consisting of pyrolytic graphite only (*fig. 18*). A difficulty with the preparation of pyrolytic graphite in the furnace described is that deposits tend to settle on the protective inner tube (*I* in *fig. 15*). After the layer has reached a certain thickness, it is strong enough to cause the tube to break upon cooling down. To avoid damage to the heater element a space of about 1.5 cm is therefore left between this tube and the element. The inner tube itself can be effectively protected by a layer of graphite felt.



Fig. 18. Pyrolytic graphite crucible for growing silicon-carbide crystals. After cooling to room temperature the thick-walled crucible breaks away from the graphite substrate owing to the difference in expansion coefficient between pyrolytic and normal graphite.

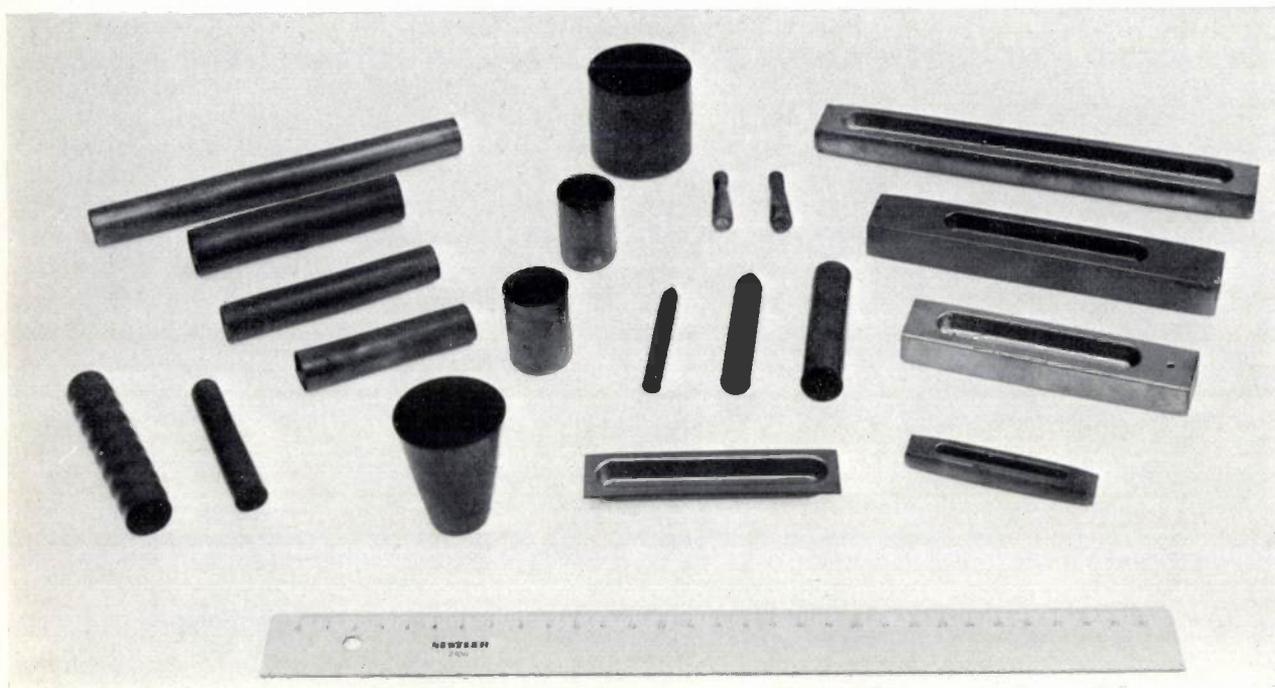


Fig. 19. Some graphite objects coated with a layer of pyrolytic graphite.

Although pyrolytic graphite settles on this felt, it can easily be replaced by a new layer of felt. The first layer of pyrolytic graphite deposited in a felt-lined furnace always has some surface irregularities, presumably caused by fresh nucleation due to deposits of fine carbon particles. Once the felt is covered with a layer of pyrolytic graphite this gives no further trouble. The difference is clearly to be seen in the two crucibles shown in fig. 17. The large crucible was made in a furnace that had only just been lined with felt, the small crucible was made in the next experiment. Fig. 19 shows various other graphite objects covered with pyrolytic graphite in our furnace.

Applications

Some of the advantages of pyrolytic graphite over normal graphite appear in its use as a crucible material. The thermal conductivity is high parallel to the wall of the crucible, but low perpendicular to the wall. The temperature gradients in a crucible of pyrolytic graphite are therefore smaller than in a normal graphite crucible. If the crucible is heated by the direct passage of electric current, the thermal conduction perpendicular to the wall is sufficiently low to ensure effective heat insulation. For this reason it is also advisable to make graphite tube furnaces from pyrolytic graphite. The thermal properties of pyrolytic graphite and the anisotropy of its electrical conductivity also favour its use as a

radiation shield in induction heating. In this case, however, the greatest benefit can be obtained from the anisotropy of the thermal conduction by heating thick-walled pyrolytic graphite crucibles with an induction current of low frequency (< 10 kHz).

Since pyrolytic graphite can be made with a very high degree of purity, it can be substituted for normal graphite in many applications in semiconductor technology (for crucibles, jigs, etc.).

For the same reason it is used as electrode material in spectrochemical analysis and also, because of its chemical resistance, as electrode material in electro-analytical chemistry. It is to be expected that as pyrolytic graphite becomes more widely known it will find an even greater range of applications.

Summary. Carbon layers deposited on substrates at temperatures above 2000°C from a carbon-containing gas have a clearly oriented graphitic structure. Carbon of this kind is known as *pyrolytic graphite*. The anisotropy of the properties of these deposits and their greater chemical inertia as compared with normal graphite widen the useful scope of carbon for scientific research and industrial applications. The article first presents a comparison of pyrolytic graphite with the ideal graphite crystal. When pyrolytic graphite is recrystallized under pressure a piece of graphite can be obtained that closely resembles the ideal graphite crystal. After a thermodynamic treatment of the carbon transport properties of various gas systems, the conditions under which pyrolytic graphite is deposited are examined. A practical method of preparation is described and a number of applications are mentioned.

The expansion ejector, a new cryogenic device

In recent years the need for continuous generation of cold at temperatures between 2 and 4 °K, and for liquid helium, has grown rapidly. In particular, developments in the field of computers, cryogenic stores, masers, superconductors, etc., present a challenge to cryogenic engineers to carry out refrigeration and liquefaction in this temperature range as simply and efficiently as possible.

Ever since interest was aroused in the liquefaction of gases the Joule-Thomson process has played a predominant role. It still is, in fact, the only process of practical importance for continuous refrigeration at temperatures of a few degrees Kelvin. In this process the gas passes through a cycle in which it expands from a high to a low pressure through an expansion valve (throttling). Provided the initial gas temperature is below the "inversion temperature" the expanding gas is cooled.

The Joule-Thomson cycle is sketched in *fig. 1*. After compression at room temperature by the compressor *C*, the gas at the high pressure p_1 flows through precoolers *P* and heat exchangers *H* to the expansion valve *E*. On leaving the expansion valve the gas at the low pressure p_2 flows through the heat exchangers back to the compressor. The precoolers cool the gas to below its inversion temperature and the high pressure input gas to the expansion valve is continuously pre-cooled in the heat exchangers by the low pressure exhaust fluid [1]. Eventually the temperature is reduced below the gas liquefaction temperature and the condensed fraction of liquid is separated out in *V*; it can either be evaporated (for producing cold) or tapped off (liquefaction).

It is inherent in this cycle that the pressure p_s at the suction side of the compressor is equal to p_2 or even less, owing to the flow resistance in the heat exchangers. The vapour pressure of helium drops rapidly with temperature, so that a very low suction pressure is required if liquid helium is to be produced at lower temperatures (*fig. 2a*). If we consider this together with the fact that a compressor for a given delivery will be larger for a lower suction pressure, while the input power increases with decreasing suction pressure, it be-

comes evident (see *fig. 2b*) that at temperatures of 3.5 °K or below, an excessively large compressor is required. The compressor required to achieve a given refrigerating capacity at 2.5 °K is ten times as large as at 4.2 °K.

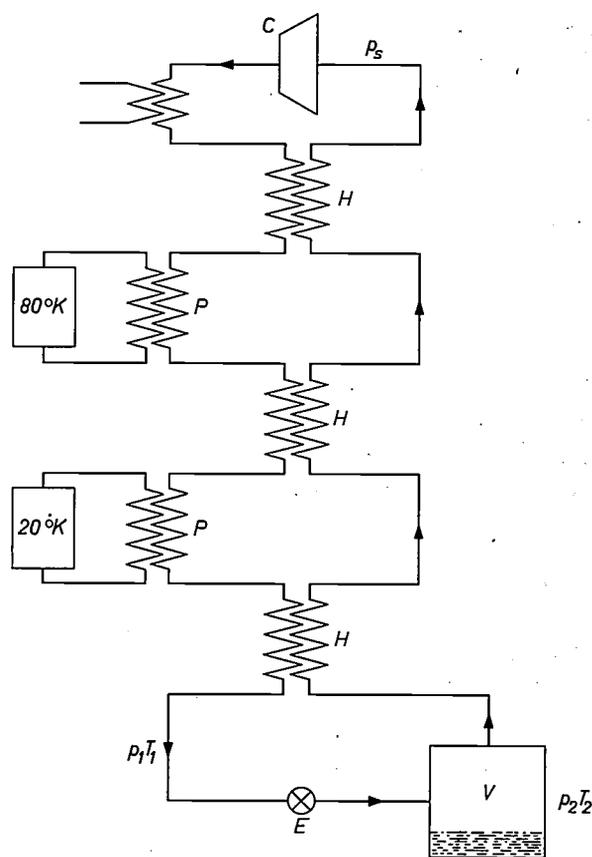


Fig. 1. The conventional Joule-Thomson cycle. The compressor *C* creates a pressure difference $p_1 - p_2$ across the expansion valve *E*. The gas, provided it is below the inversion temperature, cools down on expansion and condenses in *V*. *P* precoolers, *H* heat exchangers. The suction pressure p_s of the compressor is approximately equal to p_2 .

Investigations carried out in this laboratory have shown that this problem can be solved by a simple modification of the Joule-Thomson cycle. The central element of the modification is a device which we have called an *expansion ejector*.

First of all, we note that the conventional throttling

[1] We sometimes use the expression "fluid" instead of "gas", as the terms "gas" and "liquid" are rather vague in the region of the critical point.

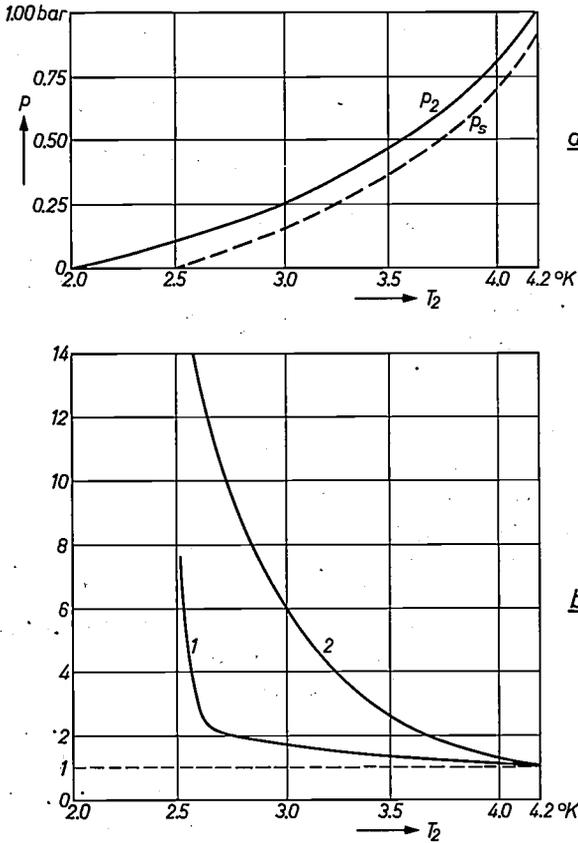


Fig. 2. a) The vapour pressure curve of helium, i.e. the pressure p_2 as a function of T_2 when there is liquid in V (fig. 1). P_s suction pressure of the compressor in fig. 1. It is assumed that the pressure drop $p_2 - p_s$ across the heat exchanger is constant at 0.1 bar for all temperatures. b) Input power (curve 1) and the dimensions (curve 2) of the compressor as a function of the required temperature T_2 in V , compared with the values at 4.2 °K.

process is not really very efficient. This may be seen by considering throttling through a small aperture (fig. 3). The high-grade energy which the gas has as a result of the high pressure p_1 is then first converted into di-

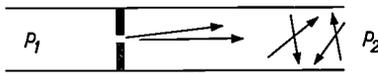


Fig. 3. Illustrating expansion through a small orifice.

rected kinetic energy (a "jet"). This directed kinetic energy is not however utilized as work or tapped off, but dissipated (converted into non-directed kinetic energy, i.e. heat). This cancels out a substantial amount of the expansion cooling in the gas, and it is only on account of certain specific properties of the gas that any cooling effect below the inversion temperature remains. The great advantage of the expansion valve,

however, is its extreme simplicity. In this it is unlike isentropic expansion which, although in principle much more effective, requires an expansion machine with moving parts. (With an ideal gas throttling would have no cooling effect, but there would be a cooling effect with isentropic expansion.)

The principle of the expansion ejector is the making use of this directed kinetic energy, which was previously wasted, to create suction power by jet action, rather like the action of a water jet pump. This suction power can then be utilized to achieve a lower temperature using the same compressor.

The modified cycle is sketched in fig. 4. Part of the

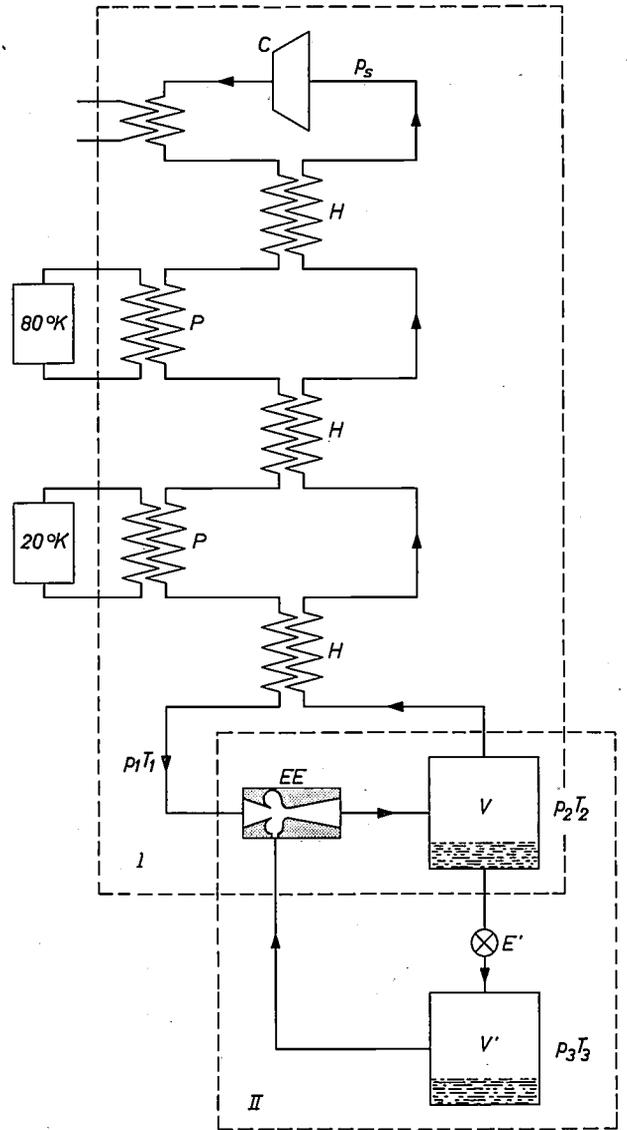


Fig. 4. The modified Joule-Thomson cycle. In addition to the primary cycle I there is a secondary cycle II. The expansion ejector EE throttles the fluid in the primary cycle and at the same time compresses the fluid in the secondary cycle from p_3 to p_2 . In E' the fluid is throttled in the secondary cycle. The pressure p_3 in V' can now be substantially lower than the suction pressure p_s of the compressor.

fluid in the vessel V is fed into a second vessel V' through an expansion valve E' . The action of the expansion ejector then reduces the pressure in V' to p_3 which is much lower than p_2 in V , so that the temperature T_3 is likewise considerably lower than T_2 . This permits the suction pressure p_s of the compressor to differ from the vapour pressure p_3 at the required low temperature T_3 . The parts of the expansion ejector can be seen in *fig. 5*. Here 1 is the jet nozzle, 2 is the suction region, 3 is the mixing zone and 4 the diffuser. The expansion ejector thus acts at the same time as an expansion valve (between p_1 and p_2) and as a "jet compressor" (between p_3 and p_2).

The principle described can be used in two ways:

a) In a machine for continuous refrigeration at extremely low temperatures. A machine of this kind operates with a closed cycle: no liquid helium is tapped off, and so helium has to be supplied. In an experimental arrangement with a pump suction pressure of 1 bar an expansion ejector suction pressure of about 50 torr has been achieved, which corresponds to a temperature T_3 of 2.3 °K.

b) In a liquefier. The liquid helium is tapped off and in the compressor gaseous helium has to be supplied to the system. The helium is tapped off at $p_3 = 1$ bar. The suction pressure p_2 can now be 2.5 to 3 bar, so that at $p_1 = 20$ bar the compressor need only have a compression ratio of 20/3 instead of 20, as required in the conventional cycle. A helium liquefier based on this principle is now being marketed by Philips[*].

The expansion ejector can not only be put to use in systems like that of *fig. 4*, but in all refrigerating processes employing an expansion valve, for example in processes where the refrigeration is produced partly by expansion machines.

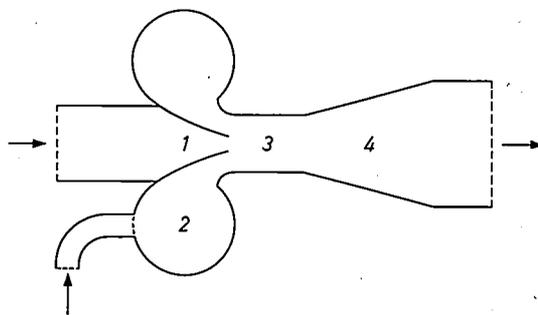


Fig. 5. The expansion ejector. 1 jet nozzle, 2 suction region, 3 mixing zone, 4 diffuser.

It is interesting to note that any positive effect of the expansion ejector means pure gain with respect to a normal expansion valve, as the ejector cannot perform less efficiently than such a valve. Studies aimed at deriving the optimum design of expansion ejector are being pursued in this laboratory.

The advantage of the expansion ejector can be summarized as follows. A required amount of cold can be produced at a given temperature more efficiently and with simpler means (i.e. a smaller compressor) than when only an expansion valve is used. The expansion ejector itself is an extremely simple device; it has no moving parts and is no larger than a match-box.

J. A. Rietdijk

Dr. Ir. J. A. Rietdijk is with Philips Research Laboratories, Eindhoven.

[*] *A description of this helium liquefier is to appear later in this journal. Ed.*

Metallurgical aspects of the alloy-diffusion method in transistor technology

P. J. W. Jochems and E. Kooi

The alloy-diffusion principle continues to be employed with considerable success in the quantity production of germanium transistors for high frequencies. Research into preparation methods based on this principle has brought to light many interesting phenomena, which often give rise to difficulties. Some of these phenomena are discussed in the following article; their presence has been most noticeable in attempts to find methods for making silicon transistors, but these phenomena are by no means absent with germanium.

The alloy-diffusion principle; p.o.b. transistors

In transistors for use at high frequencies the base has to be a layer which is extremely thin. One of the basic principles which can be applied in making such a thin layer is that of alloy-diffusion. Several tens of millions of germanium transistors have been made in the Philips Semiconductor Works at Nijmegen (Netherlands) by means of procedures based on this principle [1]. Silicon transistors are more usually made by other methods, in particular the "planar technique" [2].

Since it is necessary to use systems of three or four components, i.e. complicated systems in phase theory, various interesting metallurgical problems are met with in research into suitable techniques. Some of these problems will be reviewed in this article. First of all, however, we shall give a brief résumé of the alloy-diffusion principle itself.

Fig. 1 shows a diagram of the method used to make a germanium *P-N-P* transistor by an alloy-diffusion process. A lead pellet doped with two elements, one acting in germanium as a donor (antimony) and the other as an acceptor (aluminium), is placed on a wafer of *P*-type germanium of the resistivity required for the collector. If the wafer and pellets are heated to about 780 °C, the lead pellet melts and a pit is formed beneath it in the germanium; the size of this pit is determined by the quantity of germanium that can dissolve in the droplet of lead at 780 °C. There is also diffusion of the antimony, which penetrates the germanium fairly quickly at 780 °C. The diffusion takes place both directly from the lead droplet and by way of the vapour phase (fig. 1a). The surface layer of the germanium wafer is converted by this into *N*-type germanium. The temperature of 780 °C is maintained for as long as

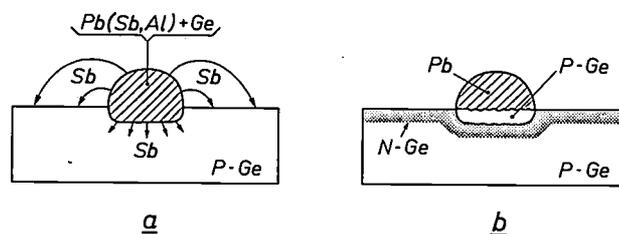


Fig. 1. Diagram showing how a transistor structure (*P-N-P*) is produced beneath a double-doped pellet of lead on a germanium wafer. *a*) The lead pellet is melted and lies in a pit of dissolved germanium (temperature 780 °C). Antimony (the donor) diffuses into the germanium, partly directly and partly by way of the vapour phase, while aluminium (the acceptor) does not. *b*) On cooling, aluminium-containing germanium is deposited in the pit (*P*-type). This is the emitter of the transistor. Beneath and around it there is a heavily antimony-doped layer of germanium (*N*-type), which can act as the base.

necessary for this layer, which will become the base of the transistor, to reach the desired thickness. When the temperature is allowed to drop, the dissolved germanium is deposited at the bottom of the pit and the lead solidifies. Since aluminium is much more soluble than antimony in solid germanium, the germanium deposited is again of type *P*, in spite of the presence of antimony (fig. 1b). The solubility of aluminium is in fact high enough to give strong doping, so that the deposited germanium has the required properties for an emitter.

Because, in such a process, the base of the transistor is formed by diffusion from a drop of molten metal — it is "pushed out" of the drop — we use the term "pushed-out base", abbreviated to p.o.b., and also refer to p.o.b. processes and p.o.b. transistors.

Clearly, only a limited number of combinations of materials can be considered for a p.o.b. process. The materials used for doping, the material carrying them (the lead in the present case) and the semiconductor (germanium or silicon) must meet a wide complex of

interdependent requirements. As we shall see below, complications arise because the pellets have to be as small as possible to increase the maximum frequency at which the transistor will operate. Further difficulties may arise because a second pellet, to act as a base contact, has to be placed close to the pellet beneath which the transistor configuration is to be formed. We shall deal with this particular problem in the final section of this article. The sections immediately following will therefore be devoted to phenomena encountered in and close to the first metal pellet. We shall begin with the case in which the semiconductor is silicon, as the metallurgical situation here is the simplest.

Silicon

The properties of materials suitable for doping were examined to see which materials dissolve well in silicon but diffuse into it slowly, and which dissolve poorly but diffuse into it quickly. Roughly speaking, the donors (As, Sb, P, etc.) correspond to the first case and the acceptors (Al, Ga, etc.) to the second. Thus the situation is exactly the opposite of that which holds with germanium. A silicon p.o.b. transistor of *P-N-P* structure cannot therefore be made, but a silicon p.o.b. *N-P-N* transistor can be made, provided that *N*-type silicon is used as starting material.

To ensure that the diffusion process takes place reasonably quickly, the temperature must be at least 1000 °C, i.e. about 250 °C higher than with germanium. Tin is suitable here as a carrier. It has a sufficiently low vapour pressure and silicon is not too soluble in it at the working temperature required, while on the other hand its low melting point (232 °C), makes it very easy to attach leads to it. (Both with Si and Ge it is in principle possible to dispense with the carrier. In general, however, this is no simpler, and the facility of easy attachment of the leads has been lost.)

Now one of the problems we mentioned above arises: it is very difficult to make doped tin pellets which are small enough (diameter smaller than 0.1 mm for 200 MHz) and yet homogeneous in composition. If, for example, about 2% of gallium and arsenic are dissolved in molten tin at 1200 °C, and the solution is then cooled quickly to room temperature, a material is obtained containing separate crystals of GaAs (fig. 2). Since the GaAs crystals can be as large as a few tens of microns, the material obtained in this way is completely unsuitable for making the homogeneous pellets required.

Similar difficulties are encountered with Ga-Sb, Al-As and Al-P combinations. The latter two form compounds which have a very high melting point (above 1600 °C) and are poorly soluble in tin. Furthermore, these compounds react readily with water vapour

or oxygen to form aluminium oxide. This greatly hinders the alloying of silicon with tin pellets containing one of these materials.

The difficulties outlined above are not completely insuperable. One procedure which may be used when working with the combination Al-As is to dope the tin initially with arsenic alone. Although Sn-As compounds are formed, these do not give rise to serious difficulty as they melt at fairly low temperatures and so dissolve in the tin. A pellet of this material is placed on a silicon wafer, the temperature is increased until the tin melts and a pit is formed. The temperature is then allowed to drop again. The aluminium is applied in the form of a suspension to the solid droplet of tin and the temperature is again raised until the tin melts, so that the aluminium can dissolve in it, and so on. (This method

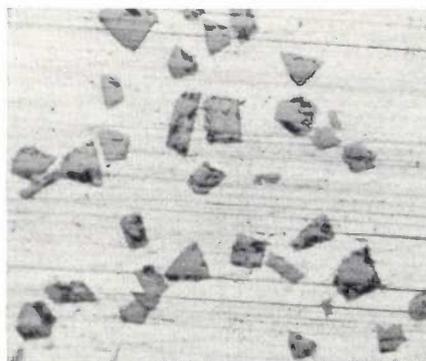


Fig. 2. Microscopic section of a piece of doped tin obtained by the fast cooling of a melt (1200 °C) containing 2% gallium and 2% arsenic. The material is not homogeneous, but contains GaAs crystals. Magnification 100×.

of doping with aluminium has already been used in the manufacture of Ge transistors; see below.)

In a variation of this process Al is added to the molten tin pellet by the way of the vapour phase; this can be done if the temperature is higher than 1000 °C. In principle, all that is necessary is to put down a little Al powder near the pellet. This method permits accurate control of the quantity of Al reaching the pellet. Fig. 3 shows an example of a p.o.b. silicon transistor with an *N-P-N* structure.

The formation of solid compounds

When applying the method that has just been outlined the concentrations can only be varied within fairly close limits. Outside these limits, other undesirable effects are met. The examination of these effects has

[1] P. J. W. Jochems, Philips tech. Rev. 24, 231, 1962/63.

[2] This is described in A. Schmitz, Solid circuits, Philips tech. Rev. 27, 192-199, 1966 (No. 7).

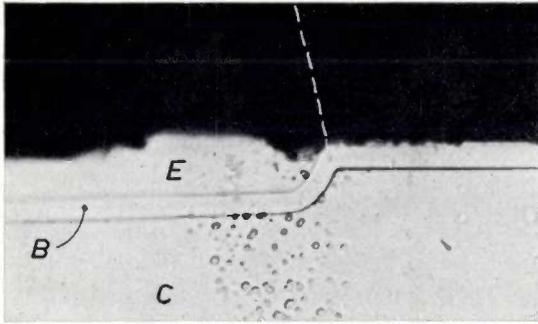


Fig. 3. Microscopic section^[3] of an experimental silicon *N-P-N* transistor made by an alloy-diffusion process. (Magnification 550 \times .) *C* collector. *B* base layer (4.5 μm thick). *E* emitter. The tin pellet from which the emitter was grown has been removed; the position of its edge is marked by the white dashed line.

however led to a better understanding of what actually happens in the Si-Sn-Al-As system. The phenomena with which we were faced were these:

- 1) Whenever the Al concentration in the molten droplet was higher than the As concentration, the operation of the Si-Sn-Al-As alloy as an emitter was often poor. Only very little As was found to have entered the alloy, which was therefore insufficiently *N*-type.
- 2) If a higher As concentration was used, the Al concentration remaining the same, the Al often did not diffuse, and as a result no base was formed.

We think that the simplest explanation for all this is that the compound Al-As is only slightly soluble in molten Sn-Si. Of course, it may be that supersaturation occurs as the temperature is lowered and the p.o.b. process takes place normally. If, however, crystalline Al-As is formed, little As will remain behind in the melt if the Al is present in excess. The silicon deposited on cooling cannot then contain a very large quantity of As. If on the other hand there is an excess of As, the melt becomes very poor in Al when the Al-As crystallizes out and there will be hardly any Al diffusion. (Apart from the formation of solid Al-As, interaction between Al and As in the melt or in the silicon can have some effect.)

Germanium

One may wonder whether difficulties similar to the ones described above for silicon will occur when alloying germanium with a doped pellet of lead as described in the introduction. An inspection of the phase diagrams for the sub-systems Pb-Al, Pb-Sb and Al-Sb will

show that even more difficulties are to be expected. For instance, aluminium is very poorly soluble in lead; the solubility is only 0.1% by wt. at 660 $^{\circ}\text{C}$, the melting point (fig. 4). It is therefore difficult to make small pellets of lead doped with Al and Sb. In practice, these difficulties are avoided by starting with a Pb(Sb) pellet and alloying the germanium wafer with it in the usual way. After this has cooled, a little finely divided Al is applied to the pellet, again in the form of a suspension, and the pellet and wafer are reheated. Some of the aluminium then dissolves in the molten lead. After cooling, the major portion of this aluminium is to be found in the germanium, since aluminium dissolves very readily in solid germanium. This process can therefore be used to provide a *P*-type emitter region sufficiently heavily doped with Al.

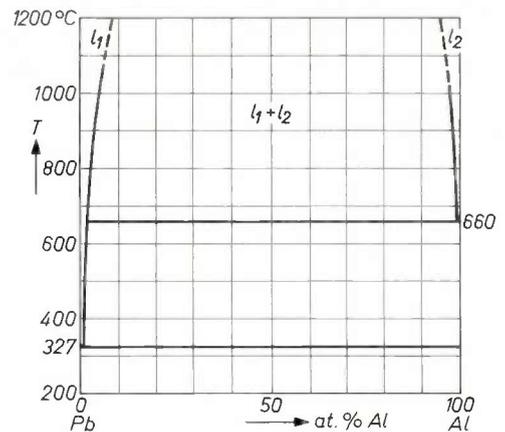


Fig. 4. Phase diagram^[4] of the Pb-Al system. Aluminium is very poorly soluble in lead. Nor does molten aluminium ($T > 660^{\circ}\text{C}$) mix very readily with molten lead (melting point 327 $^{\circ}\text{C}$). l_1 molten lead containing a little aluminium. l_2 molten aluminium containing a little lead.

The poor solubility of Al in Pb, which at first appeared to be a difficulty, is in fact a considerable advantage in the practical application of this process. It does not matter how much Al is applied to the lead pellet, since the quantity dissolved is determined only by the temperature. This quantity, and therefore the quantity finally to be found in the germanium, is completely under control.

It has been found in practice that variation of the Sb content of the original lead pellet allows a certain variation in the base resistance of the transistors being made. There are however limits to the variation that can be obtained: if the Sb content is increased to such an extent that solid Al-Sb is formed, the concentrations of Al and Sb in the melt remain constant as shown below, and the advantage becomes a drawback.

If we have a molten lead pellet containing Sb, on the germa-

[3] Prepared by B. Jansen at this laboratory using a method specially developed for the purpose; see *Solid State Electronics*, 2, 14, 1961.

[4] Taken from M. Hansen and K. Anderko, *Constitution of binary alloys*, 2nd ed., McGraw-Hill, New York 1958.

[5] See reference [1].

nium at 700 °C, and there is some Al on the pellet, we then have a system with four components: Pb, Sb, Al and Ge, and three phases: the molten lead (with Ge, Sb and Al dissolved in it), molten aluminium (with some Pb, Ge and Sb dissolved in it), and solid Ge. There is therefore a free choice of three quantities, or if the temperature and pressure are fixed, of one quantity, e.g. the lead content of one of the liquid phases. If the Sb concentration in the lead is now increased to such an extent that solid Al-Sb is formed, there are four phases instead of three and at a given temperature and pressure all the concentrations are fixed.

If the doping is with As instead of Sb, the limits to the range of variation are even closer.

Complications caused by the base contact

Besides the pellet from which the alloy-diffusion process takes place, giving the desired transistor configuration, there has to be a second pellet, placed close to the first on the germanium or silicon wafer, to act later as a base contact. With silicon, this leads to fresh difficulties as the two pellets require different doping materials: at high temperatures part of the doping material in one pellet can penetrate into the other by way of the vapour phase.

Before dealing with this, we shall briefly recall the way in which a complete germanium transistor is made.

In the manufacture of germanium transistors, the second pellet of lead is doped with antimony only, i.e. not with aluminium. Since antimony diffuses into the germanium beneath and between both pellets, there is a conducting path of N-type material between the second pellet and the base of the transistor^[5] (fig. 5).

We saw above that the first pellet is also doped initially with antimony only, and that the aluminium is added later. The process takes place therefore in the

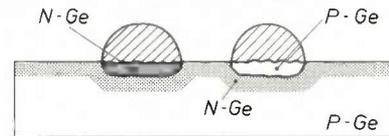


Fig. 5. Diagram of a germanium P-N-P transistor complete with base contact. The contact consists of a second lead pellet doped only with antimony and from which, therefore, N-type germanium separates out in the pit. It is connected to the base by the N-type layer in the germanium wafer (cf. fig. 1b).

following simple way. Two identical Pb(Sb) pellets are placed close together on the germanium (fig. 6), and heated to about 600 °C. The pellets melt at this temperature and some of the germanium is dissolved in them. The Ge wafer and pellets are then cooled to room temperature. The Al suspension mentioned above is then applied to one of the pellets and the wafer and pellets are heated once more, but this time to a higher temperature, e.g. 780 °C, and cooled again after a quarter of an hour. The greater part of the diffusion of Sb takes place during this second heating cycle and the Al-doped emitter is formed during the final cooling.

Experimental silicon transistors

A method similar to that just described for germanium cannot be used for the manufacture of N-P-N silicon transistors. We can of course begin with two Sn(Al) pellets, but if one of them is to be doped later with As, this material would also contaminate the other pellet, since As is fairly volatile.

We have found that it is nevertheless possible to start

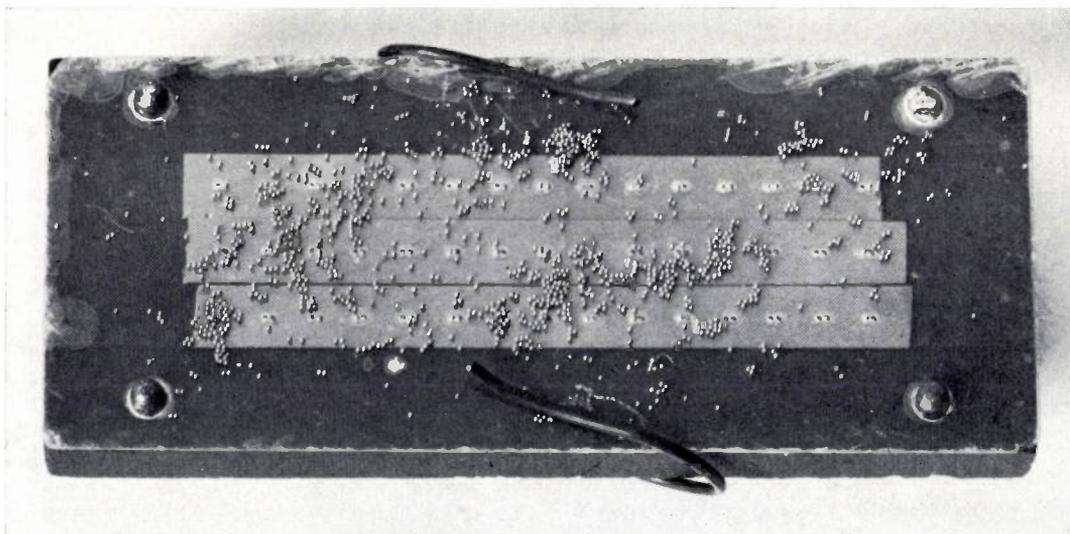


Fig. 6. Jig, as used in the laboratory for locating pairs of lead pellets on a germanium wafer. There are three strips of germanium on a graphite base. Above these strips there is a sheet of mica, located by four pins and held down by two springs. The mica sheet has pairs of holes drilled in it, and the lead pellets are brushed across the mica till all these holes are filled.

with the same material for both pellets, provided that the common doping agent is the volatile one. There are naturally certain consequences to this: one, as we shall see in a moment, is that the pellet which receives the second doping agent forms the base contact and not the transistor.

The following procedure was used. Two tin pellets containing 1% As were applied at 1000 to 1050 °C to a silicon wafer, whose surface layer had been doped with gallium to give it *P*-type conductivity (fig. 7a and b). After cooling, a little finely divided Al was then applied, as with germanium, and the wafer and pellets were reheated. If the duration and temperature of this second heating process are correct, the following occurs (fig. 7c and d). Part of the aluminium dissolves in the pellet to which it was applied (1), but some aluminium also enters the other pellet (2) by way of the vapour phase. Aluminium diffuses into the silicon from both pellets, forming a *P*-type layer in it. If the wafer and pellets are then cooled, a quite different situation is now found for each of the two pellets. There is relatively little aluminium in pellet 2 and, because of the heavier arsenic doping, the silicon deposited in the pit is of type *N*. In this case therefore, pellet 2, with the part of the silicon wafer beneath it, forms the *N-P-N* transistor.

Now, the other pellet can in fact act as the base contact. Not only is it connected to the *P*-type layer of the transistor through the gallium-doped surface layer of the silicon wafer, but the silicon deposited in the pit is also itself of type *P*, as was required. This is because AlAs forms in pellet 1 owing to the excess of aluminium, so that the As concentration in the melt is low and the silicon that crystallizes out is much more heavily doped with Al than with As — exactly the opposite situation to that in pellet 2.

It has been found in practice that silicon transistors made in this way still have too high a resistance between the semiconductor and the tin base contact. This comes about because aluminium is not very soluble in silicon and the alloyed silicon is therefore too lightly doped. The difficulty can be remedied by using boron as a doping agent for pellet 1 (as well as aluminium). The boron also acts as an acceptor in silicon and dissolves well in it. It is not possible to use boron alone and omit the aluminium, since boron does not dissolve very well in pure tin.

We can sum up by noting that the many metallurgical aspects associated with the manufacture of semiconductor devices on germanium and silicon by means of an alloy-diffusion process can give rise to many difficulties. In particular, great care is required because of the risk of the formation of chemical compounds. How-

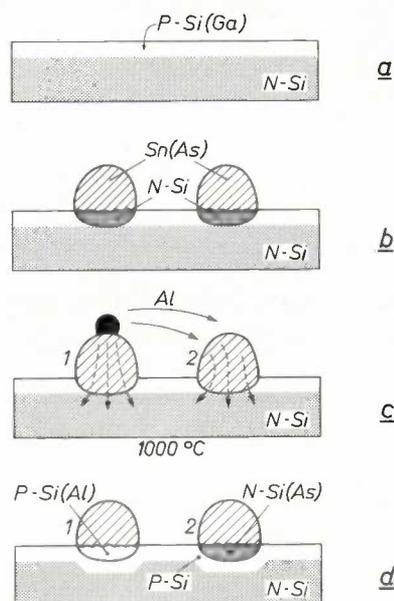


Fig. 7. Diagrammatic representation of an alloy-diffusion process for the manufacture of silicon *N-P-N* transistors.

- A wafer of *N*-type silicon is given a *P*-type surface layer by diffusion of gallium.
- Two pellets of Sn(As) are placed upon it, melted and recooled.
- Finely divided Al is applied to pellet 1 and wafer and pellets are heated to about 1000 °C. Most of the Al dissolves in pellet 1 while part of it vaporizes over to pellet 2. Aluminium diffuses from both pellets into the silicon and forms a layer of *P*-type silicon beneath them.
- On cooling, *P*-type silicon is deposited in pit 1; the lead melt in the pit is rich in Al but contains little As owing to the formation of AlAs. *N*-type silicon is deposited in pit 2. This becomes the emitter of the transistor.

ever, the metallurgical aspects can often be turned to advantage — sometimes in a rather unusual way — once a proper understanding of the background has been acquired.

Summary. Alloy-diffusion processes for high-frequency transistor manufacture, which have been applied in industry for years in making Ge transistors, have many metallurgical aspects, which sometimes give rise to difficulties. Complications often occur because the two doping materials combine to form a high-melting-point compound (GaAs, GaSb, AlAs, AlP); the minority substance then disappears almost completely from the melt. The best results are obtained with singly-doped pellets. In manufacturing Ge *P-N-P* transistors Pb(Sb) pellets are first melted on a Ge wafer. After cooling, finely divided Al is applied to one pellet of each pair, and they are melted again. The same procedure, but with Sn(As) pellets, can be used with Si. Some of the Al then enters the other pellet via the vapour phase and after diffusion forms the (*P*-type) base beneath it. AlAs is formed in the first pellet; the As concentration therefore decreases considerably, and the Si that crystallizes out also becomes *P*-type. This pellet becomes the base contact. An *N-P-N* transistor is thus obtained under the pellet which had no Al applied to it.

Current-limiting circuits for transistorized power supplies

R. Gasser and R. Hug

One of the areas of electronics in which valves have nearly completely been superseded by semi-conductors is that of stabilized d.c. power supply units. The very low internal resistance of these units makes it necessary to safeguard them against excessive current drain. In this paper the authors discuss the principles of the circuits used for this purpose.

Introduction

In stabilized d.c. power supply equipment valves have nowadays almost completely been replaced by diodes and transistors. This has brought about a reduction in size and in price of the apparatus involved. For this reason the fields of application for stabilized power supply units is continuously increasing. They are now used in places where batteries or unstabilized rectifiers were formerly employed.

A well-stabilized power supply provides a d.c. voltage which is scarcely affected by variations of the mains voltage or of the load current. Stabilization against load current variation implies that the units have very low internal resistance.

Another requirement is that the output voltage should be independent of temperature. As several semiconductor parameters depend on temperature, special attention must be paid to this point in transistorized equipment.

Circuit of a stabilized power supply

The operation of a stabilized power supply has been previously discussed in Philips Technical Review by Klein and Zaalberg van Zelst^[1]. The principles, which are those of a feed-back circuit, will be put in perspective with the aid of *fig. 1*. The unregulated voltage supply (which is nearly always a two phase rectifier) with e.m.f. U_u and internal resistance R_i is connected to the output terminals *a* and *b* via the control transistor Tr_1 . This transistor is driven by the amplifier *A*, which amplifies the difference between a fraction *k* of the output voltage U_1 and a fixed reference voltage U_r . The control transistor acts as a resistance in series with the load R_l , the value of this series resistance being affected by the difference amplifier *A*. The sign of the amplification is such that an increase of U_1 causes an increase of the equivalent resistance of Tr_1 . With a

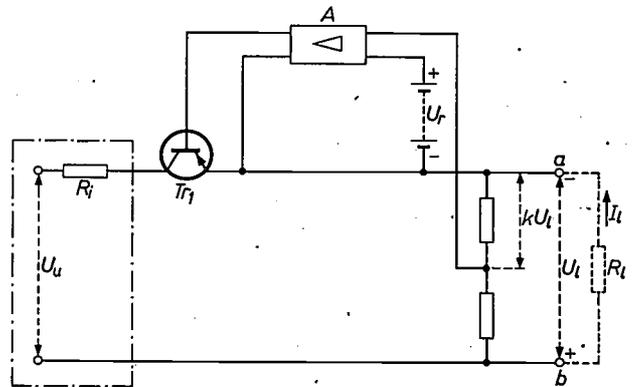


Fig. 1. Simplified diagram of a stabilized power supply unit. The unregulated voltage supply with e.m.f. U_u and internal resistance R_i is connected to the output terminals *a* and *b* via the transistor Tr_1 . Tr_1 is controlled by the amplifier *A*, which amplifies the difference of a certain fraction *k* of the output voltage U_1 and a fixed reference voltage U_r . R_l load resistance. I_l load current.

sufficiently large amplification of *A* the output voltage can be effectively stabilized in this way. This includes stabilization against variations caused by changes of the uncontrolled voltage U_u as well as those originating from a changing load current I_l . The latter effect is often called *regulation*^[1].

As the internal resistance of a well-stabilized power supply is very low, the connection of a small load resistance brings about a very large current, possibly damaging the power supply as well as the circuits connected to it. As a means of protection a fuse can be used, but in many cases a heavy overload or even a short-circuit may occur so suddenly that a fuse is too slow to prevent damage of the equipment. For this reason electronic circuits are incorporated in the power supply to prevent the current from exceeding a certain limit.

[1] See G. Klein and J. J. Zaalberg van Zelst, Combinations of valves and transistors in a stabilized 2000 V power supply, Philips tech. Rev. 25, 181-190, 1963/64, G. Klein and J. J. Zaalberg van Zelst, Instrumental electronics, Philips Technical Library 1967.

These circuits act so quickly that no harm can be caused to any component. The design of these current-limiting circuits forms the subject of this article.

Different protection methods

For controlling a current-limiting circuit an indication is needed of the value of the load current. This can be obtained by connecting a small resistance R_p in series with the load (fig. 2). The voltage across R_p controls the protection circuit P . When this voltage exceeds a certain value, P supplies a signal to one of the stages of the amplifier A . The control transistor is then affected in such a way that the required protection is obtained. The ways of operation of such protection circuits may differ considerably; we shall now deal with some of the usual methods.

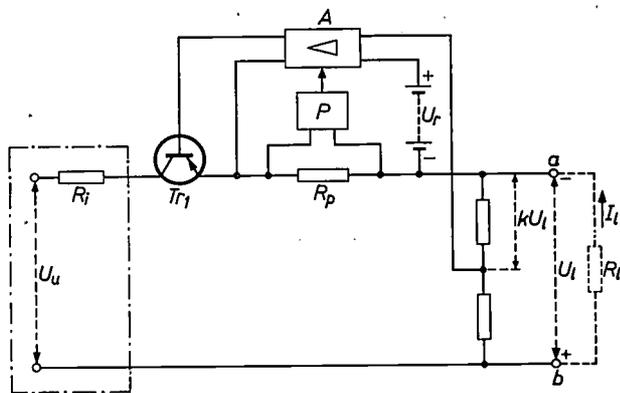


Fig. 2. To safeguard the equipment against overload, a protection circuit P is included. This circuit is controlled by the voltage across a small resistance R_p in series with the load. The circuit P delivers the appropriate signal voltage to one of the stages of the amplifier to give the required protection against overload.

Cut-out protection

With the first system the exceeding of a certain value I_{1m} of the load current I_1 causes the control transistor to be cut-off; the output voltage U_1 drops from the nominal value U_{1n} to zero and so does the current. This is illustrated in fig. 3, which shows the voltage-current characteristic of an ideally stabilized power supply provided with the protection system mentioned. Mostly the operation takes place in such a way that the cut-out condition is only terminated after pushing a special button or after switching the equipment off and then on again. If the overloading is then still present, the cut-out comes into action again.

In this way very effective protection may be obtained with a simple circuit. However, some objections to this method can be raised. It is often considered as inadmissible that a very short lasting strong current pulse,

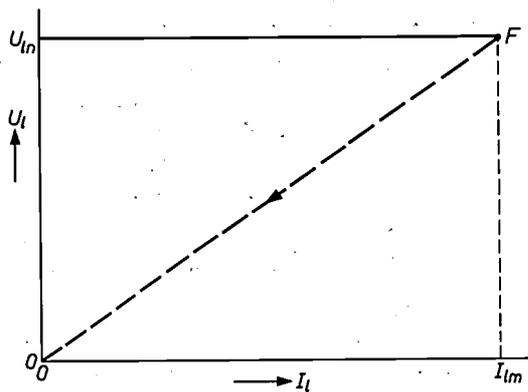


Fig. 3. Voltage-current characteristic of a stabilized power supply with cut-out protection. U_1 output voltage with nominal value U_{1n} . I_1 output current with maximum value I_{1m} . When the full-load point F is exceeded the equipment is cut-out (dashed line).

such as the one occurring when a capacitive load is connected, cuts off the power supply. This disadvantage may be partly overcome by introducing a certain time delay in the cut-out process.

Another disadvantage, which is not eliminated by a time delay, is that the power supply unit cannot be used to find out the cause of an overload; the equipment simply "refuses" to deliver a voltage to a load resistance which is too small.

Current limiting

The objections mentioned do not occur when a protection method is followed in which attainment of the maximum allowable current I_{1m} does not cut off the equipment. A decrease of the load resistance below the lowest permissible value (U_{1n}/I_{1m}) may cause the voltage to be reduced in such a way that the current is kept at the value I_{1m} (fig. 4). With this system, however, an-

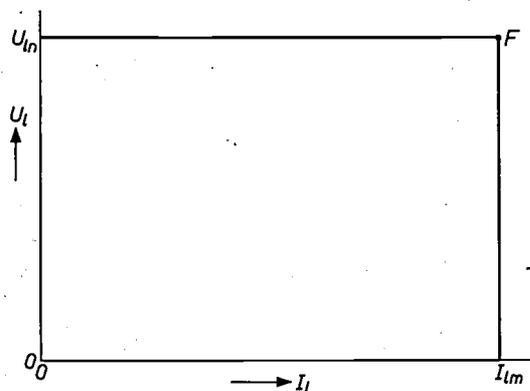


Fig. 4. Protection by a pure current-limiting circuit. When the full-load point F is reached the current does not exceed the value I_{1m} . This entails however a heavy power loss in the control transistor.

other difficulty arises. The decrease of the output voltage brings about an increase of the voltage on the control transistor. The power dissipation in this transistor may reach a far greater value than that occurring when the power supply is fully loaded (point *F*); if there is a short circuit of the terminals ($U_1 = 0$), almost the whole nominal power is dissipated in the control transistor. This must be taken into account in the design of the equipment: it may be necessary to put two or more transistors in parallel and it may be necessary to fit large cooling fins to obtain sufficient cooling. The dimensions of the apparatus could be considerably increased in this way.

Power limiting in the control transistor

Excessive power dissipation in the control transistor is prevented if a protection system is used in which a decrease of the output voltage is accompanied by a decrease of the current in such a way that the power dissipation in the control transistor is kept constant. The relation between voltage and current is now given by a line similar to the one shown in *fig. 5*. In this case no additional power losses in the control transistor need to be taken into account; when the power supply is overloaded, or even short-circuited, the dissipation in this transistor stays equal to the value at full load (point *F*). As no extra measures for cooling need to be taken, this may enable the designer to reduce the dimensions of the equipment. A more complicated circuit is however the price that must be paid. Another objection to this system is that jumps in the voltage and the current occur at the start and finish of the overload of the power supply. This is illustrated in *fig. 6*: When the load current has its maximum value (point *F*) even a small decrease of the load resistance makes the voltage and current jump to the values corresponding to point *B*. Further decrease of the load resistance is accompanied by a continuous fall of voltage and current till the short-circuit point *C* is reached. At a subsequent increase of the load resistance the voltage and current rise till point *D* is reached, and a jump to point *E* then occurs.

The irregularities mentioned may be quite troublesome. It is possible to design the circuit in such a way that the voltage and current jumps do not occur. Obviously the voltage-current characteristic should then have a shape like that shown in *fig. 7*: the straight line connecting *F* with the origin should not intersect the curved line *FC*. However, a characteristic like this can only be realized in a device having a rather low efficiency. It can be shown that the efficiency (the ratio of the power delivered to the load to the power delivered by the uncontrolled supply (see *figs. 1 and 2*) is less than 50% if such a characteristic is used.

If we denote the constant power loss in the control transistor Tr_1 by P_{tr} , the equation for the curve *FC* in *fig. 7* can be written in the following form (see also *fig. 2*):

$$U_1 = U_u - \frac{P_{tr}}{I_1} - I_1(R_1 + R_D).$$

The slope of this curve at point *F* is:

$$\left(\frac{dU_1}{dI_1}\right)_F = \frac{(P_{tr})_F}{I_{1m}^2} - (R_1 + R_D),$$

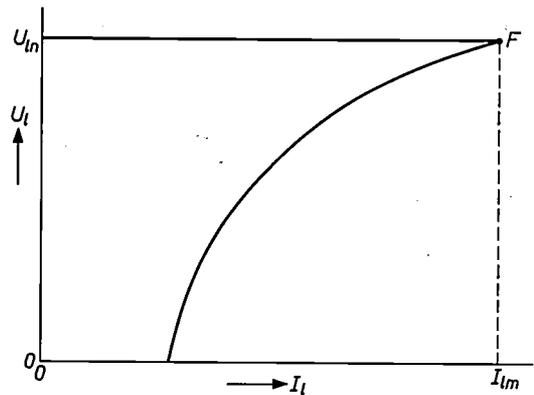


Fig. 5. Protection by a circuit which keeps the power loss in the control transistor at a constant value.

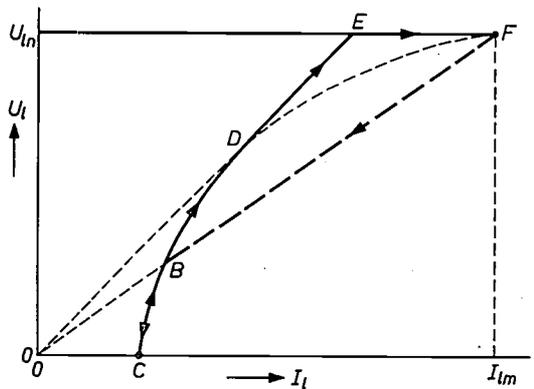


Fig. 6. With a constant power-loss system the operating point will jump from *F* to *B* and from *D* to *E*. *C* is the short-circuit point.

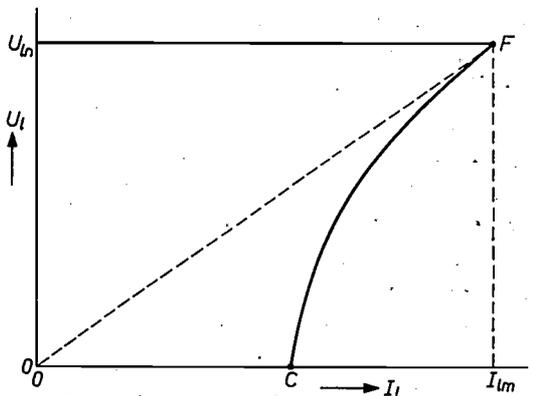


Fig. 7. With a characteristic like the one shown (*FC*), no voltage and current jumps occur. This entails however a low efficiency of the power supply unit.

$(P_{tr})_F$ being the power loss in the transistor in fully loaded condition. Putting this slope equal to U_{in}/I_{lm} or larger, the value of this power loss is found as:

$$(P_{tr})_F \geq U_{in}I_{lm} + I_{lm}^2(R_1 + R_p).$$

The full-load efficiency of the power supply is:

$$\eta = \frac{U_{in} I_{lm}}{U_{in} I_{lm} + (P_{tr})_F + I_{lm}^2(R_1 + R_p)} \leq \frac{1}{2 \left(1 + \frac{R_1 + R_p}{R_{in}} \right)},$$

with $R_{in} = U_{in}/I_{lm}$. Obviously η is always less than $\frac{1}{2}$. This low value of η also applies when the power supply is not fully loaded, the efficiency being equal to U_{in}/U_u .

Combined methods

The advantages and disadvantages of the protection methods mentioned above have led to circuits in which several systems have been combined. The characteristic of a commonly used hybrid circuit, consisting of a current limiting system and a cut-out system, is illustrated in *fig. 8*. For a continuous decrease of the load resistance, the current is first limited to the maximum permissible value I_{lm} . As has been explained above the power loss in the control transistor now increases. This increase is however limited in this case, the power supply being cut out when the output voltage has dropped by a certain amount ΔU_1 .

Another hybrid protection method is the combination of a current limiting system and a system with constant power loss in the control transistor. The corresponding characteristic is shown in *fig. 9*. Although with the protection methods of *figs. 8* and *9* the power dissipation in the control transistor may exceed the value in fully loaded condition of the power supply, the extra amount of power loss is limited and the cooling measures required are less severe than with a pure current-limiting system.

Yet another system is the one whose characteristic is shown in *fig. 10*. Here after passing the full load point F , the voltage and current decrease, following a line between the constant current and the constant power loss systems. Here again the power loss in the control transistor increases when the power supply is overloaded, but the increase is less than the one occurring with a pure current-limiting system.

Apart from protection of the power supply and the equipment connected to it, circuits designed for the constant current method of *fig. 4* have another important feature. As supply units with *voltage* stabilization have a very low internal resistance it is not possible to obtain a large current output by connecting several of these units in parallel: obviously small differences in the voltages of the units would cause the load current to be very unequally distributed over the power supplies. There could even be a reversal of the current in one or more of the power units. When however the

supply units act as *constant current sources*, no such objections can be raised to connecting several of them in parallel: the total load current being properly distributed over the supply units in this case.

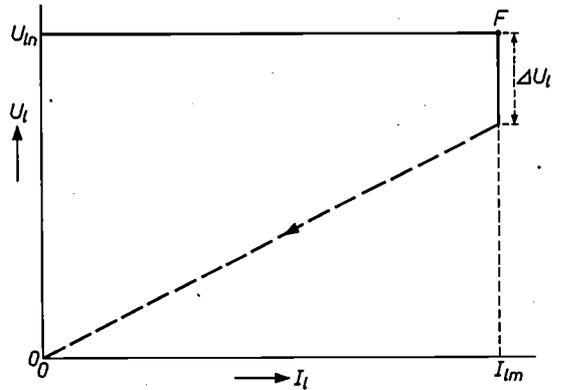


Fig. 8. Characteristic of a hybrid system. When the full-load point F has been reached the voltage first decreases, the current staying constant. At a certain value ΔU_1 of the voltage drop the equipment is cut out.

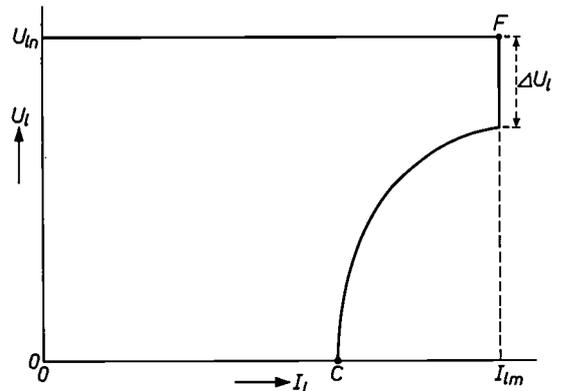


Fig. 9. Hybrid system combining a constant-current circuit and a constant-power-loss circuit.

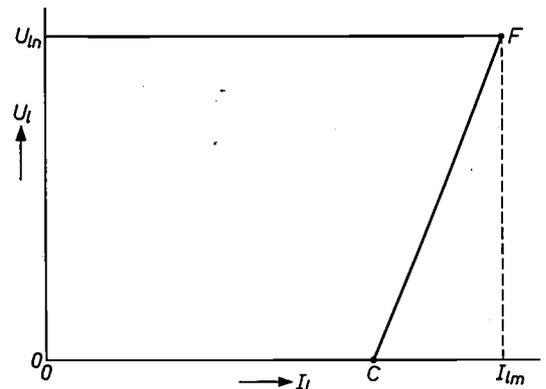


Fig. 10. Protection by a circuit with a characteristic which lies between the constant-current and the constant-power-loss systems.

Examples of protection circuits

The general remarks made above will now be illustrated by a brief discussion of two circuits for protection systems used in Philips power supply units.

Combined current-limiting and cut-out system

In fig. 11 a simplified circuit diagram is shown of the Philips power supply unit PE 4870. Fig. 12 shows a photograph of this unit with its cover removed. The output voltage is adjustable between 0.5 V and 60 V. The control transistor is again denoted by Tr_1 . Instead of one single transistor a cascade circuit of three transistors $Tr_{1...3}$ is used. This arrangement is used because such a circuit has a larger transconductance than Tr_1 alone, resulting in a better regulation of the output voltage against variations due to variations in the load current. The supply voltage for this circuit is delivered by the auxiliary (unstabilized) rectifier V_1 .

The control amplifier A is a two stage difference amplifier (transistors $Tr_{4...7}$). Its power is supplied by a second auxiliary (stabilized) rectifier, V_2 . The

reference voltage U_T is obtained from the voltage across the series circuit formed by the Zener diode Z and the diodes D_1 and D_2 . The diodes D_1 and D_2 are included to compensate for variation of the Zener voltage with temperature.

The protection circuit P contains two transistors, Tr_8 and Tr_9 , a diode D_3 and a few resistors and capacitors. Normally Tr_8 is non-conducting. If however the output current exceeds a certain value this transistor starts to conduct because of the increasing voltage on R_p . The collector of Tr_8 is connected to the base of the transistor Tr_6 in the amplifier. In this way the transistor Tr_1 is controlled and a further increase in output current is prevented. The value at which this current stabilization takes place can be adjusted by means of potentiometer R_1 to between 10% and 110% of the nominal output current of the power pack; rather sensitive loads (e.g. transistor circuits) can therefore be protected. When the output voltage has decreased to a certain value, the transistor Tr_9 starts to conduct. The additional voltage drop now occurring on R_2 causes a further decrease of

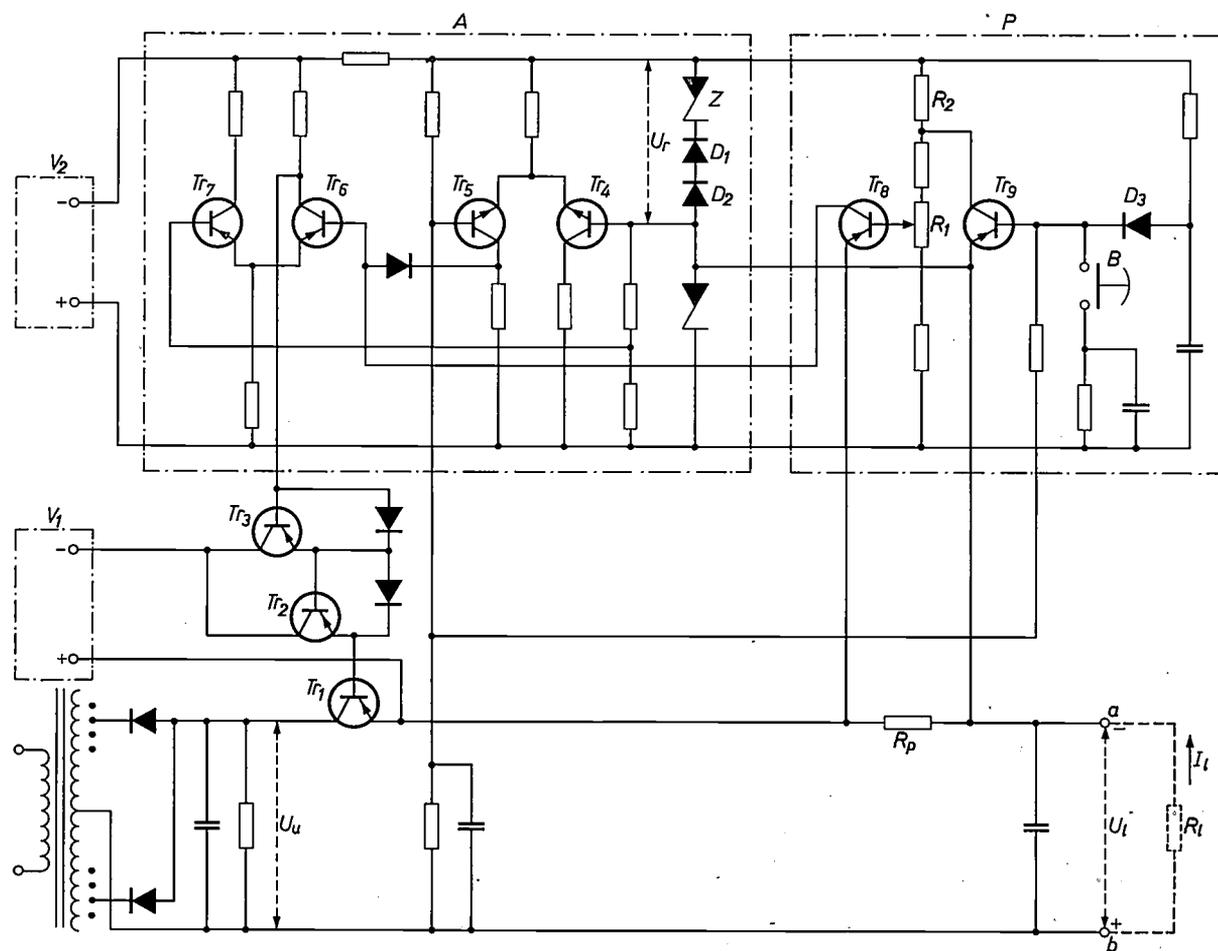


Fig. 11. Simplified circuit diagram of the Philips power supply unit PE 4870. A is a difference amplifier. P is the protection circuit which operates according to the hybrid system shown in fig. 8. The equipment can be put into operation again by pushing the button B . V_1 and V_2 are auxiliary voltage sources (separate rectifiers); V_2 is itself stabilized.

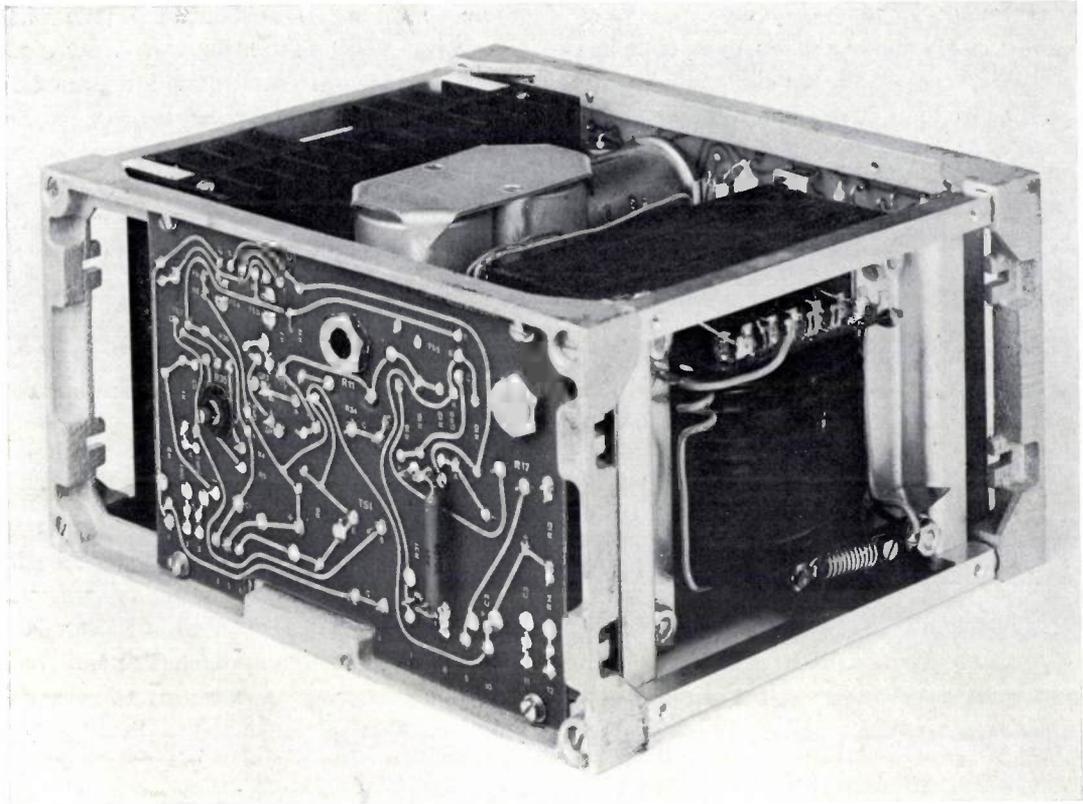


Fig. 12. The Philips power supply unit PE 4870 with cover removed. The cooling fins of the control transistor can be seen at the rear.

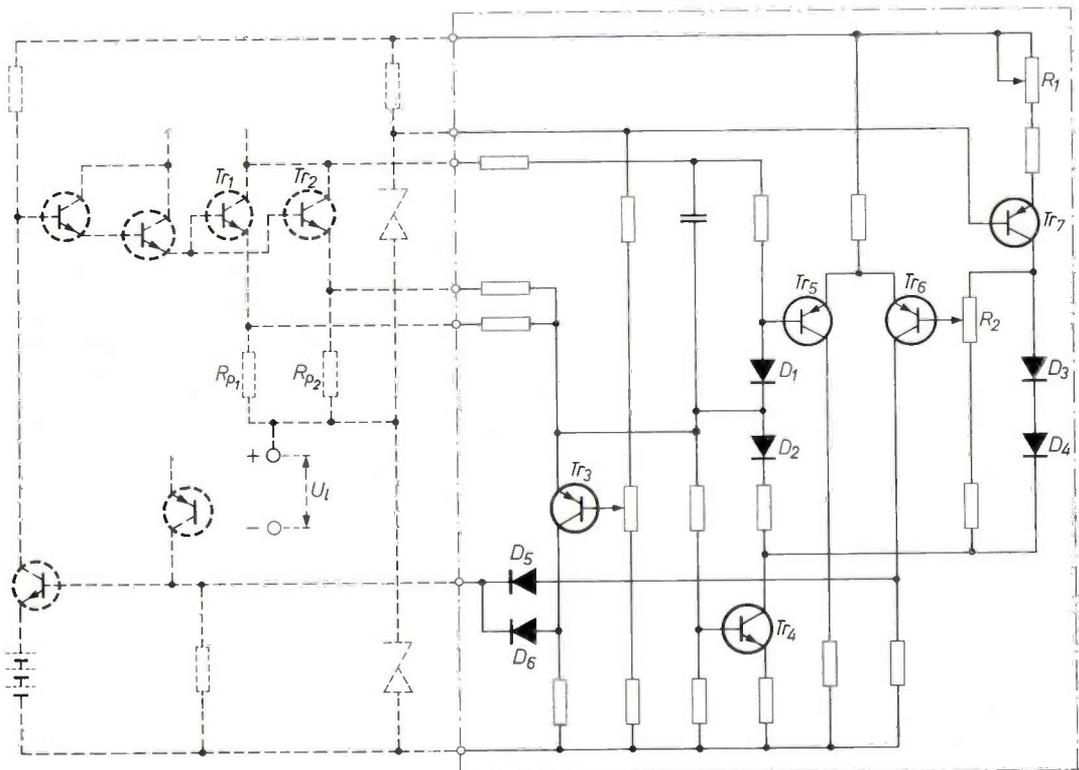


Fig. 13. Circuit diagram of the Philips protection unit PE 4891, drawn together with a part of the circuit of the Philips power supply unit PE 4868, to which it can be provided as an additional unit. The circuit operates according to the system illustrated in fig. 9 (constant-current combined with constant-power-loss). Current-limiting is brought about by the transistor Tr_3 . The constant-power-loss system is driven by the currents in the diodes D_1 and D_2 which have logarithmic characteristics.

the current in Tr_8 , thus decreasing the output current. This brings about an avalanche effect and the equipment is eventually cut out. The cut-out condition can be terminated by pushing the button B , which brings Tr_9 into the cut-off state for a moment. The same effect can be obtained by switching the equipment off and on.

Combined current-limiting and constant-power-loss system

The protection unit now to be described, type PE 4891, can be provided as an additional unit to the Philips power supply PE 4868 (voltage 0.5-30 V, current 3 A). The circuit diagram has been drawn in *fig. 13* together with a part of the circuit of the power supply mentioned.

In this power supply two control transistors Tr_1 and Tr_2 in parallel are used as well as two series resistors, R_{p1} and R_{p2} . The current-limiting process is obtained by the transistor Tr_3 in almost the same way as in the unit previously described. The regulation with constant power loss in the control transistors is brought

about by the diodes D_1 and D_2 . The current in D_1 is proportional to the collector-emitter voltage of Tr_1 and Tr_2 , whereas the current in D_2 (originating from the collector current of the transistor Tr_4) is proportional to the output current. As the characteristics of semiconductor diodes are logarithmic over a certain range, the voltage across the two diodes in series is proportional to the product of voltage and current of the control transistors, i.e. to the power dissipated in them.

The temperature dependance of the diodes D_1 and D_2 is compensated by the use of two other diodes D_3 and D_4 and a difference amplifier containing the transistors Tr_5 and Tr_6 . If large variations in temperature can occur, these elements should be housed in a constant temperature enclosure. In the unloaded condition of the power supply the current in D_2 is set to zero with the aid of Tr_7 and R_1 , and at the same time Tr_6 is brought into the cut off condition by means of R_2 . When the total voltage on D_1 and D_2 exceeds a certain value, Tr_5 is cut off and Tr_6 starts to conduct, applying the voltage via the diode D_5 to the output stage of the control amplifier. As the diode D_6 is now biased to cut-off, further action of the current-limiting transistor Tr_3 is prevented and the load current is decreased to a value giving a constant power loss in the control transistors. *Fig. 14* shows the voltage-current characteristic obtained. This is of the type discussed with *fig. 9*, providing a compromise between a current-limiting system and a constant power loss system.

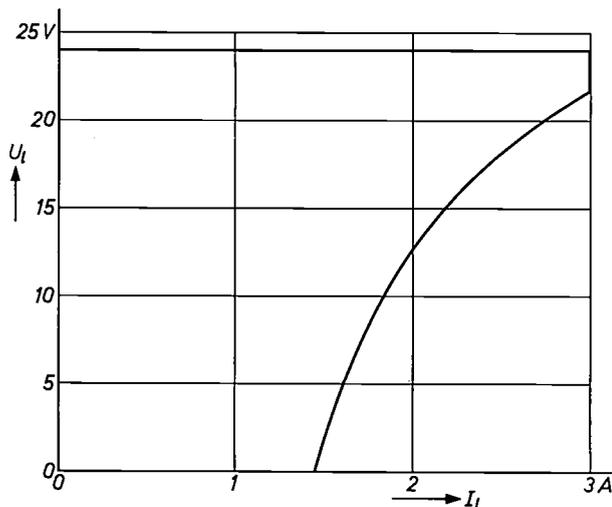


Fig. 14. Voltage-current characteristic of the Philips power supply unit PE 4868, equipped with the protection unit PE 4891.

Summary. In transistorized d.c. power supply equipment electronic circuits are needed for safeguarding the supply units and the circuits connected to them against damage due to overloading or short-circuiting. Systems which cut off the power supply when overloading occurs can be made up from simple electronic circuits, but such systems have several disadvantages in practice. If a pure current-limiting system is used, there is a large additional power dissipation in the control transistor. Circuits have therefore been developed which hold the power loss in the control transistor to a constant value. Hybrid circuits of these systems have also been made. Two examples of these circuits, as used in Philips equipment, are described.

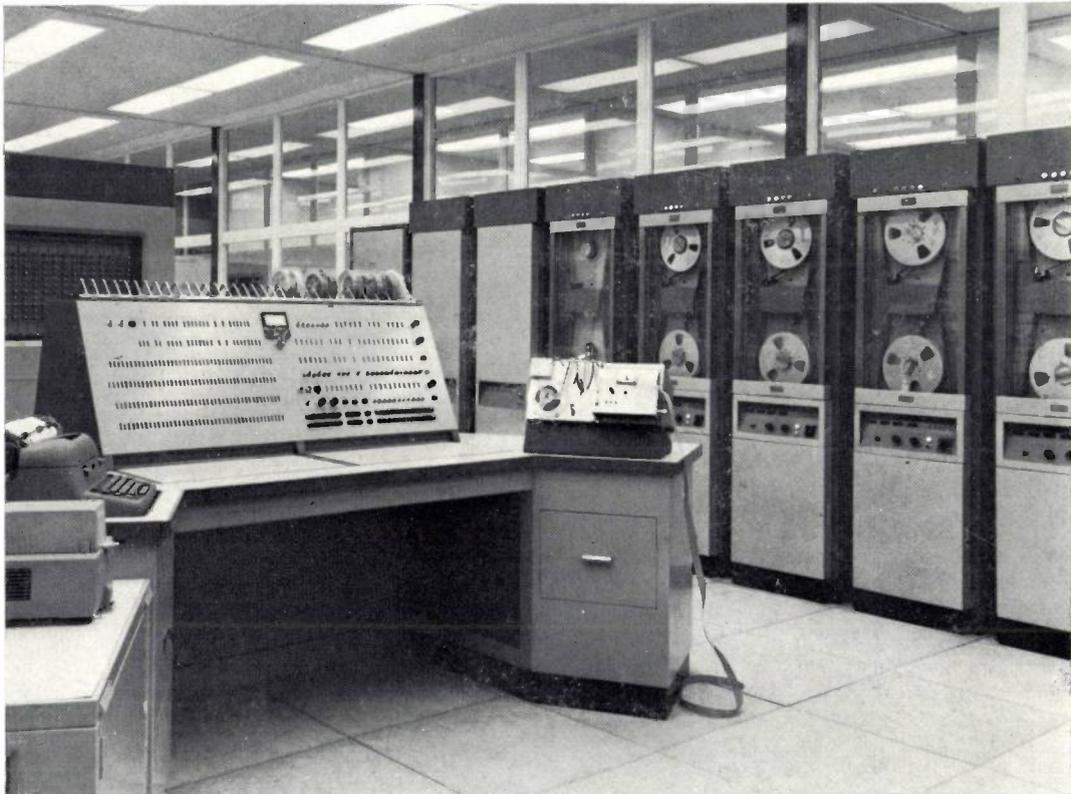
Recent scientific publications

These publications are contributed by staff of laboratories and plants which form part of or co-operate with enterprises of the Philips group of companies, particularly by staff of the following research laboratories:

Philips Research Laboratories, Eindhoven, Netherlands	<i>E</i>
Mullard Research Laboratories, Redhill (Surrey), England	<i>M</i>
Laboratoires d'Electronique et de Physique Appliquée, Limeil-Brevannes (S.O.), France	<i>L</i>
Philips Zentrallaboratorium GmbH, Aachen laboratory, Weisshausstrasse, 51 Aachen, Germany	<i>A</i>
Philips Zentrallaboratorium GmbH, Hamburg laboratory, Vogt-Kölln-Strasse 30, 2 Hamburg-Stellingen, Germany	<i>H</i>
MBLE Laboratoire de Recherche, 2 avenue Van Becelaere, Brussels 17 (Boitsfort), Belgium.	<i>B</i>

Reprints of most of these publications will be available in the near future. Requests for reprints should be addressed to the respective laboratories (see the code letter) or to Philips Research Laboratories, Eindhoven, Netherlands.

- P. Billard, J. Donjon & G. Marie:** Application de la modulation de lumière aux télécommunications. *Acta electronica* **9**, 305-313, 1965 (No. 4). *L*
- G. Blasse:** On the Eu^{3+} fluorescence of mixed metal oxides, IV. The photoluminescent efficiency of Eu^{3+} -activated oxides. *J. chem. Phys.* **45**, 2356-2360, 1966 (No. 7). *E*
- G. Blasse & A. Bril:** On the Eu^{3+} fluorescence in mixed metal oxides, III. Energy transfer in Eu^{3+} -activated tungstates and molybdates of the type Ln_2WO_6 and Ln_2MoO_6 . *J. chem. Phys.* **45**, 2350-2355, 1966 (No. 7). *E*
- G. Blasse & A. Bril:** Broad band u.v. excitation of Sm^{3+} -activated phosphors. *Physics Letters* **23**, 440-441, 1966 (No. 7). *E*
- J. Donjon & G. Marie:** Modulateurs de lumière à large bande utilisant l'effet Pockels. *Acta electronica* **9**, 315-385, 1965 (No. 4). *L*
- J. A. Geurst:** Theory of insulated-gate field-effect transistors near and beyond pinch-off. *Solid-State Electronics* **9**, 129-142, 1966 (No. 2). *E*
- W. Kischio:** Halbleitendes Cadmium- und Zinkdiphosphid. *Z. Naturf.* **21a**, 1733-1734, 1966 (No. 10). *A*
- T. Klein:** Contour deposition — a new epitaxial deposition technique for semiconductor devices and integrated circuits. *Solid-State Electronics* **9**, 959-966, 1966 (No. 10). *M*
- J. R. Mansell:** Travelling-wave phototube theory. *Int. J. Electronics* **20**, 467-488, 1966 (No. 5). *M*
- B. J. Mulder:** Photo-injection of electrons into anthracene from electrolytic electrodes. *Solid State Comm.* **4**, 615-617, 1966 (No. 11). *E*
- R. F. Pearson & R. Cooper:** Magnetic crystals that manipulate light. *New Scientist* **32**, 92-94, 1966 (No. 517). *M*
- G. W. Rathenau:** Einige Gesichtspunkte zur Wechselwirkung in nichtleitenden magnetischen Kristallen. *Z. angew. Physik* **21**, 277-282, 1966 (No. 4). *E*
- H.-D. Rüpke:** Über die Ankopplung von Mikrowellenresonatoren zur Messung von Materialeigenschaften. *Archiv elektr. Übertr.* **20**, 617-620, 1966 (No. 11). *H*
- P. C. Scholten:** Indium contacts on CdS. *Solid-State Electronics* **9**, 1142-1143, 1966 (No. 11/12). *E*
- E. Schwartz:** Empirische Synthese verlustloser, symmetrischer Zirkulatoren. *Archiv elektr. Übertr.* **20**, 621-625, 1966 (No. 11). *A*
- P. J. W. Severin & A. G. van Nie:** A simple and rugged wide-band gas discharge detector for millimeter waves. *IEEE Trans. on microwave theory and techniques* **MTT-14**, 431-436, 1966 (No. 9). *E*
- J. G. Siekman:** Lassen met een elektronenbundel. *Ingenieur* **78**, O 99-108, 1966 (No. 47). *E*
- T. L. Tansley:** Heterojunction boundary conditions. *J. appl. Phys.* **37**, 3908, 1966 (No. 10). *M*
- D. J. Vinney:** Possible travelling-wave parametric amplifier using the Gunn effect. *Electronics Letters* **2**, 357-358, 1966 (No. 10). *M*
- J. Volger:** Progress in superconductivity. *IEEE Trans. on magnetics* **MAG-2**, 159-164, 1966 (No. 3). *E*



A high-speed punched-tape reader

J. M. Visscher

Just as the computing speed of the modern electronic computer is far greater than that of the human brain, the reading speed of the punched-tape reader used with the computer is very much greater than human reading speed. If a person were able to read 2000 characters per second, like the punched-tape reader described here, he could get through the average paperback in less than 3 minutes. The design of a punched-tape reader which can read at this high speed, and yet has provision for stopping the tape very quickly, has been achieved by an ingenious combination of optical, mechanical and electronic techniques.

The punched tape and the punched card are the conventional media for conveying information to an electronic computer. A punched tape consists of a strip of material, usually paper, in which holes are punched at positions which form a rectangular pattern; see fig. 1. The rows of positions in the longitudinal direction

are called the *channels* of the tape; a combination of punched holes in a row across the tape represents a letter, figure or instruction in accordance with a particular code, and is called a *character*. In addition to the channels (the maximum number of channels is eight) the tape contains a continuous row of smaller holes, called sprocket holes. In slow tape readers these holes can be used for driving the tape, but they are used here only for indicating the position of the characters.

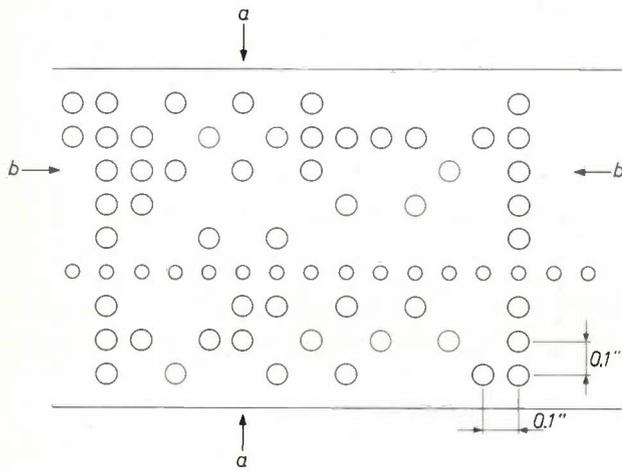


Fig. 1. Part of an eight-channel punched tape. The combination of holes in a column across the tape (e.g. *a-a*) constitutes a character. A row of places for holes along the length of the tape (for instance *b-b*) is called a channel. The small holes are the sprocket holes.

can be started or stopped in one of the following ways:

1) The tape stops for the reading of each character. The next character is not read until the computer has processed the information and the tape has moved on by one character space.

2) A number of characters are read together, in blocks, and the tape is then stopped for the characters to be processed. After this the next block is read, and so on.

3) The tape is fed continuously through the reader. It must be remembered here that the times needed for processing different characters may be very different.

To achieve a reasonable tape speed the reader must be designed in such a way that the transport time per character is a great deal shorter than the longest of these processing times. As a result, however, it may happen that the computer is still processing a certain character while the reader is already reading the next one. To avoid such a situation, one of the following measures has

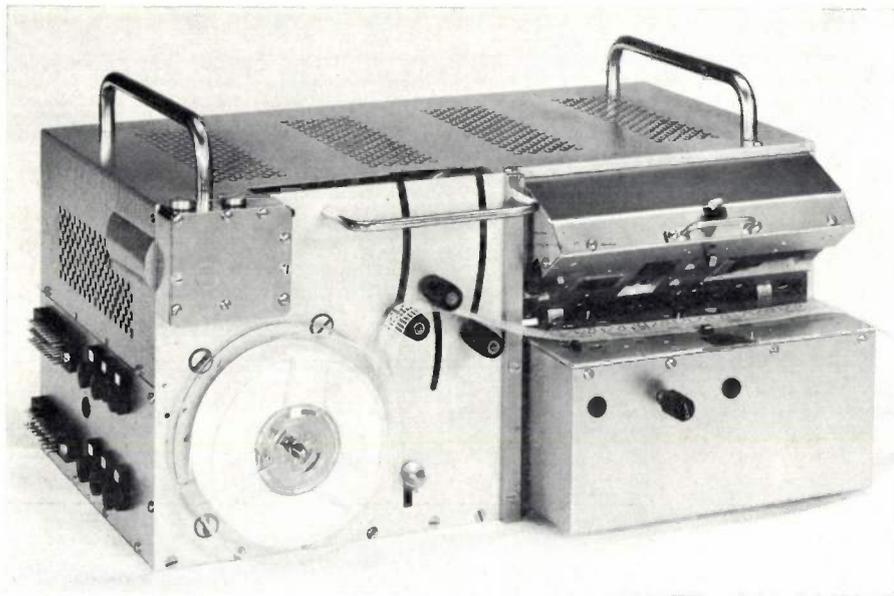


Fig. 2. The tape reader described, opened for loading the tape.

The instructions contained on the tape are passed to the computer via the punched-tape reader. In doing this, it has to transport the tape, "read" the information character by character and supply it to the computer. The characters can be read mechanically, by means of feeler pins which complete a circuit when a hole passes; capacitively, by making use of the difference between the dielectric constants of paper and air; or optically, by making use of the light-absorbing or light-scattering properties of the paper.

The computer processes the information read off the tape and at the same time derives from it the start-stop signal for the transport of the punched tape. The tape

to be adopted: a) the reader is equipped with a buffer store, in which one or more characters can be temporarily stored; b) measures are taken enabling the punched tape to be stopped very quickly within the limits of a single character which thus remains in the reading position ("stopping on a character").

A few years ago, when the development of a punched-tape reader at Philips Research Laboratories was started, it was clear that the new reader would have to be able to read at least 1200 characters per second and would also have to be adaptable to computers already in use or under development in the Company. The system with a continuously moving tape was adopted,

as this arrangement offered the highest tape speeds. Stopping on a character also had to be possible, because this removed the need for a costly electronic buffer store and moreover permitted adaptation to all possible start-stop signals. The following requirements were also taken into account:

- 1) A tape starting-time as short as a few milliseconds to give high start-stop frequencies if required.
- 2) A tape spooling system. No take-up reel is required; the tape is simply deposited in a container. The tape then has to be rewound before reading out a second time — but this is always necessary anyway as a punched tape can only be read in one direction.
- 3) A double read-out facility, in order to show up reading errors.
- 4) Continuously variable tape speed.
- 5) Simple tape loading.

To meet these requirements, the punched-tape reader

now stop the tape as quickly as possible before the new character passes right through. When the “not-ready” signal finishes, the computer again sends a “start” signal to the reader.

Operation of the tape reader

The punched-tape reader has the following main sections:

- 1) The optical reading section.
- 2) The spooling section.
- 3) The driving system.
- 4) The braking system.

With reference to *fig. 4*, which shows a diagram of the various main sections, we shall first discuss the general principles of the tape reader.

- 1) In the light path of the optical reading system there is a mask which has holes at each of the places corresponding to the holes which make up a character, in-

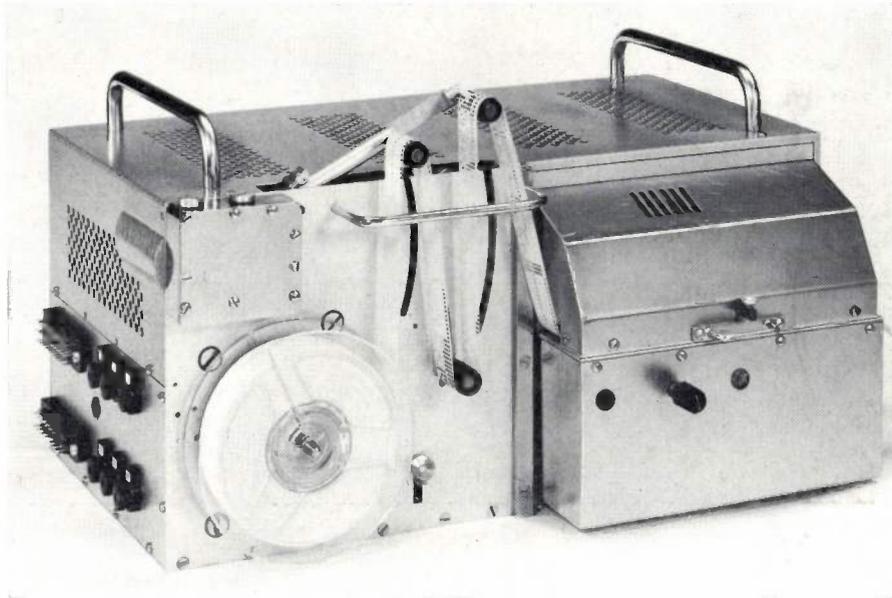


Fig. 3. The tape reader just before being put into operation. The tape has not yet been tensioned.

shown in *figs. 2 and 3* has been developed. This tape reader, which has now been in use for some years will be dealt with here.

Before describing the actual construction, it will be useful to consider in somewhat more detail the interplay between reader and computer when the tape stops on a character. During the processing of a character the computer continuously produces a “not-ready” signal. When the sprocket hole for the next character appears in the reader, the reader gives a pulse-shaped “arrival” signal. If both signals are present simultaneously, the computer derives a “stop” signal from this information and transmits it to the reader. The reader must

including the sprocket hole. The holes in the mask have the same dimensions as those in the tape. Each mask hole has a light-detector behind it. As the punched tape is transported, the light path always shows the same pattern as the perforations on the tape, so that all the characters are processed one after the other. To be able to observe the arrival of a new character in good time, and to allow as long a read-out time as possible, then for each detector, including the one for the sprocket holes, the ratio of the output current with the hole fully uncovered to the output current with the hole covered must be high. It has been found that at a stopping distance of 1 mm (i.e. about 90% of the diam-

eter of the sprocket holes) no reading errors occur if this ratio is 20 : 1.

2) The spooling section is designed so that the punched tape being fed from the reel to the actual reading section is looped on a system of adjustable rollers. This prevents delay due to the inertia of the reel when starting. When the braking is applied, too much tape is delivered, because of the inertia of the reel, and the surplus tape is taken up by an increase in the length of the loops between the rollers.

3) and 4) The tape is driven by being pressed against a permanently rotating driving roller R_d . This is done by means of a pressure roller R_p controlled by an electromagnet M_p . The tape is stopped with the aid of a very

ture of the braking magnet is completely released.

To permit stopping on a character we have allowed a stopping distance of 0.5 mm at the most. If the light-to-dark current ratio has the value quoted above, this distance gives a safety factor of about 2. The tape speed is about 3 m/s at a reading speed of 1200 characters per second. This means that to achieve a stopping distance of 0.5 mm, a deceleration of 9000 m/s² is required (assuming constant deceleration during the braking process). By way of illustration, if this deceleration were applied to a car travelling at 100 km/h (62.5 m.p.h.) the stopping distance would be only 40 mm. The rapid braking required for the punched-tape reader is obtained by letting the armature of the

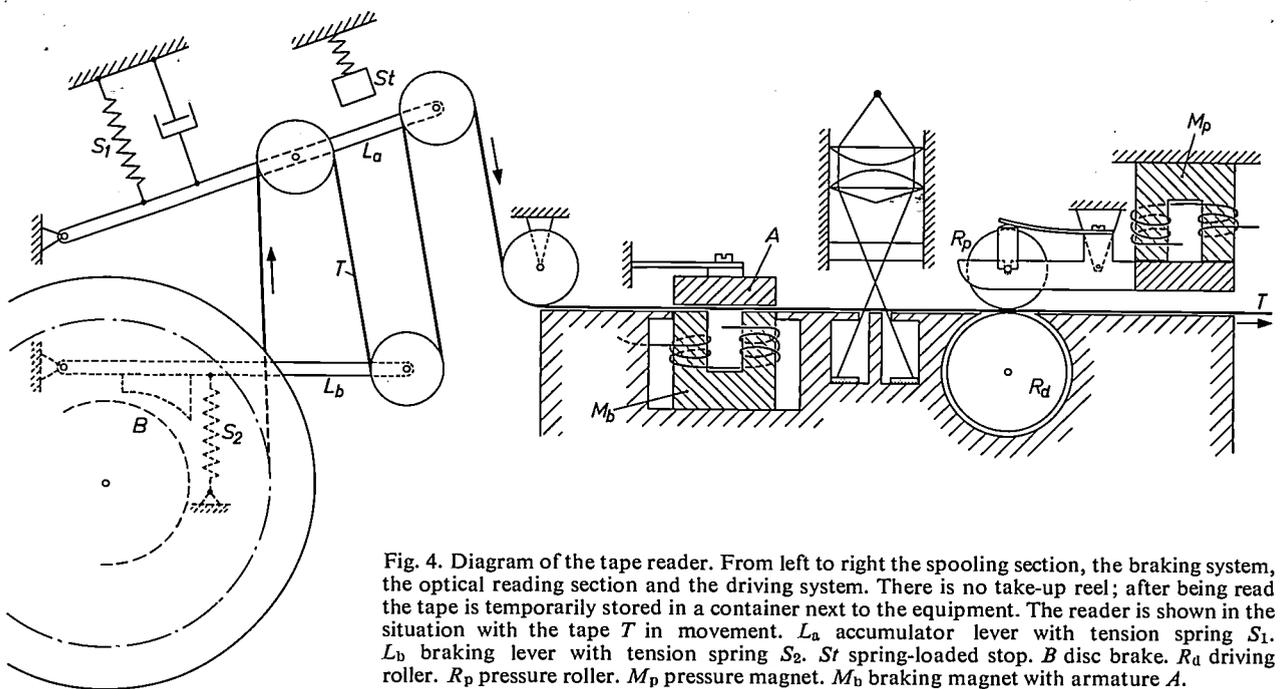


Fig. 4. Diagram of the tape reader. From left to right the spooling section, the braking system, the optical reading section and the driving system. There is no take-up reel; after being read the tape is temporarily stored in a container next to the equipment. The reader is shown in the situation with the tape T in movement. L_a accumulator lever with tension spring S_1 . L_b braking lever with tension spring S_2 . St spring-loaded stop. B disc brake. R_d driving roller. R_p pressure roller. M_p pressure magnet. M_b braking magnet with armature A .

fast magnet, the braking magnet M_b , the tape being passed through the air gap of this magnet. The "stop" signal energizes the braking magnet, while at the same time M_p is de-energized. As there is a longer mechanical delay in raising the pressure roller than in the attraction of the armature A of the braking magnet, it must be possible for the driving roller to slip on the tape to prevent the tape from breaking.

Upon starting, M_p is energized and M_b is de-energized. If these two events occurred at the same time there would be a danger that the pull exerted on the tape by the spooling system could cause the tape to slip backwards after the braking magnet has been switched off and before the pressure roller has made contact with the tape. To prevent this, the energizing of M_b is cut off with an electrical time delay, so that the pressure roller is already on the tape before the arma-

braking magnet slide permanently over the tape and by designing both this magnet and its control circuit for extremely rapid energization. The braking magnet, the reading system and the driving roller can be seen in *fig. 5*, which shows part of the reader in the open position.

After this general description we shall now examine in greater detail the construction of the main sections and the operation of the complete equipment.

The optical reading section

Since the absorption coefficient of the tape paper is low, we make use of the light-scattering properties of this paper to achieve the high light-to-dark ratio required. The optical system is designed to give a sharp image of the lamp filament immediately above a mask hole (see *fig. 6*), the width of the image being less than

the diameter of the hole. The cavity in which the photoelectric cell C — one for each mask hole — is located is made "optically black" and the beam of light is made just wide enough to cover the photoelectric cell. Now if α is the solid angle of the beam in steradians, and E_1 is the total luminous flux with open mask hole in lumens, then taking the absorption coefficient of the paper as 0.5, and assuming equal scattering of the light in all directions, the luminous flux E_2 reaching the photocell with mask hole covered is:

$$E_2 = \frac{1}{2} \frac{\alpha}{2\pi} E_1,$$

so that the light-to-dark ratio E_1/E_2 is equal to $4\pi/\alpha$. This ratio can thus be given a high value by making the solid angle of the beam very small (and suitably adapting the arrangement of the photocell).

The light-to-dark ratio achieved with our arrangement (see fig. 7) is 70 : 1. The filament of a single lamp supplies the beams of light for all the nine places of a character. The image of the filament is obtained by means of a condenser $Cond$ and a lens L . Underneath the condenser there is a prism P which produces two identical beams for reading two characters simultaneously. The lenses are all cylindrical and made of acrylic resin, permitting manufacture on a precision lathe.

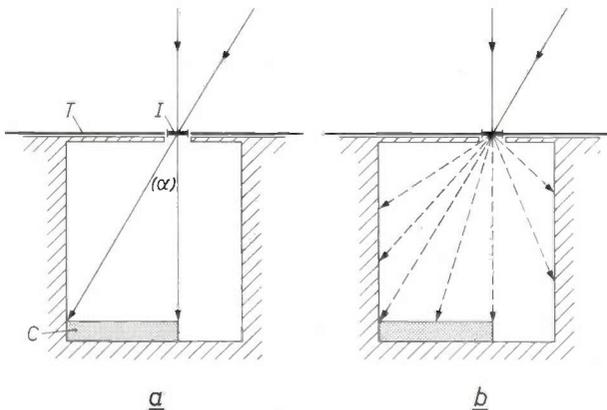


Fig. 6. Arrangement of the photocells in the optical reading section. *a*) Situation for one mask hole, uncovered by a tape hole. *b*) Situation where the mask hole is covered, so that the tape scatters the light in all directions (opening angle 2π steradians). The walls of the space containing the photocell C are made optically black. α solid angle of the beam of light. I image of the filament. T tape.

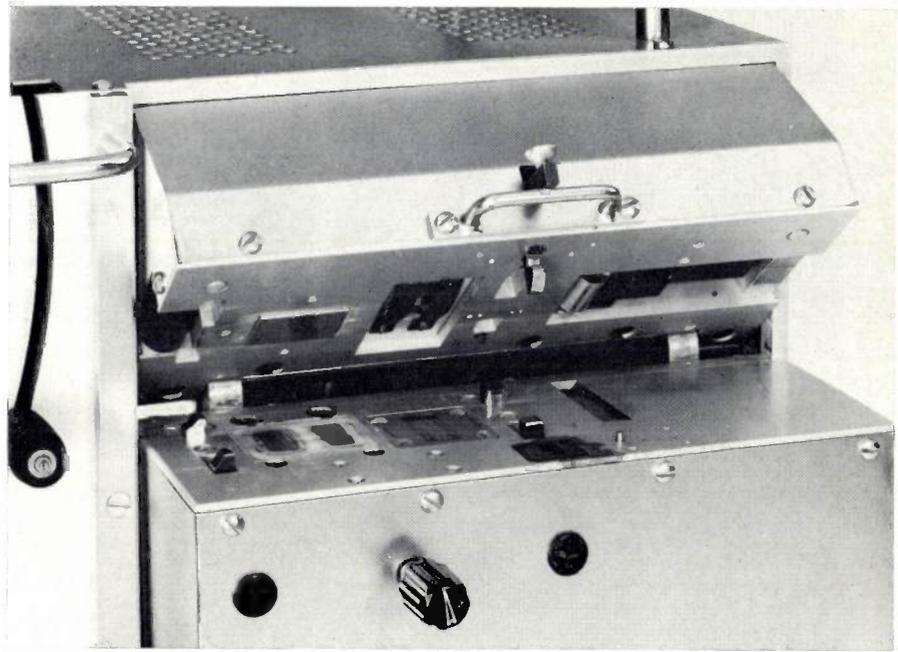


Fig. 5. The tape reader with the lid up. From left to right can be seen the braking magnet M_b , the reading system, the driving roller R_d and the pressure roller R_p .

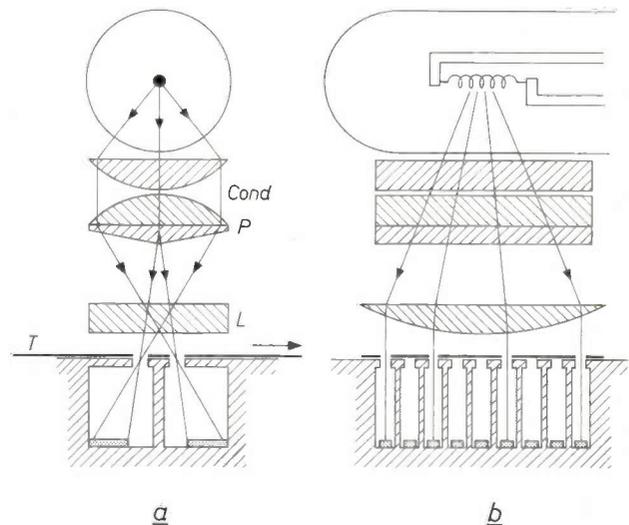


Fig. 7. Construction of the optical reading section, showing the path of the light rays. *a*) Longitudinal cross-section. *b*) Transverse section. As can be seen at the left the light rays are split into two beams by a prism P for double optical read-out. The condenser $Cond$ and the lens L are both cylindrical.

The spooling section

When the reader is in the quiescent state, the braking lever L_b is in its lowest position (fig. 4); the reel is then braked by the disc brake B . The "accumulator" lever L_a is in its upper position, resting against the spring-loaded stop St . The tape describes several loops around the rollers on the levers.

When the tape reader is started, the tape very quickly reaches the required speed. The tension in the tape causes the braking lever L_b to rise, which releases the

disc brake B sufficiently to allow the reel to start turning. The reel has much more inertia than the tape, with the result that as it begins to rotate the loops become smaller because L_a descends and L_b rises. When steady running has been achieved, the torques operating on L_b due to the tension of spring S_2 , the pressure of the disc brake and twice the tension in the tape are in equilibrium; this has the effect of keeping the tensile force in the tape approximately constant. Lever L_a takes up a position such that the pull of the tape, acting upon it at four points, is counterbalanced by the pull of the tension spring S_1 . The movement of L_a is damped by a kind of shock-absorber to stabilize to some extent the circumferential speed of the reel. The lever system is designed so that the roller of L_b can be moved between the two rollers of L_a , which simplifies the insertion of a new tape; see fig. 2.

When the tape is stopped, its tension falls; lever L_b therefore descends and the disc brake B stops the reel. In the meantime lever L_a moves a certain distance in the direction of its starting position so that the surplus tape run off the reel is taken up by the increase in the length of the loops between the rollers.

The driving system

Fig. 8a shows the driving system in greater detail, with the electromagnet M_p in the non-energized state. The driving roller R_d turns continuously at a circumferential speed v_0 . In this situation the tape is stationary, however, because the pressure roller is not in contact with it. Although the leaf spring S_3 pushes the shaft-ends of the pressure roller R_p downwards by means of the pressure piece Pr , the shaft-ends are restrained by the base of the slots which form their bearings in the armature. The pressure of the leaf spring is therefore taken up by the armature and not by the driving roller

R_d . A slightly extended tension spring S_4 pulls the armature upwards against a stop placed at such a height that the pressure roller is held just out of contact with the tape T .

When the reader receives a starting pulse from the computer, M_p is energized and we have the situation of fig. 8b. The pressure roller shaft-ends are now clear of the bottom of the slots, so that the leaf spring can force the pressure roller against the tape. When the braking magnet is de-energized, the tape is carried forward by the rotating driving roller and the pressure roller also begins to rotate. The tape is thus driven entirely by means of the rather small frictional forces (about 3.5 N) between the driving roller and the tape; we have not used the usual system in which the reel is motor driven.

Electrical and mechanical delay are both encountered on starting the driving system. The first occurs in energizing M_p ; the method used for this enables the saturation magnetization to be reached in a short time (this method will be discussed when we come to the braking magnet). The mechanical delay which occurs in swinging the armature through the angle required can be calculated approximately by assuming that the accelerating force is the magnetic attractive force of the armature less the force due to S_4 . The moment of inertia used is that of the armature plus the pressure roller, and the appropriate angle is the angle between the release of the armature and the making of contact between pressure roller and tape. The total delay time (electrical plus mechanical) is $1\frac{1}{2}$ ms. The de-energizing of the braking magnet must be given the same delay to prevent the tape from being dragged backwards; the tape thus starts to move $1\frac{1}{2}$ ms after the starting pulse.

An idea of the rapidity with which the tape thus picks

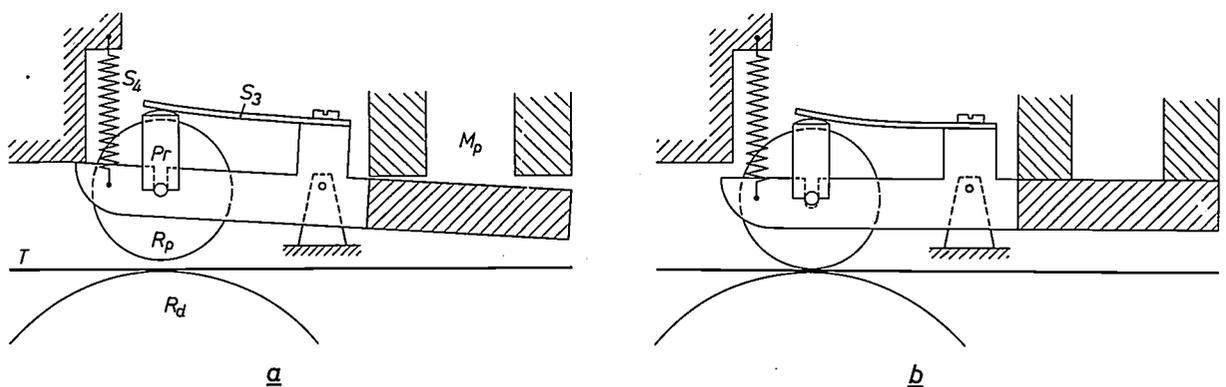


Fig. 8. The driving system in more detail. a) The magnet M_p is not energized. The force of the leaf spring S_3 is taken up via the pressure piece Pr by the armature itself, while the pressure roller R_p is held free from the tape T by the tension spring S_4 . b) The energizing of the magnet has the effect of releasing the shaft-ends of roller R_p , so that this roller is pressed by the leaf spring S_3 against the tape T and the driving roller R_d .

up speed can be obtained from *fig. 9*. This figure shows the output current from the sprocket track photocell as a function of time (the current waveform has been "squared off" by a pulse shaper). It can be seen that if a sprocket hole is visible in the starting position, the next sprocket hole appears after 6.5 ms (i.e. 5 ms after the tape began to move) and that the tape has reached its full speed of 1200 characters/s after 15 ms. At full speed a character passes in 0.83 ms.

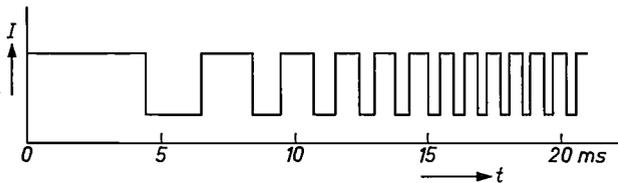


Fig. 9. Output current of the photocell that scans the sprocket holes, as a function of time, during the acceleration of the tape from rest to a speed of 1200 characters per second (corresponding to a tape speed of 3 m/s).

The tape can be started as quickly as this because the only factors of significance are the mass of the tape and the moments of inertia of the levers and the rollers on these levers.

We shall now consider the behaviour of the driving system when the tape is stopped. When the computer delivers a "stop" signal, the braking magnet M_b is energized and the magnet M_p de-energized. After the "stop" signal appears a further $1\frac{1}{2}$ ms is required before the pressure roller comes free of the tape. This time includes both mechanical and electrical delay. The electrical delay is about 1 ms, the time required for the magnetic field of M_p to decay. In the remaining 0.5 ms the armature of M_p is raised by a distance equal to the clearance between the pressure roller shaft-end and the base of the slot. This mechanical delay in the driving system when the tape is stopped can be calculated in the same way as the starting delay. This can be done by assuming that the force exerted by the leaf spring S_3 on the armature during this rotation is constant and that the magnetic attractive force drops sharply to zero. The moment of inertia which is used here is that of the armature, in this case *without* taking the mass of the pressure roller into account.

As already remarked, the delay of the braking magnet is very small. A result of this is that the rollers in the driving system are still rotating after the tape has been stopped by the braking magnet. These rollers therefore exert friction forces on the tape, and there is a danger of the tape snapping between the braking magnet and the driving system. This is avoided as follows. The pressure roller and the driving roller are

both made of metal, and an approximate estimate of the coefficient of friction between the tape and these rollers can therefore be obtained. Furthermore the force with which the pressure roller is pressed upon the tape depends only on the dimensions and the deflection of the leaf spring S_3 , and is thus quite constant and reproducible. These measures ensure that the maximum tensile stress which occurs in the tape between the braking magnet and the driving system as a result of the slipping forces cannot assume excessive values in the event of incidental variations. As the pressure roller is made of aluminium it is brought to a stop very quickly by the friction on the stationary tape. The tensile stress in the tape, originally due to both rollers, therefore drops very rapidly to half its initial value. The application of these measures prevents the tape from being broken when it is stopped.

The braking system

The required fast response of the braking system to a "stop" signal from the computer can be achieved only by minimizing the distance to be travelled by the armature of the braking magnet. This is done, as can be seen from *fig. 10*, by keeping the armature of the braking magnet *permanently pressed against the punched tape* under its own weight and the light pressure of a leaf spring S_5 , so that the tape is always kept in sliding contact with the armature and the pole shoes. By

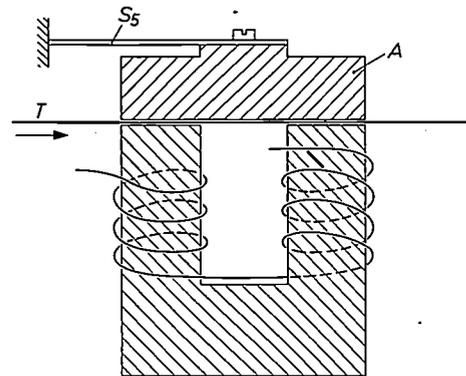


Fig. 10. The braking system. The armature A of the braking magnet, even when not energized, is always pressed against the pole shoes by the action of a light leaf spring S_5 and its own weight, so that the tape T is kept in sliding contact.

interposing very thin plates of piezo-electric material between tape and pole shoes we were able to measure the variation of the downward force on the tape as a function of time during the braking process; the result is shown in *fig. 11*. The variation in pressure on the tape which occurs immediately after the stop signal is applied is seen to be a rapidly damped periodic oscillation. The system must apparently be considered as a

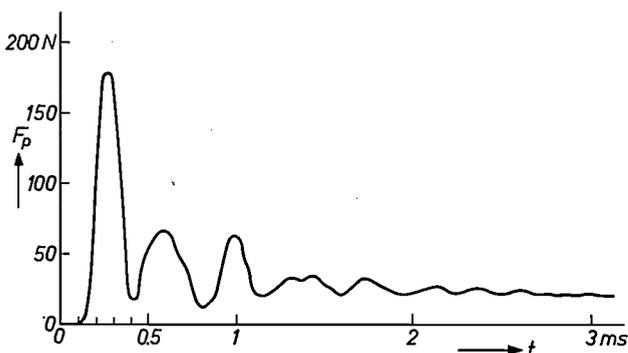


Fig. 11. The braking system force F_p , measured in newtons, as a function of time t (in ms) when the tape is stopped. The measurement was made with piezo-electric plates interposed between the tape and the pole shoes.

sprung-mass system (see fig. 12), the mass being formed by the magnet armature, and the spring action being derived from the elastic deformation of the tape, while the non-elastic deformation causes the damping. If we calculate the frequency of vibrations perpendicular to the plane of the tape (vibrations due to rotation of the armature are also possible), then for a spring stiffness calculated from the static depression of the paper, we find a vibration period of $240 \mu\text{s}$ if damping is not taken into account. From fig. 11 we find a vibration period of $400 \mu\text{s}$. The difference is explained by the uncertainty in the measurement of the spring stiffness, the frequency-reducing effect of damping on a sprung-mass system, and the apparent decrease in area of the pole shoes which occurs because the measuring plates were of smaller area than the pole shoes.

It can further be seen from fig. 11 that the armature starts to move about $110 \mu\text{s}$ after the instant at which the energizing of the braking magnet commences, as the force on the tape then begins to rise. The force reaches its maximum after another similar period. At the end of this article we shall deal at some length with the design of a magnet capable of such rapid action. We shall first calculate the stopping distance from the curve of the force on the tape (fig. 11).

Calculation of the stopping distance

To assist explanation of the calculation we shall first discuss the behaviour of the tape when it is stopped. We have already seen that a constant deceleration of about 9000 m/s^2 would be needed to obtain a stopping distance of 0.5 mm at a tape speed of 3 m/s . However, as fig. 11 shows, we are faced with a delay time t_0 of at least $110 \mu\text{s}$, during which the tape covers a distance of 0.3 mm at this speed. A deceleration of much more than 9000 m/s^2 is therefore necessary, and also a much greater force on the tape than would follow from this deceleration. It is found that when such greater forces are applied with the braking magnet, the tape behaves

quite differently from the way it would behave with purely frictional braking. In particular, we do not find the expected quadratic relationship between tape speed and stopping distance that follows if the energy to be destroyed (the kinetic energy of the tape, $\frac{1}{2}mv_0^2$) is simply equated to the braking force multiplied by the stopping distance. With this fast braking quite another treatment of this problem is required, in which the elastic properties of the tape are taken into account.

If the tape moving at a speed v_0 is suddenly or in a very short time given zero velocity at the moment $t = 0$ at the position of the braking magnet, a shock wave will move along the tape at a speed equal to the speed

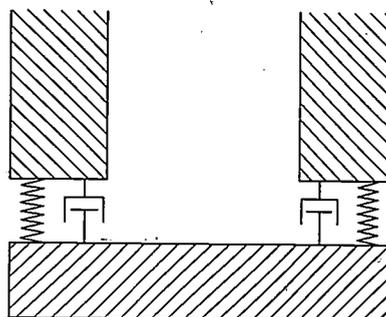


Fig. 12. Diagram of the sprung-mass system formed by the tape and the armature of the braking magnet.

of sound v_s in the tape material; see fig. 13. Between this shock wave and the braking magnet the tape is at rest, but at the other side of the shock wave the tape is in normal movement at the speed v_0 .

The part which is at rest is uniformly stretched as a result of a force F associated with a strain ϵ in the tape, which is given by [1]:

$$\epsilon = \frac{v_0}{v_s} \dots \dots \dots (1)$$

According to Hooke's law we have:

$$F = \epsilon EA, \dots \dots \dots (2)$$

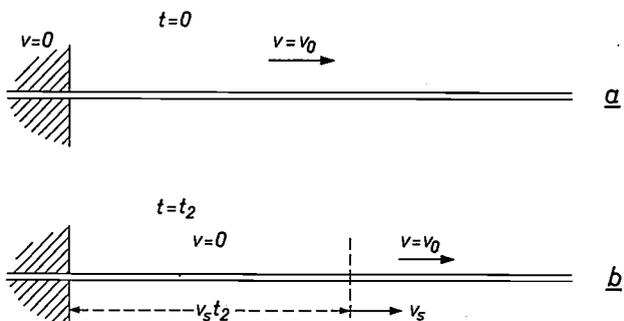


Fig. 13. When a velocity $v = 0$ is imparted to the tape travelling at a speed v_0 at the time $t = 0$ at the position of the braking magnet, a shock wave is generated which travels along the moving tape at a speed v_s . a) Situation at $t = 0$; b) situation after t_2 seconds.

where E is the modulus of elasticity of the tape material and A is the average cross-sectional area of the tape. Further, the velocity of propagation of a longitudinal vibration (i.e. the speed of sound) in a medium is given by:

$$v_s = \sqrt{\frac{E}{\rho}}$$

where ρ is the density of the medium. If we apply this to the tape and introduce the mass of the tape per unit length $m_1 = \rho A$, we find, using (1) and (2):

$$F = m_1 v_0 v_s \dots \dots \dots (3)$$

The minimum value of the tensile force which must be applied in the tape in order to stop it is therefore proportional to the original speed v_0 .

The expression found for F can now be used for calculating the stopping distance as a function of the tape speed, as the tape comes completely to rest as soon as the force in the tape at the braking magnet has reached the value of F given by (3). The total frictional force applied to both sides of the tape by the force F_p due to the armature of the braking magnet must therefore be:

$$2fF_p = F, \dots \dots \dots (4)$$

where f is the coefficient of friction between the braking magnet material and the tape. The force on the tape, represented as a function of time in fig. 11, can be represented by the idealized curve of fig. 14. The instant of stopping, $t = t_1$, is reached when F_p exceeds the value $F/2f$ given by (4). The stopping distance is the distance travelled by the tape between the times $t = 0$ and $t = t_1$.

Up to the instant $t = t_0$ there is no braking force on the tape, and therefore the speed v remains unchanged. The time between t_0 and t_1 is so short that we can disregard the reduction of tape speed due to the rapidly increasing friction during this period; we may therefore state that the stopping distance is approximately equal to:

$$x = v_0 t_1 \dots \dots \dots (5)$$

Having now calculated F from eq. (3), and given the value of f , we can determine t_1 graphically from fig. 14, which gives us the stopping distance.

It is also possible to express the straight line of fig. 14 by the equation:

$$F_p = 2.25(t - t_0)10^6, \dots \dots \dots (6)$$

and to calculate from this the relation between stop-

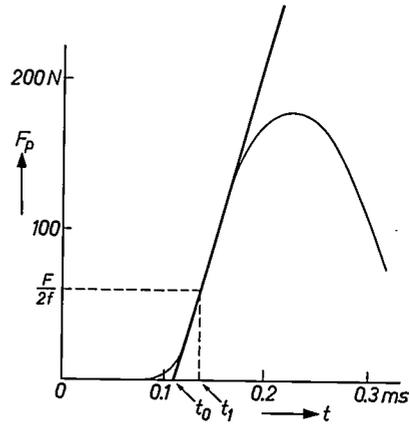


Fig. 14. Idealized curve of the braking system force as a function of time, at the beginning of the braking process. The braking magnet is energized at $t = 0$. t_0 is the delay of the braking magnet. The instant at which the tape is stopped (t_1) can be determined from the minimum force $F/2f$ which the braking system has to give in order to generate sufficient tensile force in the tape to stop the tape abruptly, as given by eq. (3).

ping distance and tape speed. From (3), (4) and (6) it follows that:

$$\frac{m_1 v_0 v_s}{2f} = 2.25(t_1 - t_0)10^6,$$

and therefore:

$$x = v_0 \left(\frac{m_1 v_0 v_s}{4.5f} 10^{-6} + t_0 \right) \dots \dots (7)$$

For the tape used by us $m_1 = 2.17 \times 10^{-3}$ kg/m and $v_s = 2.8 \times 10^3$ m/s (for unperforated tape). If we substitute this in (7), and take for f the measured value 0.12 and 119 μ s for t_0 , then we find the curve:

$$x = (0.11 v_0^2 + 1.19 v_0)10^{-4}, \dots \dots (8)$$

which corresponds very closely to the measured relationship between stopping distance and tape speed (fig. 15). When compared with the figure of 110 μ s mentioned above, the value of $t_0 = 119 \mu$ s is found to

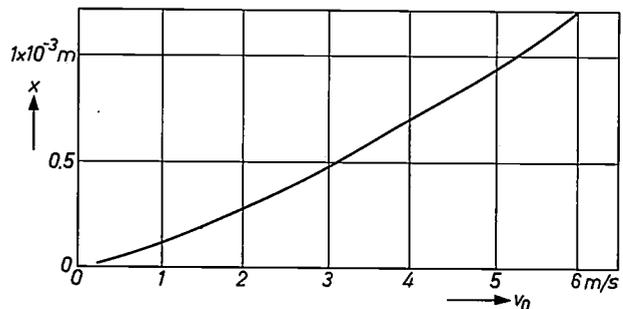


Fig. 15. The measured stopping distance x as a function of the tape speed v_0 . At $v_0 = 3$ m/s, corresponding to 1200 characters/s, the stopping distance is somewhat less than 0.5 mm, which is more than sufficient for stopping on a character. The calculated relation between stopping distance and tape speed (eq. 8) gives good agreement with this curve.

[1] See S. Timoshenko and D. H. Young, Vibration problems in engineering, Van Nostrand, Princeton N. J., 1955, 3rd edn., page 305.

be completely within the measuring accuracy, indicating considerable agreement between the measurements and the calculations.

At the required tape speed of 3 m/s, corresponding to a reading speed of 1200 characters/s, we find from fig. 15 a stopping distance slightly less than 0.5 mm. This meets the requirement (with a safety factor of 2) which we specified for the stopping distance in order to be able to stop on a character. The time t_1 in this case is about 0.15 ms.

Design of the braking magnet

The requirement to be met by the braking magnet is that the force F_p attracting the armature shall reach a specified high value in a very short time t_1 (about 100 N in 0.15 ms). From the following calculations it will appear that it is not possible to design an electromagnet whose time constant is small enough to enable the required conditions to be met when the magnet is energized. The time constant of a magnet capable of exerting a force of this magnitude is 10 ms at the very minimum. We shall show that it is nevertheless possible to achieve the short switching time required, by choosing an *extremely high* value for the ratio of the energizing voltage to the number of turns. This means that we have to energize the magnet with a much higher voltage than would be normal for such a magnet. A convenient method of calculating the magnet may be derived as follows.

The pull of an electromagnet whose armature is stationary is proportional to the square of the magnetic induction B and to the cross-sectional area A_m of the core. The maximum value of the induction B_1 is taken close to the saturation induction of the magnet material (the relation between B and H is still linear here); we therefore choose a material with a high saturation magnetization. The pull must now have reached the required value if the induction has reached the value B_1 . From this we find the required minimum value of A_m .

From the values B_1 and A_m we find the flux which must be present at the time t_1 :

$$\Phi_1 = B_1 A_m. \quad \dots \dots \dots (9)$$

For the magnetic circuit formed by the core, the armature and the air gap (see fig. 16a):

$$\Phi = \frac{ni}{W}, \quad \dots \dots \dots (10)$$

where i is the current through the coil, n the number of turns and W the reluctance of the circuit. Until saturation has been reached, W depends entirely on the dimensions A_m and l_a of the air gap, i.e. in this case on A_m and the thickness of the tape. This then determines the

value of W , and we can thus calculate from (10) the number of ampere-turns $(ni)_1$ that must be present at the time t_1 .

With the very rapid increase in the number of ampere-turns, quite large eddy currents can occur, causing a less rapid increase in the induction, so that the switching time becomes longer. To avoid this, considerable attention must be given to lamination of the magnet material: laminations only 0.1 mm thick are used, and they are insulated from each other by glass granules with a maximum diameter of 10 μ m. In fact, designing a magnet for such short switching times is similar to designing a transformer for frequencies in the region of 20 kHz.

If the maximum number of ampere-turns is known, we can calculate the second quantity on which the magnet core depends, i.e. the value of A_c , which is the area of the aperture between the two arms of the core, through which the turns have to pass (see fig. 16).

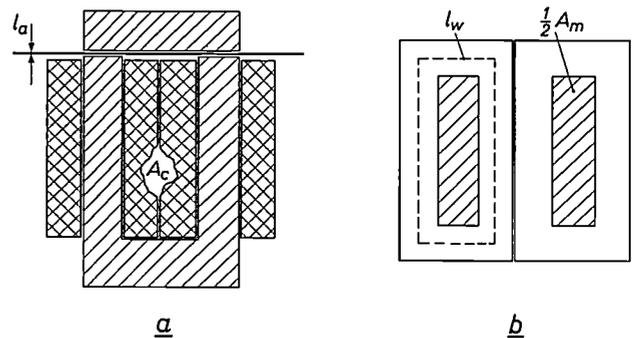


Fig. 16. The braking magnet. a) Cross-section through the yoke. l_a width of the air-gap. The turns of the winding are distributed over the two arms of the yoke. All windings pass through the central opening of section A_c . b) Perpendicular cross-section through the yoke. The average length per winding, l_w , is shown dotted. The cross-sectional area A_m of the core is the sum of the two hatched cross-sections.

From the permissible temperature increase of the magnet and the maximum switching frequency we find a maximum permissible value for the current density in the magnet winding. This current density can be expressed in terms of A_c and the number of ampere-turns:

$$j = \frac{ni}{\beta A_c},$$

where β is the space factor of the winding. Assuming that the number of ampere-turns does not exceed $(ni)_1$, we then find from the maximum permissible value of j the minimum value required for A_c .

Having thus determined the dimensions of the core, we must now design a winding which, within the specified time, can carry a current i_1 such that the number

of ampere-turns is $(ni)_1$. If the magnet is energized in the usual way by applying a voltage E at the time $t = 0$ (see fig. 17) the current through the coil increases exponentially:

$$i = \frac{E}{R} \left(1 - e^{-\frac{R}{L}t} \right), \dots \dots (11)$$

where L/R is the time constant τ of the magnet. The values of L and R are not yet known, but the ratio L/R can be calculated and is found to be proportional

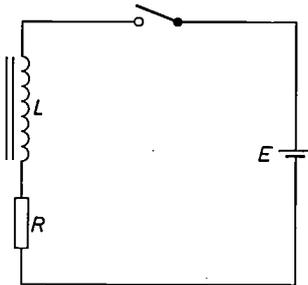


Fig. 17. The braking magnet circuit (schematic). L is the inductance and R the resistance of the magnet winding.

to $A_m A_c / l_a l_w$, where l_w is the average length of one turn (see fig. 16b). All these quantities are known, and therefore by fixing the dimensions of the core we have determined the time constant of the magnet. Calculating τ in this way for the braking magnet we find it to be much greater than t_1 . The situation then (see fig. 18) is that a current i_1 — corresponding to a number of ampere-turns $(ni)_1$ — must be reached in a time t_1 which is much shorter than τ . It is clear from fig. 18 that this can be achieved by choosing E such that the final value $i = E/R$ of the current is much greater than the required current i_1 . This final value may not be reached, however, since in determining A_c we have already assumed the maximum current. We must therefore limit the current to i_1 , otherwise the winding will burn out. The values still unknown can be calculated as follows.

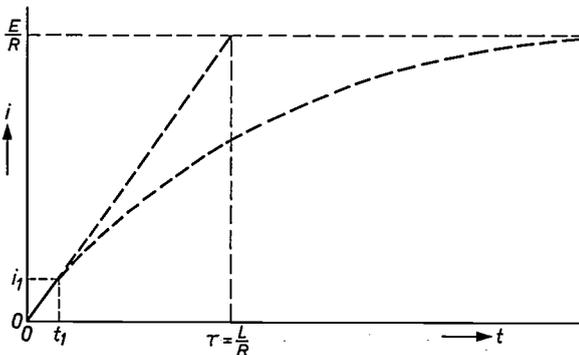


Fig. 18. Exponential curve of the current through the magnet coil. The time t_1 in which the magnet has to be energized to the required level is much shorter than the time constant τ .

In the time interval $t = 0$ to $t = t_1$ the current increase is still virtually linear. We can then write (11) as:

$$i = \frac{E}{L} t. \dots \dots (12)$$

As L is proportional to n^2 , then with the aid of (12):

$$ni = \frac{E}{cn} t,$$

where c is a known proportionality constant. Since ni at the time $t = t_1$ must be equal to or greater than the value $(ni)_1$ calculated above, it follows that:

$$\frac{E}{n} \geq \frac{c(ni)_1}{t_1}.$$

The quantities on the right-hand side are all known; if we choose a specific value for E we have then fixed the number of turns. Since these turns have to fill the area A_c , we can now calculate from n and A_c the required wire thickness. For the braking magnet described in this article, we find that $n = 77$ with a wire thickness of 0.9 mm, and $E = 50$ V and $i_1 = 3.5$ A. The value of E/n for this magnet is much greater than is usual for magnets of this type.

To limit the current through the coil to the value i_1 , a power transistor is included in the energizing circuit of the coil (fig. 19). If no base current flows through the transistor, the collector current is also zero and the

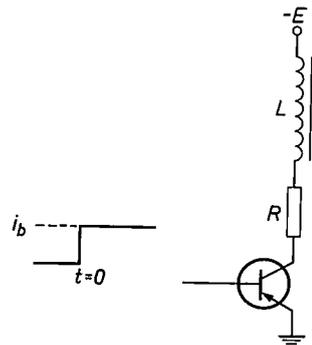


Fig. 19. Use of a transistor as switching device in the braking magnet circuit to prevent burn-out of the magnet winding. The curve shows the step-function shape of the base current i_b upon braking.

magnet is not energized. The magnet is switched on by connecting to the base a constant current source giving a current i_b whose level is such that i_c just reaches the value i_1 . The flat $i_c - V_c$ characteristic of the transistor ensures that this value is not exceeded. As the delay in a modern power transistor is negligible with respect to the switching time t_1 , interposing the transistor has no effect on the speed at which the value i_1 is reached. The

use of the transistor does set a limit to the magnitude of E , and this provides us with the last unknown factor in the design of the magnet.

A disadvantage of the circuit in fig. 19 is that during the quiescent state of the tape reader — when the braking magnet is energized — almost the entire voltage appears across the transistor which therefore has to dissipate power continuously. To overcome this disadvantage the punched-tape reader has a special energizing circuit, in which a lower voltage is applied to the transistor during the quiescent state. This circuit will not be discussed further here.

Results achieved

The punched-tape reader described above, originally designed for reading 1200 characters per second, was later made into a prototype and now proves capable of attaining a speed of 2500 characters per second. The stopping distance at this speed is 1.3 mm and stopping on a character is still just possible; this is therefore the limit of reliability. Other manufacturers have also made further progress. Readers are now on the market that can read 1800 characters per second. Since, however, these can only stop at the *next following* character an electronic buffer store is required for one character. A reader with a speed of 2000 characters per second has also appeared on the market but this requires an even more elaborate electronic buffer store.

Our prototype reader referred to has been in use since August 1964 at a maximum speed of 2000 characters per second (see title photo). A second tape reader of the same type has been in use since October 1965. The ability to stop on a character is very useful here, since the problems handled can require widely different processing times for the various characters.

Summary. The punched-tape reader described here, two models of which have been in use for a considerable time, is capable of reading-speeds up to 2500 characters per second without requiring a buffer store. This speed is possible because the tape can be stopped extremely quickly, "on" a character: if the "stop" signal is given as soon as a character appears, the same character is still visible when the tape has stopped. At a tape speed of 1200 characters/s the stopping distance is no greater than 0.5 mm. Use is made of double optical character read-out. The punched tape is transported by pressing it against a permanently rotating driving roller by means of a pressure roller controlled by an electromagnet. Braking is effected by energizing a very fast electromagnet whose armature is in permanent sliding contact with the tape, so that the distance to be travelled by the armature is extremely small. Upon braking, a shock wave is generated in the tape, and this travels along the moving tape at the speed of sound. The stopping distance calculated on this basis is found to be in good agreement with the measurements. A description is given of the optimum design of the braking magnet for rapid braking. This braking magnet is energized by a special transistor circuit which applies a very high voltage to the magnet for a very short time, enabling the desired current to be reached in a much shorter time than might be expected from the time constant of the magnet.

The skin effect

H. B. G. Casimir and J. Ubbink

- I. Introduction; the current distribution for various configurations
- II. The skin effect at high frequencies
- III. The skin effect in superconductors

"It was discovered by mathematical reasoning that when an electric current is started in a wire, it begins entirely upon its skin, in fact upon the outside of its skin; and that, in consequence, sufficiently rapidly impressed fluctuations of the current keep to the skin of the wire, and do not sensibly penetrate its interior.

Now very few (if any) unmathematical electricians can understand this fact; many of them neither understand it nor believe it. Even many who do believe it do so, I believe, simply because they are told so, and not because they can in the least feel positive about its truth of their own knowledge. As an eminent practitioner remarked, after prolonged scepticism, 'When Sir W. Thompson says so, who can doubt it?'"

These were the words of Heaviside in a plea for the use of mathematical methods in 1891. Now, seventy-five years later, the skin effect is such common knowledge that one sometimes thinks one understands it even without "mathematical reasoning".

The expression for this effect put forward in 1886 by Rayleigh and now often used as a matter of course, does however have its limitations. For example, its application to pure metals at very high frequencies leads to incorrect results, as noted by H. London in 1940. Superconductors are another special case, where the formula predicts an infinitely thin skin layer. These are a few of the problems which will be dealt with in this article, showing as far as possible their inter-relationship. The article is divided into three parts, the first of which follows here.

I. Introduction; the current distribution for various configurations

If a direct current flows in a conducting wire, it will be distributed uniformly over the cross-section. With alternating current, however, the current distribution is not homogeneous and, if the frequency, conductivity and dimensions of the conductor satisfy certain conditions, to be dealt with later, the current flows mainly in a thin layer at the surface of the conductor. This phenomenon is called the skin effect. It is an electrodynamic effect, that is to say, it is a result of the way in which time-varying electric and magnetic fields and electric currents are interrelated. The skin effect phenomenon is quite different from the action of a Faraday cage, for instance, which acts as a barrier to a static electric field purely and simply as a result of the fact

that the charges in the wall of the cage are mobile.

The tendency of the current to flow at the surface is closely connected with the stable character of the electromagnetic phenomena. We see this in the following way. Let us assume that inside the metal there is a filamentary current I which is increasing in strength (*fig. 1*). This current is associated with a rotational magnetic field H around it which is also increasing. A changing magnetic field induces a rotational electric field E which, in turn, induces a current in the metal. According to Lenz's law, the direction of E is such that it opposes the increase of I , thus keeping the situation stable. The figure shows that simultaneously at some distance from I a current is generated parallel to I . The net result therefore is that the current is forced outwards. Furthermore, we see that this effect increases with frequency (E is larger the more rapid the change

Prof. Dr. H. B. G. Casimir is a member of the Board of Management of N.V. Philips' Gloeilampenfabrieken; Dr. J. Ubbink is with Philips Research Laboratories, Eindhoven.

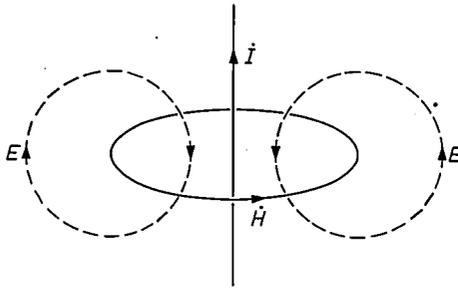


Fig. 1. A filamentary current I is accompanied by a magnetic field H . As I and H increase, an electric field E is produced whose direction near I is such as to oppose the increase in I , whereas further away from I the induced field E produces a current in the conducting medium parallel to I .

of H) and with conductivity (the larger the conductivity, the larger the current caused by E).

In this first article, we shall consider in some detail the configuration of the current for one or two special situations.

The skin effect can be a disadvantage in transporting alternating current energy along a wire or cable. To keep the resistance low, the cross-section of the conductor should be as large as is practicable but the effect of increasing the diameter is far less than with direct current. For alternating currents it is advantageous to use hollow cables (in power engineering), or braided cable (in radio engineering). At microwave frequencies the skin effect can be put to good use: the effect makes it possible to transport and store electromagnetic energy without radiation losses by using closed waveguides and resonant cavities.

At high frequencies the skin layer may be regarded as a layer screening electromagnetic radiation incident upon the metal: as a result of the conducting properties of the metal the radiation penetrates into the metal no further than the depth of the skin layer. Even a single electron, whether bound or free, possesses screening properties to a certain extent: incident radiation is scattered by the electron so that the power travelling straight on is less than the incident power.

As ω (the angular frequency) or σ (the conductivity) increases, the penetration depth δ decreases. In simple electron theory, σ is proportional to the mean free path l of the conduction electrons. As σ or ω increases, there will therefore be an instant at which δ becomes smaller than l . The current density at a given point will then no longer be determined simply by the local field intensity and the static conductivity, and the simple theory of the skin effect will no longer apply. This situation is referred to as the "anomalous skin effect".

As the frequency increases, other effects may become significant, namely, relaxation effects: the electron is

subject to many cycles of the alternating field between two collisions and within the mean time that it spends in the skin layer. Broadly speaking, the field then "sees" in effect a layer of free electrons.

Finally, at still higher frequencies, the "plasma frequency" of the metal will be reached above which the metal becomes transparent to the radiation. These skin effect complications at high frequencies form the subject of part II.

Metals in the superconducting state form a special class of conductors. These will be discussed in part III. We consider the two-fluid model, in which the electrons are divided into two types, normal and superconducting. The superconducting electrons, although they dissipate no energy, do have screening properties (even at $\omega = 0$). They therefore cause a skin effect: fields can penetrate only to the London penetration depth, which is independent of the frequency, and the "super-current" in this layer does not cause energy losses. The normal electrons within this layer do, however, absorb electromagnetic energy for $\omega \neq 0$, giving some (small) high frequency losses. There is, however, a frequency limit above which the superconducting electrons also absorb energy. This absorption is due to the transfer of electrons from the superconducting to the normal state by the radiation, via a quantum process. A superconductor differs very little from an ordinary metal at frequencies above this frequency limit (which lies in the microwave range).

Since the skin effect is based entirely upon the dynamic properties of electromagnetic fields and currents as given by Maxwell's equations, it will be useful to set down the four equations here:

$$\text{curl } \mathbf{H} = \partial \mathbf{D} / \partial t + \mathbf{J}, \quad \dots \dots \dots (1)$$

$$\text{curl } \mathbf{E} = -\partial \mathbf{B} / \partial t, \quad \dots \dots \dots (2)$$

$$\text{div } \mathbf{B} = 0, \quad \dots \dots \dots (3)$$

$$\text{div } \mathbf{D} = \rho. \quad \dots \dots \dots (4)$$

The following points should be noted:

- a) In what follows we shall in general regard the material as a medium with a given relative dielectric constant and permeability ϵ_r and μ_r (so that in the material $\mathbf{D} = \epsilon \mathbf{E}$, $\mathbf{B} = \mu \mathbf{H}$, with $\epsilon = \epsilon_r \epsilon_0$, $\mu = \mu_r \mu_0$), in which the free electrons carry the current. ϵ_r and μ_r are generally of the order of unity for non-ferromagnetic materials. Interesting complications which may arise when μ_r becomes much greater than unity (ferromagnetism) will be discussed in the last section of part I.
- b) Over a very wide frequency range, the term $\partial \mathbf{D} / \partial t$ (the "displacement current") in the metal is negligible with respect to the current density \mathbf{J} and may therefore be ignored. When \mathbf{J} can be represented simply as $\sigma \mathbf{E}$,

this amounts to taking ω as negligible with respect to σ/ϵ . For copper at room temperature for instance, $\sigma \approx 10^8 (\Omega\text{m})^{-1}$ and $\sigma/\epsilon \approx 10^{19} \text{ s}^{-1}$, which is very much higher than the frequencies with which we shall be dealing in this article. $\partial\mathbf{D}/\partial t$ can become comparable to \mathbf{J} only in the relaxation range, where \mathbf{J} becomes smaller than $\sigma\mathbf{E}$; even then $\partial\mathbf{D}/\partial t$ begins to become really significant only at frequencies near the plasma frequency.

Parallel wires

The case of a number of parallel wires lying in a plane and connected in parallel may be used as a simple illustration of the essential features of the skin effect. The currents in the wires affect one another by induction so that the current in the innermost wires is less than in the outer wires.

Let us consider three equidistant wires, 1, 2 and 3, connected together at their ends and connected to an a.c. source of angular frequency ω (fig. 2).

If the frequency is so low that induction effects can be ignored, the currents I_1, I_2 and I_3 through 1, 2 and 3 are all equal in phase and amplitude. This is no longer the case at higher frequencies. Consider the circuit (1, 2) formed by wires 1 and 2. There is a flux Φ_3 , produced by I_3 , through this circuit. An increase in I_3 induces an e.m.f. in (1, 2) opposing I_2 just as it does I_3 , since I_2 and I_3 are on the same side of (1, 2). An increase in I_3 therefore tends to oppose I_2 and to reinforce I_1 ; a decrease in I_3 tends to reinforce I_2 ; this has the result that I_2 lags slightly in phase behind I_1 and I_3 . Once there is a difference between I_1 and I_2 , $I_1 - I_2$ represents a circulating current in (1, 2) producing a flux which, in turn, largely determines the difference between I_1 and I_2 . The net result is that I_2 lags in phase behind I_1 and I_3 but has virtually the same amplitude.

The situation is easier to grasp at very high frequencies because of the consideration that the net flux through a circuit must be almost zero, because any net flux would produce almost infinitely high e.m.f.'s. Let us assume that the fluxes through (1, 2) as a result of I_1, I_2 and I_3 are Φ_1, Φ_2 and Φ_3 respectively (fig. 2b). Owing to the geometry and as I_1 and I_3 are equal in phase (and also in magnitude), due to symmetry, Φ_1 and Φ_3 are in opposite phase. Because the net flux is zero, Φ_2 must also be in phase or in antiphase: because 1 and 2 are symmetrical with respect to the zone (1, 2) and 3 is further away, $\Phi_1 = AI_1, \Phi_2 = AI_2$ and $\Phi_3 = aI_3$, with $a < A$. It then follows from $\Phi_1 = \Phi_2 + \Phi_3$ and $I_3 = I_1$, that $AI_1 = AI_2 + aI_1$, hence $I_2 = I_1(A - a)/A$. I_2 therefore has a smaller amplitude than I_1 , and there are no phase differences.

For a quantitative calculation let us assume that

the wires are thin and long compared with their separation: $r_w \ll a \ll L$ (r_w is the radius of the wires, L their length and a is the spacing). It follows from (1), using Stokes's theorem, that if H is the field at a distance x from a wire carrying a current I :

$$2\pi xH = I, \quad \text{hence} \quad H = I/2\pi x,$$

and the total flux through a surface bounded by two values of $x, x = p$ and $x = q$, is:

$$\int_p^q B dS = \int_p^q \mu_0 H dS = \int_p^q \frac{\mu_0 I L}{2\pi x} dx = \frac{\mu_0 I L}{2\pi} \left[\ln x \right]_p^q.$$

The total flux between 1 and 2 is therefore:

$$\begin{aligned} \frac{\mu_0 L}{2\pi} \left(I_1 \ln \frac{a}{r_w} - I_2 \ln \frac{a}{r_w} - I_3 \ln \frac{2a}{a} \right) &= \\ &= \frac{\mu_0 L}{2\pi} \left[(I_1 - I_2) \ln \frac{a}{r_w} - I_3 \ln 2 \right]. \end{aligned}$$

Let us assume that the fields and currents vary with time as $\exp(j\omega t)$. It then follows from (2) and Stokes's

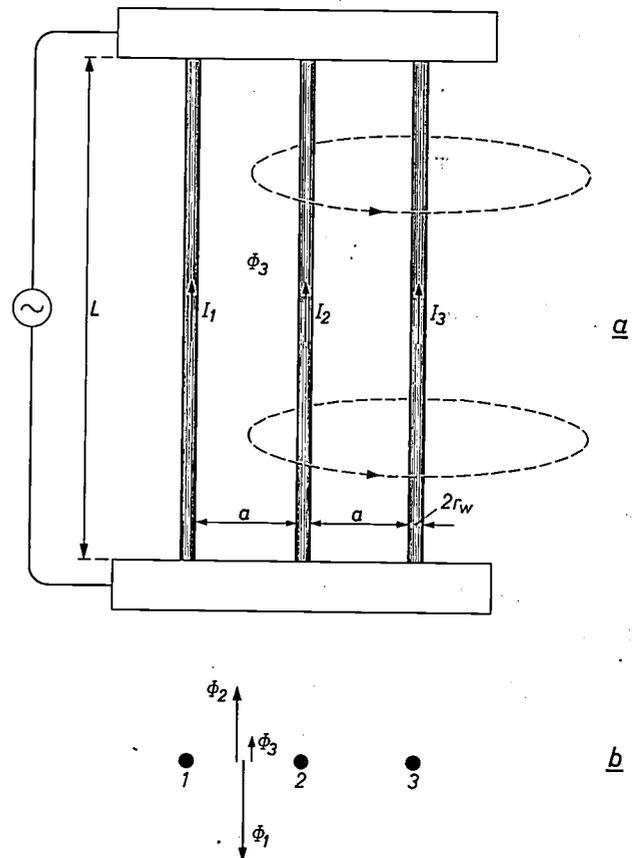


Fig. 2. a) Three parallel wires, connected electrically in parallel, of length L , diameter $2r_w$ and spacing a . Φ_3 is the flux between the first two wires caused by the current in the third. b) Qualitative indication of the fluxes through the circuit (1, 2) at very high frequencies.

theorem, with $B = \mu_0 H$, $E_1 = R_l I_1$, $E_2 = R_l I_2$ ($R_l =$ resistance per unit length of the wires) that:

$$R_l(I_1 - I_2)L = -\frac{j\omega\mu_0 L}{2\pi} \left[(I_1 - I_2) \ln \frac{a}{r_w} - I_3 \ln 2 \right],$$

or, substituting $I_3 = I_1$:

$$\frac{I_2}{I_1} = 1 - \frac{\ln 2}{\alpha}, \dots \dots \dots (5)$$

with

$$\alpha = \ln \frac{a}{r_w} - j \frac{2\pi R_l}{\mu_0 \omega} \dots \dots \dots (6)$$

The combined effect of the geometrical and electrical factors is contained in the parameter α .

As the frequency varies from 0 to infinity, the locus of α is parallel to the imaginary axis and I_2/I_1 describes a semicircle in the complex plane as sketched in fig. 3. Although, under the stated condition $a \gg r_w$, the effect can never become particularly large, it still shows a few essential traits of the skin effect, as may be seen from (5) and (6) and fig. 3; the "inner current" is smaller than the "outer current" and lags behind it; at low frequencies there are phase differences only, which increase as the frequency increases and the resistance decreases. At very high frequencies the distribution of current between the wires, and hence the field distribution outside the wires, is independent of

the frequency. On going to higher frequencies we must bear in mind that R_l is the *effective* resistance per unit length, and that at higher frequencies it no longer corresponds to the resistance at zero frequency — the cause of this being, of course, the skin effect in each wire.

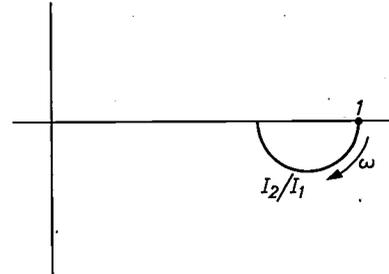


Fig. 3. The current ratio I_2/I_1 describes a semicircle in the complex plane as the frequency passes from zero to high values.

It is still practicable to calculate the current distribution under the set conditions for four wires. The result becomes more and more complex as the number of wires increases: if there are n wires, $n - 1$ equations with $n - 1$ unknown quantities have to be solved.

The current in the wires for three, four and five wires, with $\alpha = 3$, is shown in fig. 4. This real value of α will

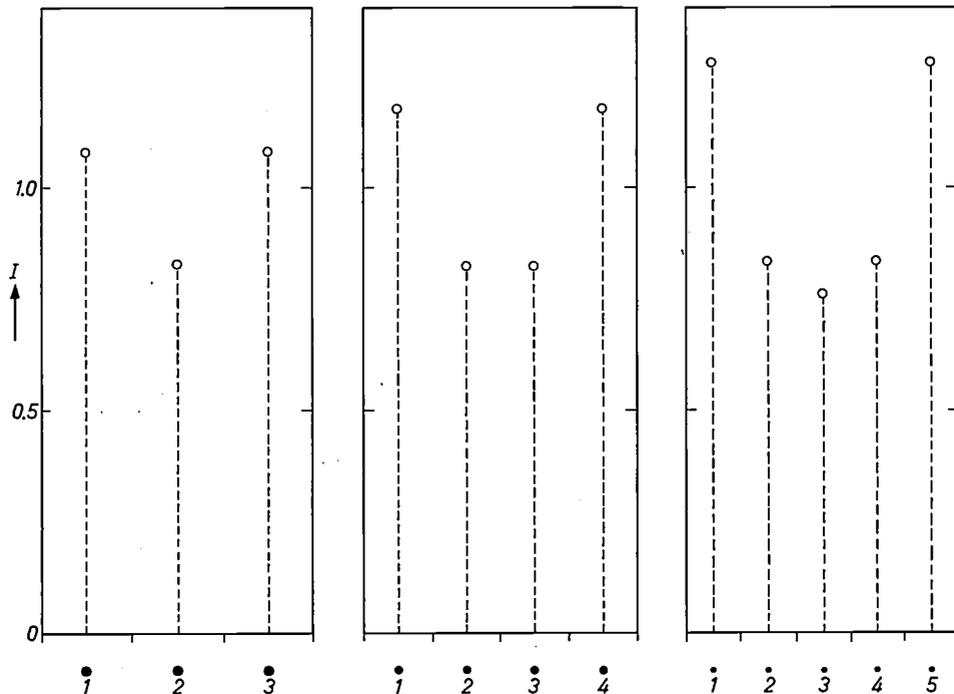


Fig. 4. The distribution of an alternating current over three, four or five wires, with $\alpha = 3$. The wires are shown below in cross-section, distributed over the same width in the three cases. The current in each wire is shown vertically. The current is normalized to give an average "current density" of 1 (the average current per wire). The fact that α is real implies high frequencies and the value 3 corresponds to a wire spacing/wire diameter ratio of about 10.

be realized at very high frequencies where the imaginary part becomes insignificant (there are therefore no phase differences between the currents). The value 3 for the real term corresponds to a value of about 20 for a/r_w . In each case the wires are distributed over the same width and the total current is chosen such that the average current per wire (a kind of current density) is unity.

Although, as we have seen, it is possible to calculate the "skin effect" for a small number of parallel wires, there is little sense in performing this type of calculation for an increasing number of wires in order to arrive, in the limit, at the effect in a solid conductor. The analysis of this case is much better dealt with by considering the conductor from the start as a continuum with continuously distributed fields and currents, as below.

Wire of circular cross-section

The distribution of an alternating current over the cross-section of a cylindrical wire is the standard example in discussing the skin effect. As the problem has circular symmetry, we look for a circularly-symmetric solution in which the electric field E and the current density J are parallel to the axis of the wire and the magnetic field H is perpendicular to a plane passing through the axis of the wire (fig. 5). E , H and J are simply functions of the distance r from the axis of the cylinder, with time dependence $\exp(j\omega t)$. Applying Stokes's theorem to an area with a contour a of radius r , it follows from (1) that for H in the wire:

$$2\pi r H = \int_0^r 2\pi r' J dr' .$$

Differentiating with respect to r :

$$H + r \frac{dH}{dr} = Jr . \quad \dots \quad (7)$$

If we now apply equation (2) and Stokes's theorem to a surface with a contour b , we find:

$$E_1 - E_2 = -\frac{dE}{dr} dr = -j\omega\mu H dr .$$

With $J = \sigma E$, it follows that:

$$H = -j \frac{1}{\omega\mu\sigma} \frac{dJ}{dr} . \quad \dots \quad (8)$$

Differentiating (8) and substituting for H and dH/dr in (7) gives a differential equation for J :

$$\frac{d^2 J}{dr^2} + \frac{1}{r} \frac{dJ}{dr} - \frac{2j}{\delta_k^2} J = 0 , \quad \dots \quad (9)$$

where
$$\delta_k^2 = \frac{2}{\omega\mu\sigma} ; \quad \dots \quad (10)$$

δ_k is a quantity with the dimension of a length, and is called the (classical) "skin depth". Equation (9) cannot be solved by elementary means. However, if a dimensionless complex variable

$$x = (1 - j) \frac{r}{\delta_k} \quad \dots \quad (11)$$

is introduced the equation reduces to:

$$x^2 \frac{d^2 J}{dx^2} + x \frac{dJ}{dx} + x^2 J = 0 , \quad \dots \quad (12)$$

a Bessel differential equation of order zero. The only solution which remains finite at $x = 0$ is the Bessel function of the first kind. The absolute value and argu-

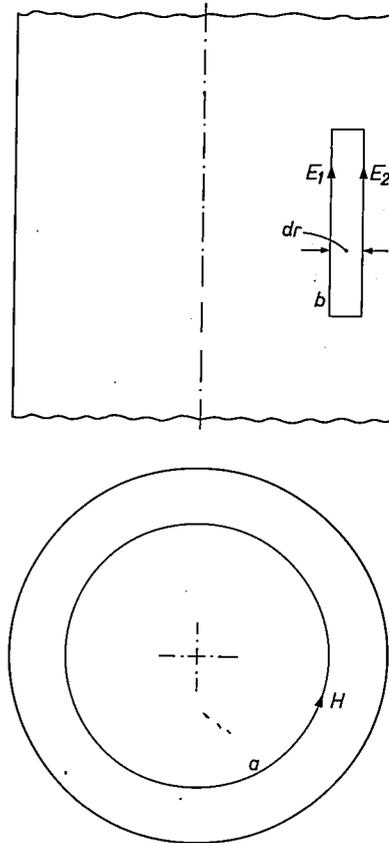


Fig. 5. Diagram relating to the calculation of the current distribution in a cylindrical wire. Above: longitudinal cross-section. Below: transverse cross-section.

ment of this Bessel function are tabulated, e.g. in the Jahnke-Emde tables, for a variable of the complex form (11). The amplitude of J (the modulus of the Bessel function) is shown in fig. 6 as a function of r/δ_k for $0 < r/\delta_k < 10$; the amplitude multiplied by the cosine of the phase (the argument of the Bessel function) is shown by the dashed curves. This last curve gives an idea of the distribution of the current at a given time. It can be seen from the figure that no

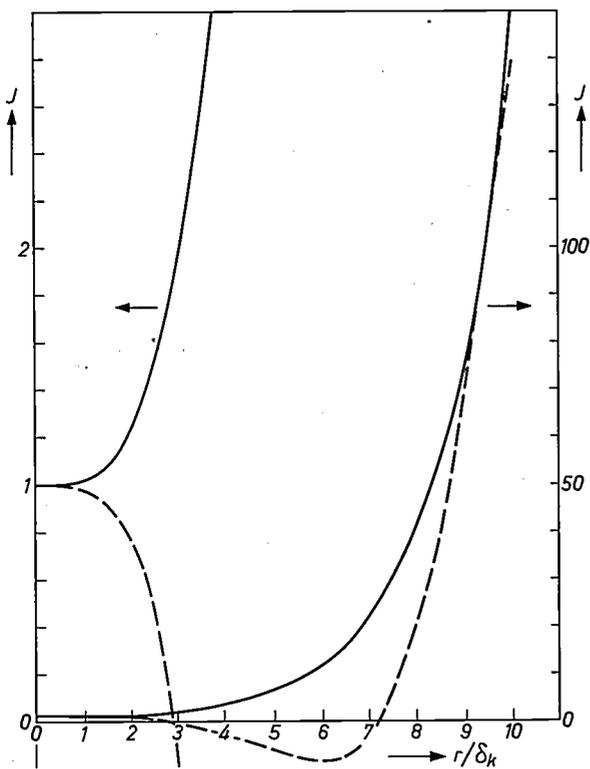


Fig. 6. The current distribution in a cylindrical wire. The amplitude of the current density is plotted vertically and r/δ_k horizontally. The dashed line shows the current distribution at an arbitrary instant. Both curves are also given for small r/δ_k with the vertical scale enlarged 50 times.

significant amplitude or phase variations occur while $r/\delta_k < 1$, or, in other words, the skin effect is not encountered in wires thinner than the skin depth.

For $r \gg \delta_k$ we obtain a relatively large increase in the current density. If we try $J = \exp(Cx)$ as a solution of (12), we see that the relative value of the second term decreases as x increases. We can therefore consider the extreme case in which the second term is neglected and in this case the simple exponential function is, in fact, a solution. The current is now substantially at the surface of the wire, i.e. $R \gg \delta_k$, where R is the radius of the wire. It is therefore convenient to introduce the variable $z = R - r$. Returning to (9) and omitting the second term, we find immediately as a solution:

$$J = J_0 \exp [-(1 + j)z/\delta_k], \dots (13)$$

J_0 being the current density at the surface. Bearing in mind that this still has to be multiplied by $\exp(j\omega t)$, it will be seen that the variation of the current density is given by an attenuated wave travelling inwards. The attenuation is very large: at the first point where the phase is opposite to that at the surface, $z/\delta_k = \pi$, the amplitude is $\exp(-\pi) = 0.05$ times that at the surface.

The following rules of thumb (which can easily be

checked with (10)) may be used to give an idea of the skin depth at various frequencies. For copper at room temperature ($\sigma = 0.6 \times 10^8 \Omega^{-1}\text{m}^{-1}$, $\mu_r = 1$, so that $\mu = \mu_0 = 1.26 \times 10^{-6} \text{ H/m}$) the penetration depth is a) $\delta_k = 1 \text{ cm}$ (or, more accurately, 0.9 cm) at 50 Hz; b) $\delta_k = 1 \text{ micron}$ at a wavelength of 7 cm (microwave region).

The skin depth in the microwave region is thus very small. Experimental verification of equation (10) in the microwave region therefore requires very careful attention to the surface quality. Equation (10) can, for example, be tested by measuring the Q (quality factor) of a resonant cavity. $1/Q$ is a direct measure of the power dissipation in the wall and the dissipation is directly related to the skin depth. The measured Q of a resonant cavity is generally considerably lower than that predicted by (10), partly because the grooves produced during machining are deeper than the skin depth, so that the surface becomes effectively much larger. Little improvement is achieved by polishing such a surface. It is however possible to approach the theoretical value of Q very closely by taking the greatest possible care in machining the surface. Gevers [1] obtained 98% of the theoretical Q for a resonant cavity machined with a feed much smaller than the radius of curvature of the point of the tool: the tool used had a radius of about 100 μm , and the feed was about 1 μm .

Conductor of arbitrary shape

If we consider a conductor of any shape in which an alternating current is flowing, then in order to calculate the distribution of the current in the conductor, we have to find solutions of Maxwell's equations in the conductor and in the space around it that are compatible at the bounding surface.

Once more we shall restrict ourselves to a single frequency low enough to permit the displacement current to be neglected and again assume that the currents and fields have the time dependence $\exp(j\omega t)$. If we now take the curl of (1) and substitute $\mathbf{J} = 0$ outside the conductor and $\mathbf{J} = \sigma\mathbf{E}$ and $\text{curl } \mathbf{E} = -j\omega\mu\mathbf{H}$ (from (2)) inside it, we find (since $\text{curl curl} = \text{grad div} - \Delta$ and $\text{div } \mathbf{H} = 0$):

$$\text{outside the conductor: } \Delta\mathbf{H} = 0, \dots (14)$$

$$\text{inside the conductor: } \Delta\mathbf{H} = j\omega\mu\sigma\mathbf{H}. \dots (15)$$

It is not practicable to search for solutions to these equations that are compatible at an arbitrary boundary surface. The case of the cylindrical wire was so simple because the solutions for the regions inside and outside

[1] M. Gevers, Measuring the dielectric constant and loss angle of solids at 3000 Mc/s, Philips tech. Rev. 13, 61-70, 1951.

the wire can in fact be obtained independently of each other, provided that we restrict ourselves to solutions with circular symmetry. The matching of the two solutions is no problem, for the phase and amplitude for both solutions are constants at the boundary plane and only these constants have to be matched. The external solution is not affected by the radial current distribution in the wire.

If we now consider the limiting case in which the skin layer is thin compared to the dimensions of the conductor and the radii of curvature of the surface, we can simplify the problem even if the conductor does not have circular symmetry.

The problem then breaks down into two parts:

- a) How do field and current vary in a direction perpendicular to the surface?
- b) How do field and current vary along the surface?

Let L be a length much greater than the skin depth, but considerably smaller than the smallest radius of curvature of the surface. L is then much smaller than the conductor itself or the apparatus exciting the magnetic field, such as a coil. The variation in the magnetic field outside the conductor over a distance L , particularly the component perpendicular to the surface, is then very slight. Now consider a flat "box" at the surface (length and width $\approx L$, thickness $\approx \delta_k$) completely containing the skin layer (see fig. 7). The component perpendicular to the surface, H_n , varies very little over the upper wall of the box. From (3) and Gauss's theorem, the inward flux passing through

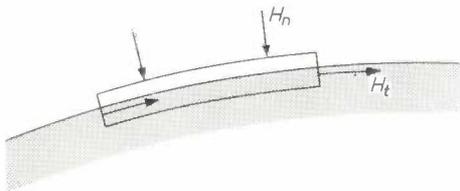


Fig. 7. The magnetic fields at the surface of a flat volume element enclosing the skin layer.

the walls of the box is the same as the outward flux, and, since no flux passes through the lower wall, $H_n L^2$ is, at most, of the order of $H_t \delta_k L$ (or even less, because the fluxes through the side walls virtually compensate one another). H_n is thus at most about $(\delta_k/L)H_t$. In the limiting case where $\delta_k \ll L$, we can therefore disregard the component perpendicular to the surface with respect to the tangential components: the magnetic field is tangential at the surface.

Part (a) of the problem has now become very simple. We choose a local co-ordinate system with origin in the surface and z -axis perpendicular to the surface. Be-

cause the fields vary very little along the surface, $\Delta \mathbf{H} \approx \partial^2 \mathbf{H} / \partial z^2$, so that in the metal:

$$\frac{\partial^2 \mathbf{H}}{\partial z^2} = j\omega\mu\sigma \mathbf{H},$$

from which it follows, with the boundary condition that the fields are zero in the interior, that

$$H = H_0 \exp [-(1 + j)z/\delta_k],$$

in conformity with the current distribution already found in (13).

Part (b) of the problem can be formulated as a potential problem with a simple boundary condition, since $\text{curl } \mathbf{H} = 0$ in the region outside the conductor. It follows from this that \mathbf{H} can be derived from a magnetic potential φ , i.e. $\mathbf{H} = \text{grad } \varphi$. Since $\text{div } \mathbf{H} = 0$ we have at once $\Delta \varphi = \text{div grad } \varphi = 0$. The boundary condition for this potential is $H_n = \partial \varphi / \partial z = 0$. In other words, the magnetic equipotential surfaces terminate at right angles to the surface of the conductor. It should be noted that we must be prepared to find a potential which is not a single-valued function of position. For every revolution around a current-carrying wire, there is a certain increase in the magnetic potential.

If we now restrict ourselves to a cylindrical, but not necessarily circularly-symmetric, conductor, question (b) can be reduced to a potential problem in another way. Under these conditions the current distribution along the surface is the same as the distribution of the charge over the surface of an insulated, charged conductor of the same shape. In other words, the boundary condition is that the potential is constant at the surface of the conductor.

That this is so may be seen from the following. A magnetic field may be described by a vector potential \mathbf{A} such that

$$\text{curl } \mathbf{A} = \mathbf{H}; \quad \dots \dots \dots (16)$$

for \mathbf{A} we choose the gauge:

$$\text{div } \mathbf{A} = 0. \quad \dots \dots \dots (17)$$

It follows at once from (1) (with $\partial \mathbf{D} / \partial t = 0, \mathbf{J} = 0$ and $\text{curl curl} = \text{grad div} - \Delta$) that

$$\Delta \mathbf{A} = 0.$$

The solution that interests us is the one in which the current is parallel to the longitudinal direction of the conductor (the x direction), the magnetic field is perpendicular to it and all the quantities are independent of x (the problem is evidently two-dimensional). To describe such situations, a vector potential \mathbf{A} in the longitudinal direction: $A_x = A, A_y = A_z = 0$, is a

convenient choice. Equation (16) is then equivalent to $H_y = \partial A / \partial z$, $H_z = -\partial A / \partial y$.

Let us again take at some point on the surface a local system of axes, with the z axis perpendicular to the surface. In our limiting case $\delta_k \ll L$ the magnetic field has, as we saw, no vertical component at the surface: $H_z = 0$, so that $\partial A / \partial y = 0$, which means that A is constant at the surface. The two-dimensional problem thus reduces to finding a scalar function $A(y, z)$ which satisfies $\Delta A = 0$ outside the conductor and is constant at the surface.

Flat strip

We shall use the above to calculate the current distribution over the width of a flat strip for the case of a very small skin depth. If we consider a flat strip (width B , thickness D) in which an alternating current flows in the longitudinal direction, we can distinguish between various frequency ranges. At very low frequencies (direct current) the current is uniformly distributed over the cross-section. At higher frequencies there is a range in which the current is still uniformly distributed over the thickness ($\delta_k \gg D$) but no longer over the width. The calculation of the distribution over the width in this situation has been carried out by Belevitch, Gueret and Liénard [2]. At very high frequencies the current flows in a skin layer that is thin compared to the strip thickness ($\delta_k \ll D$), and here, again the distribution is not uniform over the width. It is this distribution that we shall now calculate.

We take $\delta_k \ll D$, so that the above considerations apply, and $D \ll B$. With this latter condition we can idealize the strip as infinitely thin so that its cross section becomes a line and our boundary condition becomes: A is constant over the line. The two-dimensional potential problem can now be solved by conformal mapping. We take co-ordinates X, Y in the plane of our problem. Consider the complex variable $Z = X + jY$ and write down an analytical function w of Z :

$$w = u + jv = w(Z).$$

This relation maps the Z -plane on to the w -plane. The relationship $\Delta A = 0$ remains valid in the w -plane for it follows from the Cauchy relations for analytical functions,

$$\frac{\partial u}{\partial X} = \frac{\partial v}{\partial Y}, \quad \frac{\partial v}{\partial X} = -\frac{\partial u}{\partial Y},$$

that:

$$\Delta_z A = \left[\left(\frac{\partial u}{\partial X} \right)^2 + \left(\frac{\partial v}{\partial Y} \right)^2 \right] \Delta_w A,$$

in which

$$\Delta_z = \frac{\partial^2}{\partial X^2} + \frac{\partial^2}{\partial Y^2}, \quad \Delta_w = \frac{\partial^2}{\partial u^2} + \frac{\partial^2}{\partial v^2},$$

so that the requirement $\Delta_z A = 0$ becomes $\Delta_w A = 0$. Let us assume that the line on which A is constant is $Y = 0, -1 \leq X \leq +1$. Now, by choosing a relation between w and Z which maps this line on to a circle, we reduce our problem to a simple circular-symmetric problem. Such a relation is

$$Z = \frac{1}{2} (w + 1/w).$$

The circle $|w| = 1$ described by $w = \exp(j\vartheta)$ as ϑ varies is converted into $Z = \cos \vartheta$, i.e. $X = \cos \vartheta, Y = 0$, which describes our line. Taking $w = r \exp(j\vartheta)$, then for a circular-symmetrical A outside the circle, $\Delta A = 0$ is equivalent to:

$$\frac{\partial^2 A}{\partial r^2} + \frac{1}{r} \frac{\partial A}{\partial r} = 0.$$

The solution is $A = C \ln r$, in which C is an arbitrary constant. From (1) and Stokes's theorem the surface current density J_s , i.e. the current per unit width of surface $J_s = \int J dz$, is equal to the tangential magnetic field, and hence:

$$J_s = \lim_{r \rightarrow 0} \frac{\partial A}{\partial Y}.$$

For points close to the surface, $r \approx 1$: let us put $r = 1 + \epsilon$ ($\epsilon \ll 1$). Then:

$$Y = \text{Im } Z = \frac{1}{2} (r - 1/r) \sin \vartheta \approx \epsilon \sin \vartheta$$

and

$$\frac{\partial A}{\partial Y} = C \frac{\partial \ln r}{\partial Y} = \frac{C}{r} \frac{\partial r}{\partial Y} = \frac{C}{r} \frac{\partial \epsilon}{\partial Y} = \frac{C}{(1 + \epsilon) \sin \vartheta};$$

hence

$$J_s = \frac{C}{\sin \vartheta} = \frac{C}{\sqrt{1 - X^2}}.$$

The current along one face of the strip is:

$$I_0 = \int_{-1}^{+1} \frac{C dX}{\sqrt{1 - X^2}} = C\pi$$

(the total current is $2I_0$, since both faces carry current). To facilitate comparison with the parallel-wire problem we normalize so as to obtain an average surface current density of unity. We therefore put $C = 2/\pi$ (the strip $Y = 0, -1 \leq X \leq +1$, has a width of 2). J_s is plotted in fig. 8 for $C = 2/\pi$.

We note that, under the conditions $\delta_k \ll D \ll B$, the current distribution over the width is independent of the frequency and the conductivity. Further, the solution given here forms an asymptotic approximation in the range of (lower) frequencies where, in contrast with the above conditions, the skin depth is large in comparison with the thickness, but small in comparison with the root of the product of width and thickness [2].

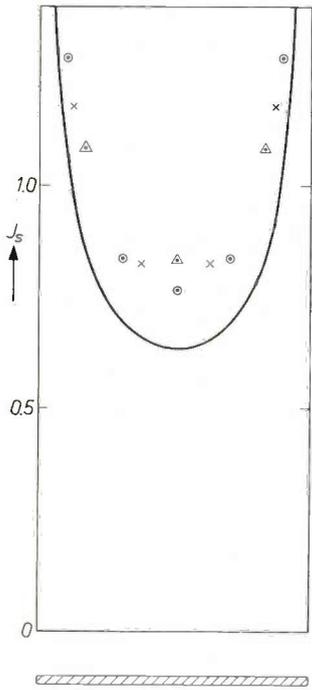


Fig. 8. The distribution of the surface current over the width of a flat strip (width B , thickness D) with $\delta_k \ll D \ll B$ (the solid curve). The curve is normalized to give a "surface current density" of unity. The points indicated by triangles, crosses and circles, taken from fig. 4, represent the current distribution over three, four or five wires respectively for $d = 3$ (cf. fig. 4).

H is constant. As the field must vanish for $z \rightarrow \pm \infty$, $H = 0$ everywhere outside the strips; we assume that $H = H_0$ between the strips. Inside each strip:

$$\frac{d^2 H}{dz^2} = j\omega\mu\sigma H.$$

With the boundary conditions $H = 0$ for $z = a + d$ and $H = H_0$ for $z = a$, we find for the field in the "upper" strip ($z > 0$):

$$H = H_0 \frac{\sinh a(a + d - z)}{\sinh ad}, \dots (18)$$

in which $a = (1 + j)/\delta_k$. The field in the lower strip (with $H = 0$ for $z = -a - d$, and $H = H_0$ for $z = -a$) follows from this by replacing z by $-z$. For the current density in the upper strip we find (with $J = dH/dz$):

$$J = -aH_0 \frac{\cosh a(a + d - z)}{\sinh ad}, \dots (19)$$

and that in the lower strip follows by replacing z by $-z$ and multiplying the entire expression by -1 . The current per unit width in the upper strip is:

$$\int_a^{a+d} J dz = \int_a^{a+d} \frac{dH}{dz} dz = [H]_a^{a+d} = -H_0. \dots (20)$$

No complications arise in taking the d.c. limit of (19), $\delta_k \rightarrow \infty$, i.e. $a \rightarrow 0$. Expanding (19) and neglecting all but first order terms in a , $J = -H_0/d$, as is consistent with (20). The system may be considered as a deformed single-turn coil.

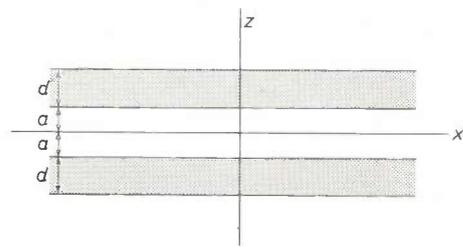


Fig. 9. Two strips of thickness d at a spacing $2a$ extending to infinity in length and width.

One might be inclined to think that the problem of the flat strip could be reduced to a one-dimensional problem by considering the limiting case of an infinitely wide strip (where one need consider only the coordinate perpendicular to the plane of the strip). This is however not possible because the contribution towards the magnetic field from distant current strips cannot be neglected. The field due to a current strip of width dX at a distance X (measured along the width) is proportional to dX/X , and the integral of such a term diverges as the boundary is extended to infinity. Thus the width cannot be made infinitely large, and the co-ordinate along the width cannot be eliminated by such means.

This *can* however be done for the case of two infinitely wide strips placed opposite to each other and carrying current in opposite directions. We shall deal with this briefly.

A one-dimensional problem: two opposite and infinitely wide strips

Suppose that two strips of thickness d at a distance $2a$ extend to an infinite distance in both length and width (see fig. 9). In this system, the current can flow lengthwise (equal but opposite currents in the two strips) with the magnetic field along the width, the current and field depending only on the co-ordinate z perpendicular to the plane of the strips. Even if the condition $\delta_k \ll d$ is not fulfilled this one-dimensional problem is very simple.

Equation (1) becomes $J = dH/dz$. In the region between and outside the strips, $J = 0$ and therefore

Ferromagnetic sphere

In ferromagnetic metals the skin effect occurs in a frequency range completely different from that in the non-ferromagnetic metals. For a metal in which $\mu_r = 5000$ and $\sigma = \frac{1}{5}\sigma_{\text{copper}}$, the value of $\mu\sigma$ is 1000 times greater than it is for copper, so that the skin depth is only 1 mm at frequencies as low as 1 Hz.

[2] V. Belevitch, P. Gueret and J. C. Liénard, Le skin-effet dans un ruban, Rev. HF 5, 109-115, 1962.

In the previous sections we always thought of the alternating current in the conductor as being produced by a current source connected to it. It is also possible, however, to study the skin effect in a conducting body in an alternating electromagnetic field, the current being produced by induction. We shall now consider the case of a ferromagnetic sphere in an alternating magnetic field. An interesting complication is that the tendency of the alternating field and therefore of the flux to be forced outwards by the skin effect is opposed by the ferromagnetism of the sphere, which tends to concentrate the flux. Certain peculiarities in the behaviour of iron particles in a high frequency field may be seen as a conflict between these opposing tendencies [3].

For the sake of simplicity we assume that the material has a permeability μ_r which is independent of the frequency. In reality, μ_r is in general a function of the frequency. Employing the usual notation for complex permeabilities, $\mu_r = \mu_r' + j\mu_r''$, the μ_r' tends to become unity at high values of ω , and μ_r'' exhibits one or more peaks as a function of the frequency, these peaks representing losses. We shall not take these effects into account.

Let us now consider a sphere of radius R in an initially uniform magnetic field (fig. 10). Using Stokes's

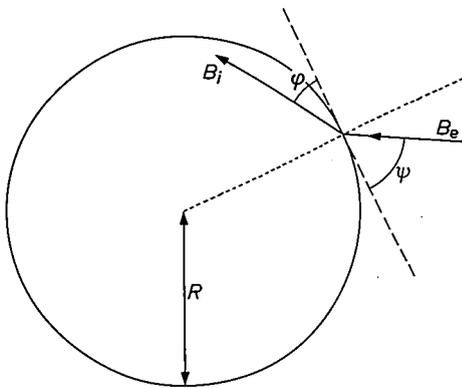


Fig. 10. Refraction of the lines of magnetic induction at the surface of a ferromagnetic sphere.

and Gauss's theorems and equations (1) and (3), one arrives at the well-known boundary conditions that must apply at the surface (the suffixes n and t refer to the normal and tangential component, i and e refer to internal and external):

$$B_{ne} = B_{ni},$$

$$H_{te} = H_{ti} \text{ or } B_{te} = B_{ti} / \mu_r,$$

from which it follows that the "refraction" of the

magnetic flux lines (which are continuous through the surface) is given by:

$$\tan \psi = \mu_r \tan \varphi,$$

where

$$\tan \varphi = B_{ni} / B_{ti}, \quad \tan \psi = B_{ne} / B_{te}.$$

In a static magnetic field, $\tan \varphi$ has a value of the order of unity (the field within the sphere is homogeneous), so that for a ferromagnetic material $\tan \psi \gg 1$: the lines of flux outside the sphere terminate at right angles to the sphere. (For a diamagnetic material, on the other hand, they are substantially tangential to the surface.)

Let us now consider alternating fields of frequencies such that $\delta_k \ll R$: the fields are confined to a layer of thickness δ_k . We may now estimate $\tan \varphi$ to be of the order of δ_k/R , so that:

$$\tan \psi \approx \frac{\mu_r \delta_k}{R}.$$

For a ferromagnetic material we can now distinguish three cases (fig. 11):

- a) $R \ll \delta_k$, "low frequency". The pattern of the flux lines is identical to the static case; it is homogeneous inside the sphere, and outside the sphere it is characterized by flux lines perpendicular to the surface (ferromagnetic pattern).
- b) $\delta_k \ll R \ll \mu_r \delta_k$, "medium frequency". Inside the sphere the field is concentrated in a skin layer, and outside the pattern is still ferromagnetic.
- c) $\mu_r \delta_k \ll R$, "high frequency". Inside the sphere the field is concentrated in a skin layer, and outside the pattern is diamagnetic.

If we introduce a critical frequency defined by σ and R (but independent of μ_r),

$$\omega_c = \frac{2}{\mu_0 \sigma R^2}, \quad \dots \dots \dots (21)$$

the boundaries between regions (a), (b) and (c) are the frequencies given by:

$$(a) \quad \frac{\omega}{\omega_c} \ll \frac{1}{\mu_r},$$

$$(b) \quad \frac{1}{\mu_r} \ll \frac{\omega}{\omega_c} \ll \mu_r,$$

$$(c) \quad \mu_r \ll \frac{\omega}{\omega_c}.$$

In the case of a sphere 1 mm in diameter (i.e. $R = 5 \times 10^{-4}$ m) and $\sigma = \frac{1}{5} \sigma_{\text{copper}} = 1.2 \times 10^7 (\Omega\text{m})^{-1}$, $\omega_c \approx 5 \times 10^5$ rad/s. For such a sphere, with $\mu_r = 5000$, the frequency boundaries $\omega = \omega_c / \mu_r$ and $\omega = \mu_r \omega_c$ of the medium-frequency region become: $\omega = 100$ and $\omega = 2.5 \times 10^9$ rad/s, corresponding roughly to 20 Hz and 400 MHz.

[3] H. B. G. Casimir, Philips Res. Repts. 2, 42-54, 1947.

The difference between the three cases shows up particularly in the power dissipated by the sphere. This depends on μ , σ and ω in different ways in the three cases. If we borrow from magnetostatics the result that (because of demagnetization effects) the magnetic flux density in a sphere with a high μ_r in an originally homogeneous static magnetic field H is $B = 3\mu_0 H$ (i.e. independent of μ_r), the energy dissipation for case (a), $R \ll \delta_k$, may easily be found. Consider an

Let J_m be the amplitude of J ; then $\overline{(\text{Re } J)^2} = \frac{1}{2} J_m^2$. Integration over the sphere gives:

$$P = \frac{6\pi}{5} \omega \mu H_m^2 R^3 \left(\frac{R}{\mu_r \delta_k} \right)^2,$$

where H_m is the amplitude of H .

The dissipation in cases (b) and (c) may be estimated in the following way. The current now flows through a thin broad surface band. This band has a length of

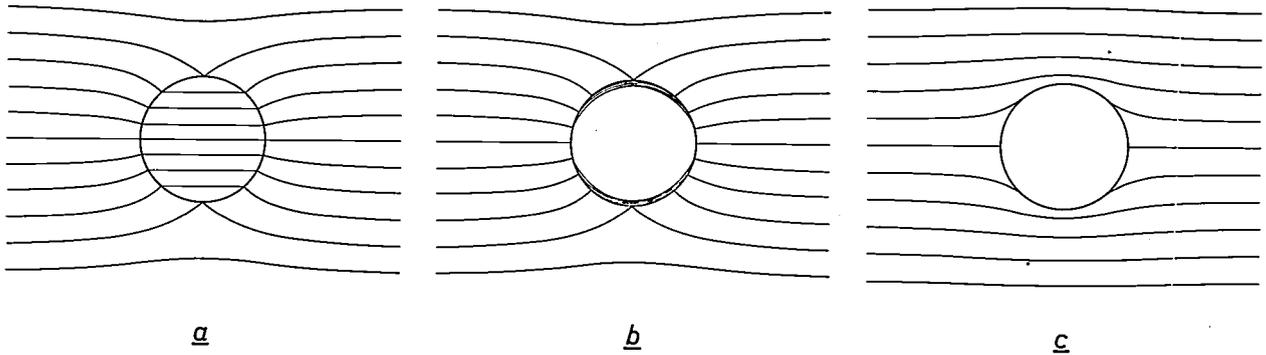


Fig. 11. Lines of magnetic induction inside and outside a ferromagnetic sphere in an initially homogeneous alternating magnetic field in three frequency regions. a) "Low-frequency": $R \ll \delta_k$, b) "medium-frequency": $\delta_k \ll R \ll \mu_r \delta_k$, c) "high-frequency" $\mu_r \delta_k \ll R$.

elementary ring in the sphere as shown in fig. 12. We once more assume that the field and current are proportional to $\exp(j\omega t)$. With Stokes's theorem, equation (2), $J = \sigma E$ and $B = 3\mu_0 H$, we find for the current density in the ring $J = -\frac{3}{2} j\omega \sigma \mu_0 r H$. The actual instantaneous current is $\text{Re } J dr dz$ and the resistance is $2\pi r / \sigma dr dz$, whence the energy dissipation per second is

$$dP = \overline{(\text{Re } J dr dz)^2} \frac{2\pi r}{\sigma dr dz} = \frac{2\pi r}{\sigma} \overline{(\text{Re } J)^2} dr dz.$$

about $2\pi R$ and a cross-section of about $\delta_k R$. The heat developed per second is roughly:

$$P \approx \overline{(\text{Re } J \delta_k R)^2} \cdot \frac{2\pi R}{\sigma R \delta_k} = \frac{\pi J_m^2 R^2 \delta_k}{\sigma}.$$

If Φ is the total flux through the sphere (amplitude Φ_m), then from (2) and Stokes's theorem:

$$2\pi R J = -j\omega \sigma \Phi,$$

so that:

$$P \approx \frac{\omega \Phi_m^2}{2\pi \mu \delta_k}.$$

The flux through the sphere differs in the two cases (b) and (c). In case (b) the external field pattern and hence the flux through the sphere are the same as in the static case, so that $\Phi = \pi R^2 \times 3\mu_0 H$. In case (c) the flux lines are tangential to the surface, so that no demagnetization effects occur. Therefore, in the skin layer, $B = \mu H$ and $\Phi = 2\pi R \delta_k B = 2\pi R H \mu \delta_k$. Substituting for the flux in each case gives:

$$b) P \approx \frac{9\pi}{2} \omega \mu_0 H_m^2 R^3 \left(\frac{R}{\mu_r \delta_k} \right),$$

$$c) P \approx 2\pi \omega \mu_0 H_m^2 R^3 \left(\frac{\mu_r \delta_k}{R} \right).$$

An exact calculation [3] gives the same results for these limiting cases except that the numerical factor

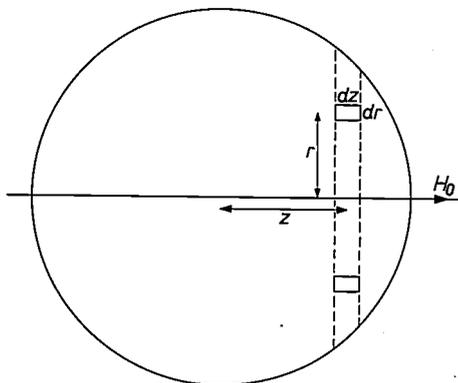


Fig. 12. Annular volume element (radius r at distance z from the centre) of a sphere, used in calculating the power dissipated in case (a), $R \ll \delta_k$. The axis of the ring is parallel to the magnetic field.

$9\pi/2$ in (b) becomes 3π and the factor 2π in (c) becomes $3\pi/2$. If we put

$$6\pi\mu_0 H_m^2 R^3 = W \quad \text{and} \quad \mu_r \delta_k / R = \alpha,$$

and once more introduce ω_c according to (21), and using (10), the final result is:

For case (a), i.e. $R \ll \delta_k$ or $\omega/\omega_c \ll 1/\mu_r$:

$$P = \frac{1}{5}\omega W \mu_r \alpha^{-2} = \frac{1}{5}\omega_c W \left(\frac{\omega}{\omega_c}\right)^2 \dots (22a)$$

For case (b), i.e. $\delta_k \ll R \ll \mu_r \delta_k$ or $1/\mu_r \ll \omega/\omega_c \ll \mu_r$:

$$P = \frac{1}{2}\omega W \alpha^{-1} = \frac{1}{2}\omega_c W \mu_r^{-1/2} \left(\frac{\omega}{\omega_c}\right)^{3/2} \dots (22b)$$

For case (c), i.e. $\mu_r \delta_k \ll R$ or $\mu_r \ll \omega/\omega_c$:

$$P = \frac{1}{4}\omega W \alpha = \frac{1}{4}\omega_c W \mu_r^{1/2} \left(\frac{\omega}{\omega_c}\right)^{1/2} \dots (22c)$$

In fig. 13 P is plotted against ω/ω_c for each of these cases, on logarithmic scales, for a given value of μ_r , namely μ_{r1} . The factors $\frac{1}{5}$, $\frac{1}{2}$ and $\frac{1}{4}$ have been omitted in plotting this diagram.

A peculiar phenomenon resulting from this interplay of skin effect and ferromagnetism is "temperature hysteresis" [4]. If a conducting body is brought into a high-frequency field, the temperature assumes a value such that the heat developed is equal to the heat radiated. The heat radiated is highly temperature-dependent ($\propto T^4$ or T^5). In the ferromagnetic case, the following phenomenon can now occur. The body is gradually brought near to a high-frequency coil. Initially the temperature increases gradually, and the body begins to glow a pale red. At a certain point the body will suddenly become white-hot. If it is now gradually removed again, it will continue to glow white until far beyond the point where it first became white hot and then suddenly reverts to the pale red colour.

This behaviour may be explained in the following way. As a result of its dependence upon μ_r , the heat developed, P , is highly temperature-dependent in the region of the Curie temperature T_c . Let us suppose that μ_r has a high value μ_{r1} when $T < T_c$ and is unity when $T > T_c$. P is also plotted for $\mu_r = 1$ in fig. 13 (then region (b) is non-existent). We see from this that there is a frequency range in which P increases abruptly when T is raised above T_c . We consider a frequency in this range.

Fig. 14 gives a diagram of the heat developed, P , and the heat radiated, P_e , plotted against the temperature on logarithmic scales. The curve for P_e is a straight line (with a slope of 4 or 5). P exhibits an abrupt rise at

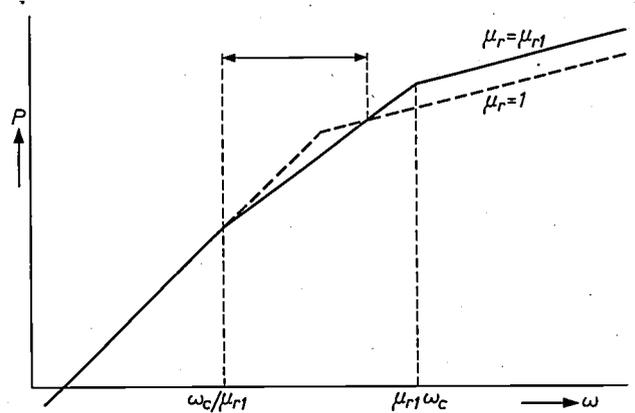


Fig. 13. The power P dissipated in a ferromagnetic sphere in an alternating magnetic field as a function of the angular frequency ω , both plotted on a logarithmic scale, according to equations (22a, b, c) (the factors $\frac{1}{5}$, $\frac{1}{2}$ and $\frac{1}{4}$ being omitted). The solid line applies for $\mu_r = \mu_{r1} \gg 1$ ($T < T_c$), and the dashed line for $\mu_r = 1$ ($T > T_c$). The double arrow indicates the range in which "temperature hysteresis" may be expected.

the Curie point. If the amplitude of the high-frequency field is increased, P shifts upwards, or, effectively, P_c shifts downwards with respect to P . The temperature that the body assumes is given by the intersection of the two curves. As the amplitude increases the cycle passes

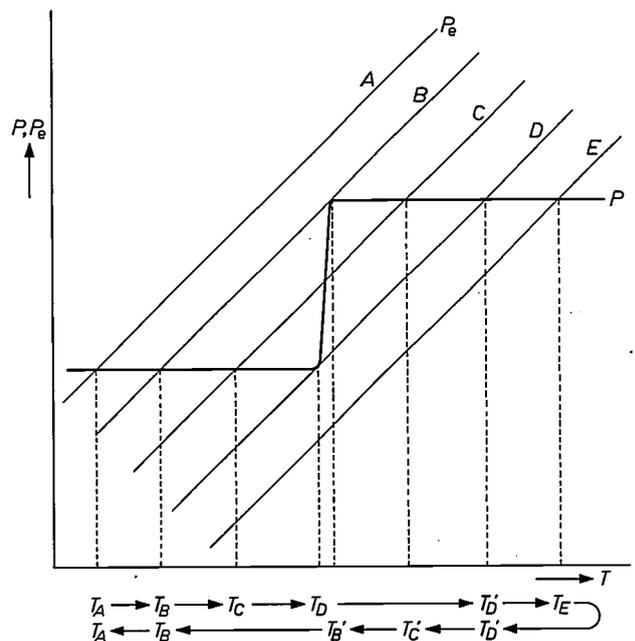


Fig. 14. The heat developed, P , and the heat radiated, P_e , as a function of the temperature T (logarithmic scales) for a ferromagnetic sphere in an alternating magnetic field with a frequency at which P rises abruptly when μ_r jumps from μ_{r1} to 1 at the Curie temperature (in the range indicated by the double arrow in fig. 13). A point of intersection of P and P_e defines a steady state. When the intensity of the alternating field is increased, P shifts upwards, and thus P_e shifts downwards with respect to P , so that situations A, B, C, D and E arise successively. On traversing the full cycle $A \dots E \dots A$, the temperatures corresponding to situations between B and D are different on the forward journey from those on the return journey.

[4] See: J. L. Snoek, New developments in ferromagnetic materials, Elsevier, Amsterdam 1947.

through situations *A*, *B*, *C*, *D* and *E*. Between *B* and *D* there are two stable equilibrium temperatures (see, for example, situation *C*). As the amplitude increases, the temperature rises abruptly from T_D to $T_{D'}$ at the moment that the line *D* leaves the lower bend in curve *P*. As the amplitude decreases the temperature remains high until, as *B* leaves the upper bend in curve *P*, it drops back from $T_{B'}$ to T_B . The frequency range in which temperature hysteresis may be expected (indicated in fig. 13 by a double arrow) is given by the conditions:

$$P(a, \mu_r = 1) \gg P(b, \mu_r = \mu_{r1})$$

and

$$P(c, \mu_r = 1) \gg P(b, \mu_r = \mu_{r1}),$$

where the *a*, *b* and *c* refer to our three frequency regions. Using (22a, b, c) we find:

$$\frac{25}{4\mu_{r1}} \ll \frac{\omega}{\omega_c} \ll \frac{1}{2} \sqrt{\mu_{r1}}.$$

For our sphere (diameter 1 mm, $\sigma = 1.2 \times 10^7 \Omega^{-1} \text{ m}^{-1}$, $\mu_{r1} = 5000$) this means that the frequency *f* must be in the range $100 \text{ Hz} \ll f \ll 3 \text{ MHz}$.

Summary. This first part of three articles on the skin effect contains a general introduction followed by a discussion of the current distribution for several configurations. The essence of the skin effect is illustrated by the case of a number of parallel wires in one plane, connected in parallel. The standard problem of the cylindrical wire is used to introduce the concept "classical skin depth". The problem of the distribution of the current over a conductor of arbitrary shape is stated in general terms. If the skin layer is thin, the problem of the distribution in the surface is reduced to a potential problem. For a cylindrical wire of any

cross-section, the surface current distribution is the same as the charge distribution over the surface of a charged conductor of the same shape. This consideration is used in the problem of the flat strip. The problem becomes purely one-dimensional for two strips which together form a flat coil. Finally the case of a ferromagnetic sphere in a magnetic alternating field is discussed. The combination of skin effect and ferromagnetism can lead to "temperature hysteresis", an effect in which the temperature variation of the sphere is hysteretic with respect to an increasing and a decreasing field amplitude.

An experimental reflex klystron for 1.5 mm wavelength

The series of millimetre-wave reflex klystrons developed at Philips Research Laboratories^[1] was extended some time ago by the addition of an experimental tube for a wavelength of 1.5 mm (fig. 1), which can give an output power of about 25 mW. The tube can be mechanically tuned from about 200 GHz to about 225 GHz or more (fig. 2). With electronic tuning by variation of the reflector voltage, as used in frequency modulation, the 3 dB points are about 200 MHz apart.

The successful development of the new tube is primarily due to the availability of cathodes of much higher permissible current density than previously ob-

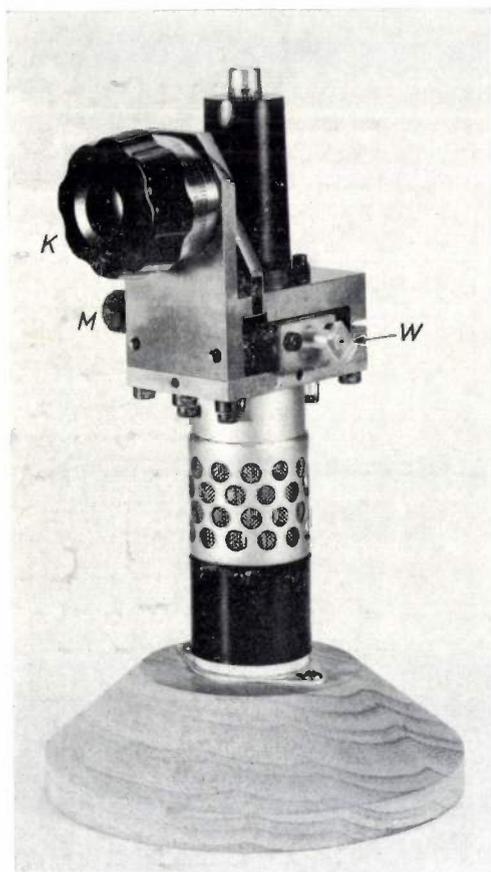


Fig. 1. Experimental reflex klystron for 1.5 mm wavelength. *K* tuning knob. *M* knob for adjusting the matching plunger (fig. 3). *W* output waveguide. The reflector connection is at the top of the tube.

The principal electrical data are:

beam voltage	2500 V
beam current	~ 14 mA
reflector voltage	-20 to -600 V
cathode current density, average	12 A/cm ²
maximum, at the centre	27 A/cm ²
output power P_0 at centre of the frequency band	~ 25 mW
mechanical tuning range	~ 15%
electronic tuning range	about 200 MHz

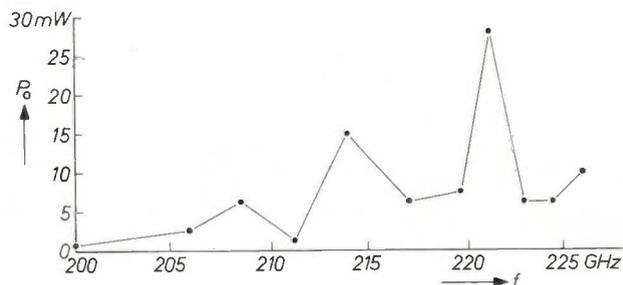


Fig. 2. Output power P_0 as a function of frequency f with mechanical tuning, measured on a randomly selected tube. The maximum value of P_0 is 25 mW or more.

tainable, i.e. the osmium dispenser cathodes^[2], and also to a simplification in the design of the r.f. section, and to the finding of a good solution to the heat dissipation problem. Other contributory factors were the progress made in the art of making extremely small resonant cavities, and the fact that it has been found possible to use higher electron transit times in the reflector space than have previously seemed feasible.

Fig. 3 shows a diagram of the cross-section of the new tube. The "tuning plate", which is also shown separately in fig. 4, is simpler in design than that of the 2.5 mm tube but is electrically approximately equivalent. The design chosen is the result of extensive investigations on scaled-up versions of the r.f. section of the 2.5 mm tube (the scaled-up versions were designed for the 3 cm wavelength range as plenty of test gear is available and the dimensions of equipment are conve-

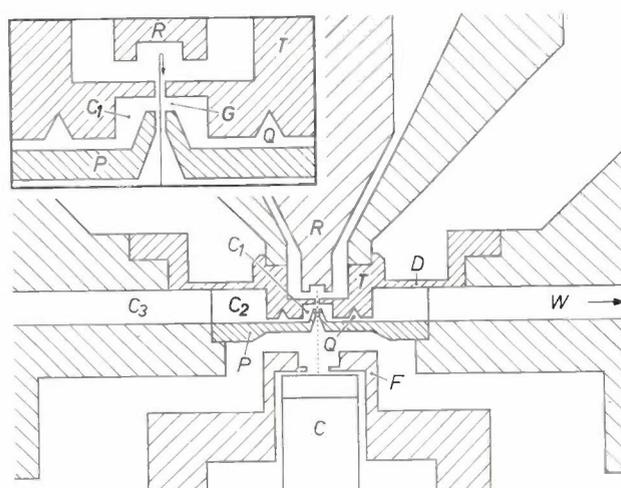


Fig. 3. Cross-section of the central section of the tube, along the axis. *C* cathode. *F* focusing electrode. *T* tuning plate, with a flexible diaphragm *D*. *R* reflector. *P* copper plate with cone. *C*₁ resonant cavity. *G* interaction gap (see inset). *C*₂ annular space. *C*₃ waveguide terminated by matching plunger (at the left outside the figure). *C*₂, *C*₃ and the gap *Q* (a $\frac{1}{4}\lambda$ transformer) together constitute the matching transformer between *C*₁ and the output waveguide *W*.

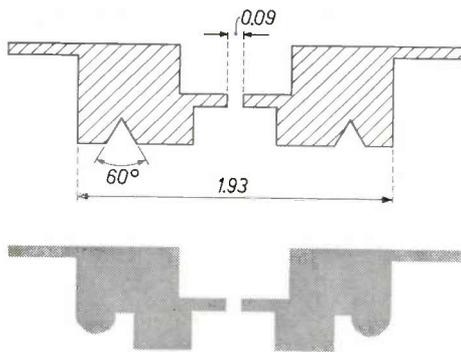


Fig. 4. Tuning plate of the new tube (top). Its shape is much simpler than that for the 2.5 mm tube (below; scaled down for 1.5 mm for comparison).

nient). These experiments showed that the new tube has a wider tuning range than would have been obtained if the tuning plate of the 2.5 mm tube had been linearly scaled down (fig. 5). Because of its simpler shape the tuning plate of the new tube can in fact be made more easily than that of the 2.5 mm tube and with better reproducibility.

The electrical characteristics of the gap Q (this is in fact a radial waveguide between the tuning plate T and the plate P , fig. 3), i.e. the gap connecting cavities C_1 and C_2 , are of very great importance. In the first place, the signal excited in C_1 has to be able to travel along this gap without serious attenuation. On the other hand, however, the characteristics should be such that the beam cannot excite oscillations in the complicated resonant cavity formed by C_1 , Q and C_2 ; the tube could then oscillate inefficiently and at far too low a frequency.

These requirements have to be met by using a tuning plate which is sufficiently simple in design that the extremely small dimensions present no insuperable technical difficulties.

It was also found that the cone of the lower cavity wall could not be designed by simple linear scaling from the 2.5 mm tube. A cone scaled from the 2.5 mm version would get far too hot, as almost all of the beam power is dissipated in this part of the tube. In the development of an earlier experimental 2 mm tube [3] a cone had been designed which was relatively larger than the cone of the 2.5 mm tube, but which had practically the same electrical properties. This cone was easily able to handle the power dissipated in it (40 W), and was therefore suitable for linear scaling for a 1.5 mm tube (fig. 6).

To make the cone of the lower cavity wall for the 1.5 mm tube, a new process had to be developed; this process included extrusion and spark machining [4]. New methods were also found necessary for making other small components, in particular the reflector.

[1] B. B. van Iperen, Reflex klystrons for wavelengths of 4 and 2.5 mm, Philips tech. Rev. 21, 221-228, 1959/60.

[2] P. Zalm and A. J. A. van Stratum, Osmium dispenser cathodes, Philips tech. Rev. 27, 69-75, 1966 (No. 3/4).

[3] Not published.

[4] This work was done by M. A. van den Ban, F. M. J. Onstenk and the late J. W. Rommerts of this laboratory.

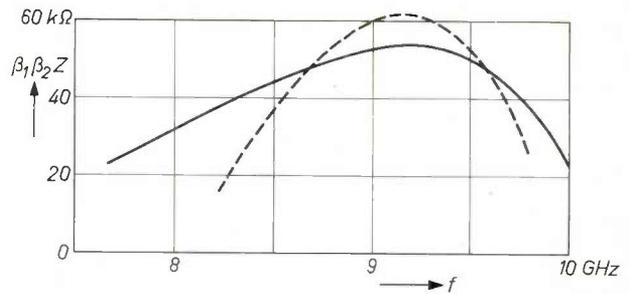


Fig. 5. Plot of $\beta_1 \beta_2 Z$ against frequency. This quantity is a measure of the output power for scaled-up versions (for 3 cm wavelength) of the r.f. section of the 2.5 mm tube. (β_1 and β_2 are the coupling factors between the forward and reflected beams respectively and the cavity; Z is the impedance of the unloaded cavity at resonance.) One version was fitted with a tuning plate of the shape used in the 1.5 mm tube (full line), the second with a plate of the other shape (dashed line).

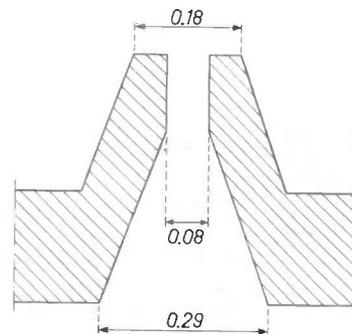


Fig. 6. Scale drawing of the cone of the lower cavity wall. All dimensions are given in mm.

Since there is a voltage of about 3000 V between the tuning plate and the reflector, it was thought inadvisable to make the distance between these components smaller than in the tubes for 4.0 and 2.5 mm. As in these earlier tubes, the spacing is 0.17 mm. The ratio N of the electron transit time in the reflector space to the period of the r.f. voltage across the interaction gap is therefore a little higher at the same reflector voltage. In principle a high N should correspond to a high output power. The value of N cannot be increased indefinitely, however, because in the long run there would be debunching owing to space-charge forces and differences in the transit times of electrons describing different trajectories. Reflecting the beam is therefore ineffective if it takes too long. At the value of N in the 1.5 mm tube described here — found from measurements to be $11\frac{3}{4}$ — these undesirable effects are apparently not yet large enough to counteract the advantages of the high value of N .

G. H. Plantinga
Th. J. Westerhof

Ir. G. H. Plantinga, now with the Philips Semiconductor Works, Nijmegen, was formerly with Philips Research Laboratories; Th. J. Westerhof is with Philips Research Laboratories, Eindhoven.

Recent scientific publications

These publications are contributed by staff of laboratories and plants which form part of or co-operate with enterprises of the Philips group of companies, particularly by staff of the following research laboratories:

Philips Research Laboratories, Eindhoven, Netherlands	<i>E</i>
Mullard Research Laboratories, Redhill (Surrey), England	<i>M</i>
Laboratoires d'Electronique et de Physique Appliquée, Limeil-Brévannes (S.O.), France	<i>L</i>
Philips Zentrallaboratorium GmbH, Aachen laboratory, Weisshausstrasse, 51 Aachen, Germany	<i>A</i>
Philips Zentrallaboratorium GmbH, Hamburg laboratory, Vogt-Kölln-Strasse 30, 2 Hamburg-Stellingen, Germany	<i>H</i>
MBLE Laboratoire de Recherche, 2 avenue Van Becelaere, Brussels 17 (Boitsfort), Belgium.	<i>B</i>

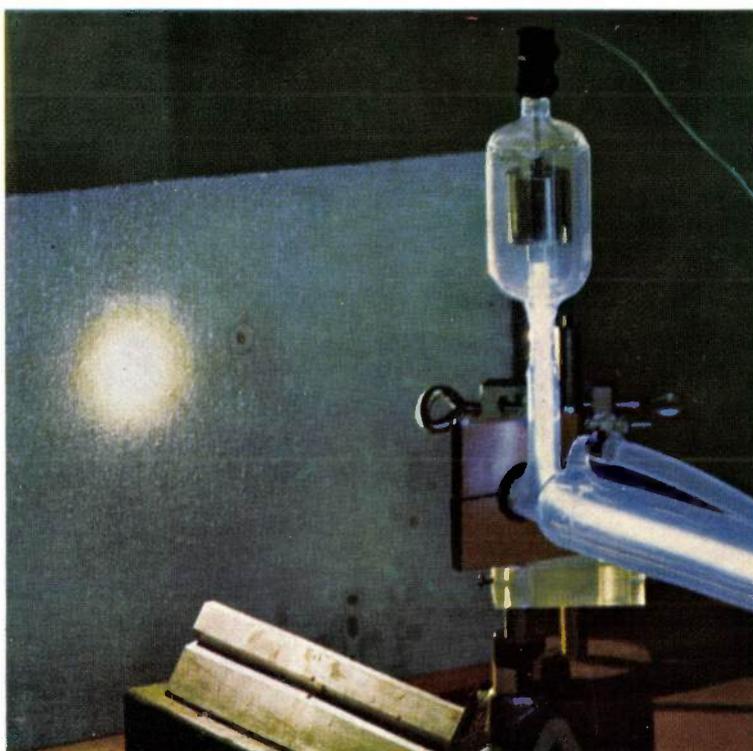
Reprints of most of these publications will be available in the near future. Requests for reprints should be addressed to the respective laboratories (see the code letter) or to Philips Research Laboratories, Eindhoven, Netherlands.

- D. Andrew:** A cycloidal path mass spectrometer with wire-wound electric field structure. 1965 Trans. 3rd Int. Vacuum Congress, Stuttgart, Vol. 2, Part 2, p. 527-533; Pergamon Press, Oxford 1967. *M*
- K. H. Beckmann:** Zur Bildung oxidischer Deckschichten auf Germaniumanoden in alkalischen Elektrolyten. Berichte Bunsenges. phys. Chemie 70, 842-849, 1966 (No. 8). *H*
- R. Bleekrode:** Pressure dependence of rotational intensity distribution in C₂ "Swan" emission from low-pressure oxyacetylene flames. J. chem. Phys. 45, 3153-3154, 1966 (No. 8). *E*
- G. Buchta:** Miniaturized broadband E-Tee circulator at X-band. Proc. IEEE 54, 1607-1608, 1966 (No. 11). *H*
- J. Burmeister:** Kristallzüchtung von Wüstit, Fe_{1-x}O. Mat. Res. Bull. 1, 17-26, 1966 (No. 1). *A*
- H. B. G. Casimir:** Note on macroscopic theory of optical rotation and double refraction in cubic crystals. Philips Res. Repts. 21, 417-422, 1966 (No. 6). *E*
- R. David & A. Venema:** Pump speed measurements in a new type of cryopumped vacuum system. 1965 Trans. 3rd Int. Vacuum Congress, Stuttgart, Vol. 2, Part 3, p. 577-579; Pergamon Press, Oxford 1967. *E*
- M. J. Folkes & G. Winsor:** Size and shape effects in magnetic thin film storage elements. Int. J. Control 3, 513-533, 1966 (No. 6). *M*
- Y. Genin:** Principes du guidage perturbatif sub-optimal et méthodes d'approximation polynomiale des matrices de guidage. Rev. MBL 9, 86-103, 1966 (No. 2). *B*
- C. A. A. J. Greebe & W. F. Druyvesteyn:** Local theory for helicon resonances in flat metal boxes. Philips Res. Repts. 21, 423-431, 1966 (No. 6). *E*
- N. Hansen & W. Littmann:** Über die Bestimmung der Haftwahrscheinlichkeit von Gasen an reinen Metalloberflächen. 1965 Trans. 3rd Int. Vacuum Congress, Stuttgart, Vol. 2, Part 2, p. 465-471; Pergamon Press, Oxford 1967. *A*
- Y. Haven, A. Kats & J. S. van Wieringen:** Optical absorption and paramagnetic resonance of colour centres in X-rayed α -quartz containing germanium. Philips Res. Repts. 21, 446-476, 1966 (No. 6). *E*
- E. E. Havinga:** Magnetic interactions between Mn³⁺ ions in perovskites. Philips Res. Repts. 21, 432-445, 1966 (No. 6). *E*
- E. Kooi:** The surface charge in oxidized silicon. Philips Res. Repts. 21, 477-495, 1966 (No. 6). *E*
- E. Kooi & M. V. Whelan:** On the role of sodium and hydrogen in the Si-SiO₂ system. Appl. Phys. Letters 9, 314-317, 1966 (No. 8). *E*
- D. J. Kroon:** A tail-Dewar for liquid helium. J. sci. Instr. 43, 831, 1966 (No. 11). *E*
- N. D. Richards:** The design of large cryotron memories. IEEE Trans. on magnetics MAG-2, 394-398, 1966 (No. 3). *M*
- R. P. van Stapele, J. C. M. Henning, G. E. G. Harde-man & P. F. Bongers:** Direct measurement of weak exchange interactions in pairs of Co²⁺ ions in Cs₃ZnCl₅. Phys. Rev. 150, 310-314, 1966 (No. 1). *E*
- T. L. Tansley:** Forward bias current-voltage characteristics for a heterojunction in which tunnelling dominates. Phys. Stat. sol. 18, 105-112, 1966 (No. 1). *M*

Sealed-off high-power CO₂ lasers

W. J. Witteman

No longer is the laser "a device in search of an application": it has evolved in a very short time from a laboratory novelty into an industrially produced device of considerable practical importance, which has already found numerous applications and has great potentialities. Its present applications include such diverse topics as eye-surgery, machining, and the transmission of information. The article below describes a development in the field of high-power infra-red lasers (10 μm) which has made it possible to produce easily manageable instruments.



Introduction

A gas laser may be regarded as an optical resonant cavity or interferometer filled with a gas that serves as active medium. By an active medium we mean here a medium in which, for any pair of excited states of the atoms or molecules, the population density of the upper energy level is greater than that of the lower one; in other words there are more atoms or molecules per unit volume in the higher energy state than in the lower one, which is just the opposite situation to that of thermal equilibrium. This "inversion" of energy level populations can be brought about, under favourable conditions, in a gas discharge. A gas laser may therefore be a gas discharge tube, both ends of which are closed by flat or spherical mirrors; these mirrors form the boundaries of the resonant cavity.

Dr. W. J. Witteman is with Philips Research Laboratories, Eindhoven.

Now if the radiation whose energy quanta correspond to the difference in energy between the two levels is reflected strongly enough by the mirrors, and if the frequency of this radiation is one of the natural frequencies of the resonant cavity, laser action may be initiated. In this process the transitions from the upper to the lower level, giving rise to the emission of radiation quanta, do not take place spontaneously but under the influence of electromagnetic radiation of the same wavelength already present (stimulated emission). Radiation later generated is thus in phase with the radiation already present, and a coherent beam is obtained which continues to increase in intensity until the losses have become equal to the output. Part of this beam, called the laser beam, can be "coupled out", i.e. allowed to escape from the resonant cavity. The beam is generally coupled out by means of semi-transparent mirrors.

Until recently all laser action which had been discovered was related to transitions between *electron* levels. In lasers of this kind the average power of the laser beam is generally a few milliwatts and the efficiency (the ratio of this power to the electrical power expended in the discharge) is low. In gas lasers of this type (e.g. the helium-neon laser) the efficiency is of the order of 10^{-4} .

There are three reasons for the low efficiency of laser action based on electron transitions: 1) The difference in energy between the levels concerned is small compared with the excitation energy of the upper level. 2) This excitation energy is many times higher than the average energy of the free electrons in a gas discharge. This means that only a small fraction of the collisions between these electrons and atoms is effective; only the electrons whose energy is greater than or equal to the excitation energy can excite the required level. In most collisions electron energy is converted into thermal energy of the gas or non-relevant levels are excited. 3) Owing to the Doppler effect resulting from the thermal movement of the atoms, the radiation originating from transitions between two specific electron levels shows a certain spread in frequency. Of all the excited atoms only a small fraction will have a velocity such that the frequency of the radiation they would emit in returning to the ground state is equal to one of the resonant frequencies of the optical resonant cavity. If the upper level has a long life, then for most of the excited atoms an instant will be reached when, as a result of collisions, the thermal velocity will have exactly reached an appropriate value. As the process involves *stimulated emission*, the transition will then take place immediately, so that ultimately nearly all the excited particles contribute to the laser process. With electron transitions, however, the life of the upper level is usually so short (about 10^{-7} s) that the transition will generally have taken place by spontaneous emission before the thermal velocity has been able to reach an appropriate value.

These phenomena, which considerably reduce the efficiency, are far less troublesome when vibrational and rotational levels of molecules are excited. Vibrational and rotational transitions lie as a rule in the infra-red, and the energy levels are therefore much more favourably situated for laser action. The excitation energy of a vibrational level acting as an upper level is much lower than the average electron energy, so that nearly all electrons in the discharge can contribute to the excitation of the upper level. Furthermore, the life of a populated upper level is so long that, when laser action occurs, the transition back to the lower level takes place almost entirely by means of stimulated emission; thermal vibrational relaxation and

spontaneous emission are negligible here. In a laser containing a suitable molecular medium it is thus possible to generate a very intense laser beam with good efficiency (up to 20% and beyond). For example, with a laser of this kind about 2 metres long a beam with a continuous power of 100 W can be generated. If even higher power is required it is simply necessary to increase the length of the tube.

Lasers of such high power can find very interesting applications in industry. The laser beam with its low divergence can be concentrated by means of mirrors or lenses into an area which is only a few wavelengths across. Depending on the reflection and the local thermal conductivity of the material, extremely high localized temperatures can be reached. In the hypothetical case where reflection and thermal conductivity are completely absent, a temperature higher than 50 000 °K would be obtainable. In practice, materials with a high melting point, such as tungsten, mica, quartz and asbestos, can easily be cut with a focused laser beam.

In cases where exceptionally high temperatures are not required, a laser heat source may be preferable to more conventional types. A laser may be particularly useful, for example, where the heat generation has to remain strictly limited to the region of interest. Moreover, it does not give rise to impurities and does not exert a pressure on the workpiece. Infra-red lasers are also coming into use for "diagnostic" purposes in plasma research, and their use is being considered for communications applications. Quite apart from very high intensity and efficiency, these lasers can be very useful for all kinds of infra-red research, particularly in the far infra-red ($\lambda > 20 \mu\text{m}$); in the past relatively little attention has been paid to this region owing to the lack of sources of sufficiently high intensity.

Partly because of this, there has been intensive research in the last two years on molecular lasers. In 1964 it was discovered that during a pulsed discharge in water vapour, there was a laser action [1] which could be attributed to rotational and vibrational transitions. Subsequent investigations showed that, given a suitable laser design and using certain other molecular gases, *continuous* laser action could be obtained [2]. If the primary objective is high power, carbon dioxide has been found to be the most suitable gas. In early experiments with this gas, which delivers infra-red radiation of about $10.6 \mu\text{m}$, a continuous beam of 1 mW was produced [3].

It was later found that the power of CO₂ lasers can be substantially increased by adding nitrogen and helium to the gas and by carefully selecting the gas-discharge and design parameters, such as the length and diameter of the resonant cavity, the radius of curva-

ture of the mirrors and the method of coupling out the beam. A continuous high-power beam could only be obtained with such a laser, however, if the gas mixture was continuously replenished during operation. It was therefore necessary to use an *open system*. For practical applications this was of course less attractive.

A short time ago we found that it is quite possible to achieve a high output power with a *sealed-off system* without having to accept a shorter operating life. This requires the addition of water vapour to the CO₂ as well as the gases already mentioned, and the use of special materials for the electrodes and for the laser tube. When the correct dimensions are chosen an efficiency varying between 16 and 21% can be obtained with a sealed-off system, depending on the radiant output power, while a continuous output of about 1 W/cm³ gas can be obtained. Life tests have shown that the addition of water vapour can increase the useful life of a sealed-off CO₂ laser — i.e. the time in which the intensity decreases to 90% of its initial value — to more than 1000 hours.

In this article we shall describe some experimental sealed-off CO₂ lasers which we have made and investigated. Before doing so we shall present a qualitative treatment of the processes underlying the operation of a CO₂ laser and of the part played in these processes by the three gases added to the CO₂ [4].

The inversion mechanism

Laser action at vibrational-rotational levels

The molecular vibrations of CO₂ are superpositions of three fundamental vibrations (*fig. 1*): the symmetrical valence vibration, the bending vibration and the asymmetrical valence vibration. The vibrational modes of the lowest excited states of these vibrations are denoted by 10⁰0, 01¹0 and 00⁰1 respectively [5]. The energy level diagram of a few vibrational states is shown in *fig. 2*. Under certain gas discharge conditions these molecular vibrations are excited. The excited states thus produced are not permanent, for as a result of collisions between the molecules and of spontaneous emission the molecule returns to its ground state. The more easily the excitation takes place and the longer the life of a particular vibrational state, the greater will be the fraction of molecules found in this state. In other words, given stationary external conditions this energy level is then found to have a higher population density (or, for brevity: population). The gas discharge conditions can now be chosen in such a way as to make the population of the 00⁰1 level greater than that of the other levels shown in *fig. 2*. This is chiefly due to the relatively long life of the vibrational state 00⁰1. An active medium can thus be obtained.

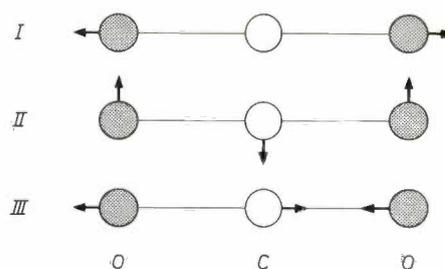


Fig. 1. Diagrammatic representation of the three fundamental vibrations of a carbon dioxide molecule. *I* symmetrical valence vibration. *II* bending vibration. *III* asymmetrical valence vibration.

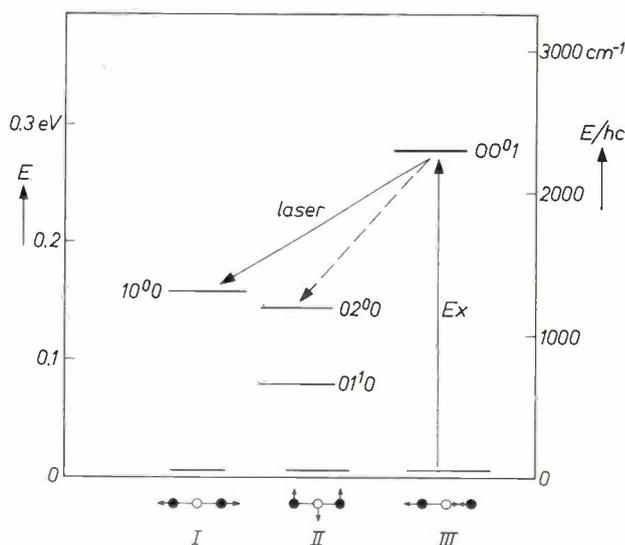


Fig. 2. Energy level diagram of the three vibrational oscillations (cf. *fig. 1*) of carbon dioxide in the case where the electrons are in the ground state. The left-hand vertical scale gives the energy E in eV, and the right-hand one shows E/hc (h is Planck's constant and c the velocity of light) in cm^{-1} . Only two or three of the lowest vibrational levels are shown. The vertical arrow Ex represents the excitation; the fully-drawn arrow between the levels of the states 00⁰1 and 10⁰0 indicates the transition responsible for the laser action (wavelength approximately 10.6 μm). Laser action can also occur at the transition indicated by the dashed arrow.

The laser action found for CO₂ is based on transitions between the 00⁰1 state as upper level and the 10⁰0 or 02⁰0 state as lower level. These transitions are indicated by arrows in *fig. 2*. In normal circum-

[1] A. Crocker, H. A. Gebbie, M. F. Kimmitt and L. E. S. Mathias, *Nature* **201**, 250, 1964.

[2] W. J. Witteman and R. Bleckrode, *Phys. Letters* **13**, 126, 1964.

[3] C. K. N. Patel, *Phys. Rev.* **136**, A 1187, 1964.

[4] For a more quantitative treatment see W. J. Witteman, *Philips Res. Repts.* **21**, 73, 1966 (No. 2).

[5] In the convention for denoting the vibrational state of a molecule the three numbers represent the vibrational quantum numbers of the vibrations *I*, *II* and *III* respectively. The superscript of the middle number is the quantum number of the angular momentum of the C-atom. The vibration need not necessarily take place in a plane; the atoms may describe circular or elliptical orbits, the O-atoms describing their orbits in a direction opposite to that of the C-atom. If the molecule is not in rotation, the total angular momentum of the O-atoms is equal and opposite to that of the C-atom.

stances the transition indicated by the fully-drawn arrow is the dominant one. This is so because the relevant transition probability is greatest while at the same time, owing to a remarkable coincidence, the inversions of the populations of the levels cannot differ much from each other, in both transitions. This is connected with the fact that the two lower levels are at nearly the same height. If one of them should become more populated than the other during the operation of the laser, this inequality will very quickly be removed as a result of collisions ("coupling between populations of the levels"). Although during the laser process based on the transition $00^01 \rightarrow 10^00$ the relevant inversion becomes smaller, the strong coupling between the two lower levels will also reduce the inversion of the transition $00^01 \rightarrow 02^00$, which will thus remain roughly equal to the inversion of the laser transition. Only at low gas pressure, where the coupling of the lower levels becomes much weaker due to the relatively low number of collisions, can any appreciable laser action be obtained at the transition $00^01 \rightarrow 02^00$ [3] as well as at the other one. The operation of our lasers, in which the gas pressure is relatively high, is therefore based almost entirely on the transition $00^01 \rightarrow 10^00$. As we saw in figs. 1 and 2, in this laser process a vibrational quantum of the excited asymmetrical valence vibration of a molecule is transferred as a smaller quantum to the symmetrical valence vibration of the same molecule, while the difference in energy is converted into radiation.

Another means of suppressing laser action at unwanted transitions consists in the application of selective reflection, arrangements being made to ensure that only the wanted radiation is reflected by the mirrors bounding the resonant cavity [9]. With this method, laser action can even be obtained at transitions between higher excited vibrational levels.

In addition to the vibrations we have to consider the rotations of the gas molecule. In CO_2 the spacing of the rotational levels is roughly 1000 times smaller than that between vibrational levels. For every vibrational state there are many populated rotational levels. Transitions like those illustrated diagrammatically in fig. 2 by an arrow are in fact always transitions between a rotational level belonging to the one vibrational state and a rotational level belonging to the other ("vibrational-rotational transitions"), the rotation quantum number j of the lower level being always greater by unity than that of the upper level (P -branch of the spectrum [7]). In general the laser radiation will not therefore be monochromatic, but will consist of a number of spectral lines with wavelengths roughly 20 nm (200 Å) apart. Since the rotational states interchange very easily, the population of the rotational levels belonging to a particular vibrational state can be defined

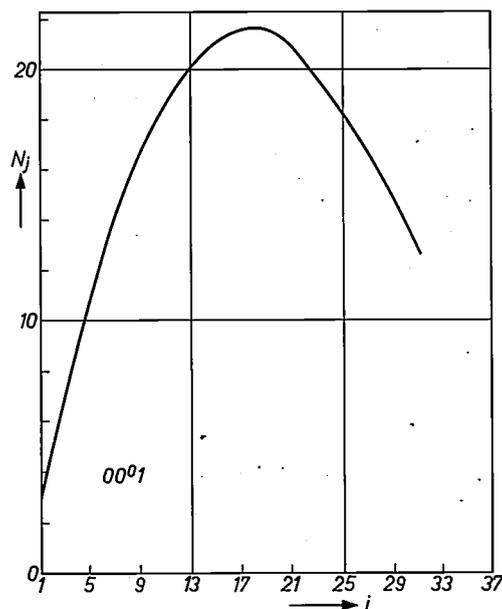


Fig. 3. Population density N_j of the rotational levels belonging to the vibrational state 00^01 (in arbitrary units) as a function of the rotational quantum number j .

by Boltzmann's equation (fig. 3). The spectral lines with the greatest intensity are usually due to the transitions $P(j=18)$, $P(20)$ and $P(22)$ of the $00^01 - 10^00$ band of the vibration-rotation spectrum of CO_2 .

In certain forms of laser only a few rotational levels are suitable for laser action. Owing to the high rate at which the transitions between individual rotational levels take place and the relatively long life of the vibrational state of the upper level, the energy present in the other rotational levels can nevertheless be made available for stimulated emission by means of rotational transitions. We have therefore found that the total beam intensity is independent of the number of rotational levels involved in the laser action.

Addition of nitrogen

It is evident that the emission of radiation increases when more molecules with an excited upper level are produced per unit time. A high emission can be achieved by adding nitrogen to the carbon dioxide gas; at a collision between a vibrationally excited N_2 molecule and a CO_2 molecule the N_2 molecule readily gives up its vibrational energy and the CO_2 molecule becomes vibrationally excited [8]. In practice, therefore, the gas discharge conditions are arranged in such a way that it is mainly the nitrogen vibrations that are excited.

The high efficiency of the CO_2 excitation via excited nitrogen is due to the combination of the following three favourable circumstances. 1) Nitrogen molecules in the ground state are easily excited into a vibrational state by electrons of 2 to 3 eV [9]. 2) At a collision they readily transfer their energy to CO_2 , since the vibration

quanta of nitrogen are virtually identical with those of the 00⁰1 vibrational mode of CO₂. 3) They do not, on the other hand, readily lose their energy in other ways: in the first place, the vibrational energy is not easily converted into translational energy, and furthermore there is no electric dipole moment that might give rise to the spontaneous emission of radiation. It has been found experimentally that the addition of nitrogen makes the intensity of a CO₂ laser beam several times greater.

Addition of water vapour

In a steady state the radiation emitted also increases when the population of the lower level falls. Steps must therefore be taken to ensure that molecules excited into the 10⁰0 vibrational state by the emission of a radiation quantum return as quickly as possible to the ground state.

If there are no other gases than nitrogen and carbon dioxide present, these molecules are returned to the ground state by collisions (collision relaxation). In this process the vibrational energy of the symmetrical valence vibration is first transferred to the bending vibration after which, again as a result of a collision, the energy of the bending vibration is imparted to the translational movement of the colliding molecules [10]. Compared with the other processes essential to laser action (excitation of nitrogen, energy transfer from nitrogen to carbon dioxide and stimulated emission), this relaxation process is the slowest and therefore, in the given conditions, determines the intensity of the radiation emitted. It has been calculated that on an average no fewer than about 5×10^4 collisions are required for the transition of one CO₂ molecule from the vibrational state 10⁰0 to the ground state.

In the introduction we noted that the output power can be improved either by using a flowing gas or by adding water vapour. In both cases the depopulation of the lower level is accelerated; when a flowing gas is used this acceleration is simply due to the fact that a relatively large number of molecules with a populated lower level are removed. (Moreover, we think that the gas flow is useful for the removal of harmful impurities released from the glass wall and the electrodes during the gas discharge, and for replenishing the gas that dissociates during the discharge.)

In a certain sense, the "water-vapour" acceleration that we have discovered [11], which has made it possible to construct sealed-off CO₂ lasers, is the reverse of the effect which nitrogen has on the population of the upper level. The lowest vibrational level of a water molecule lies close to the lower level of the laser transition in CO₂. Because of this, a water molecule is easily able, during a collision, to take over the vibrational energy of a CO₂ molecule which is in the 10⁰0 state. A water molecule vibrationally excited in this way very quickly loses its vibrational energy by collisions with other water molecules; this is due to the very strong dipole interaction. In this process the vibrational energy is converted into translational energy.

It has been found experimentally that the admixture of as little as 0.2 torr H₂O has such a marked effect that the output power is no longer determined by the rate at which the lower level is depopulated. In the optimum case the addition of water vapour can more than double the efficiency [4].

The CO₂ laser with four components

The mechanism of the CO₂ laser as described so far can be regarded as a chain consisting of the following four molecular processes (fig. 4):

- Excitation of nitrogen molecule vibrations by electrons in the gas discharge.
- Transfer of vibrational energy from nitrogen molecules to CO₂ molecules.
- Stimulated emission of electromagnetic radiation at the vibrational-rotational transitions which we have described.
- Transition of the CO₂ molecule to the ground state, by conversion of the vibrational energy of the symmetrical valence vibration into translational energy due to collisions with H₂O molecules.

In a laser of this kind the emission of radiation is determined by the production of molecules with an

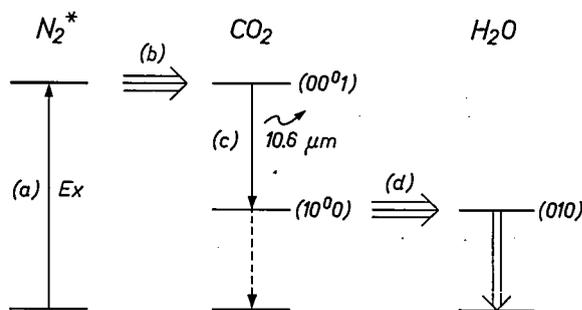


Fig. 4. Diagrammatic representation of the mechanism underlying the operation of a CO₂ laser containing N₂, CO₂ and H₂O. The double and treble arrows denote processes that take place relatively fast and very fast.

[6] G. Moeller and J. Dane Rigden, *Appl. Phys. Letters* **8**, 69, 1966 (No. 3).

[7] Transitions where $\Delta j = 0$ or $\Delta j = -1$ (the Q- and R-branches of the spectrum respectively) do not occur in our experimental conditions.

[8] N. Legay-Sommaire, L. Henry and F. Legay, *C.R. Acad. Sci. Paris* **260**, 3339, 1965.

[9] J. C. Y. Chen, *J. chem. Phys.* **40**, 3513, 1964.

[10] W. J. Witteman, *J. chem. Phys.* **35**, 1, 1961 and **37**, 655, 1962.

[11] W. J. Witteman, *Phys. Letters* **18**, 125, 1965.

excited upper level (process a). We have just seen that the presence of water vapour greatly accelerates the depopulation of the lower level. The emission of radiation is therefore not determined by process (d). The same applies to process (b): it can be calculated that relatively few collisions are necessary on an average for the transfer of vibrational energy from a nitrogen molecule to a CO_2 molecule. (This is partly because there does not have to be an initial conversion into translational energy.)

It had already been found in research on an open flowing-gas laser that the maximum radiation emission can be increased by the addition of helium to the gases mentioned above [12]. This also proved to be the case in our sealed-off lasers. In view of the foregoing considerations, the added helium must have a favourable effect on the rate of process (a). Exactly how this happens, however, is not yet known: it is probably connected with the improvement of the gas discharge conditions. At any rate, we have found that the addition of helium increases the electron temperature, except at very low currents. In our sealed-off lasers we found that the addition of helium had the effect of roughly trebling both the discharge current at which the radiation emission is at a maximum and the maximum radiation emission itself.

Just as there are gases that improve laser action there are others that have the opposite effect. Careful attention must therefore be paid to the purity of the gases used.

Since the generation of vibrationally excited nitrogen, which sets limits to the emission of radiation, is due to the electrons in the discharge, increasing the current should in the first place result in increased emission of radiation. On the other hand, however, the probability of a nitrogen molecule being vibrationally excited by an electron shows a sharp maximum at an electron energy of about 2.5 eV. Now in a $\text{CO}_2\text{-N}_2\text{-H}_2\text{O}$ mixture it has been found from double probe measurements that raising the current has the effect of lowering the electron temperature. If the initial conditions are such that as many electrons as possible have an energy of about 2.5 eV, this means that raising the current will soon lead to a lower radiation emission. If helium is then added to the gas mixture, the electron temperature rises, provided the current is not too small. The maximum emission of radiation is then found at a higher current and is greater than it was before.

The decrease of laser action in the presence of certain impurities is likewise due to the lower electron temperature, the ionization potentials of these impurities being lower than those of N_2 and CO_2 and thus governing to a considerable extent the gas discharge conditions.

Some experimental sealed-off CO_2 lasers

Laser with variable coupling-out

Two factors besides the inversion process which are of importance in obtaining a powerful laser beam

are the method of extracting the beam (coupling-out) and the shape of the region in the resonant cavity in which laser action takes place. These determine the "laser patterns", i.e. the spatial density distribution of the radiation energy in the beam. The shape of the resonant cavity is principally determined by the radius of curvature of the mirrors used and their distance apart. To limit the losses the mirrors should scatter and absorb as little radiation as possible. Most success in the infra-red region has so far been obtained with quartz mirrors coated with an evaporated gold film; for wavelengths of about $10\ \mu\text{m}$ these mirrors have a reflection coefficient of about 98%. Unlike mirrors with evaporated dielectric films — originally only used for lasers operating in the visible range — mirrors with a gold film give no transmission, so that they are not suitable for coupling-out.

Attempts were at first made to solve the coupling-out problem for infra-red lasers by making a hole in one of the gold mirrors and closing it with a material that transmits infra-red radiation. This method has however two serious drawbacks. In the first place a hole in the mirror promotes the formation of laser patterns in which the maximum energy density is found mainly by the hole, so that the part of the radiation that is coupled out is far below the maximum. If the mirrors are given an appropriate radius of curvature it is in fact possible to ensure that part of the radiation is coupled out for all laser patterns [13]. The second drawback is that there are always serious diffraction losses when a coupling hole is used.

Our coupling-out system, used in the laser illustrated in *fig. 5*, does not have these drawbacks [14]. The coupling-out factor can be varied from 0 to more than 50%. As can be seen, the boundaries of the optical resonant cavity do not coincide with the ends of the discharge tube, the left-hand mirror being some distance away from the end. The left-hand end of the tube is terminated by an optically flat plane-parallel germanium plate, the angle between its perpendicular axis and the optical axis of the mirror system being equal to the Brewster angle ($75^\circ 59'$). Since germanium absorbs hardly any infra-red radiation at a wavelength of $10.6\ \mu\text{m}$, all radiation whose plane of polarization is perpendicular to the plane of the drawing therefore passes unhindered through the germanium crystal.

For coupling-out a second germanium plate is used, which is also located inside the resonant cavity but outside the tube. The angle β at which this radiation is incident on the coupling-out plate differs by a few degrees from the Brewster angle, so that partial reflection occurs and part of the beam leaves the resonant cavity. The fraction of the incident radiation reflected from this plate becomes larger the greater the

deviation of β from the Brewster angle. Using a laser like the one shown in fig. 5 we were able to make a quantitative study of the laser mechanism [15].

It is evident that the same fraction is reflected of the radiation returning from mirror M_2 and incident on the other face of the plate. This reflected radiation, however, goes in the opposite direction. By fitting a second plane mirror M_4 it is possible to combine the two coupled-out beams if required.

If the angle β is set to a value very close to the Brewster angle and the difference between the two

curvature of 240 cm. The distance between the two mirrors can be varied; the radiation emission is maximum when the mirrors are 270 cm apart. The discharge tube is 200 cm long and has a diameter of 24 mm.

The partial pressure of the water vapour can be varied from 10^{-3} torr to 1 torr by freezing out part of the water present in a side tube. Fig. 6 shows the maximum output power as a function of the partial pressure of the water vapour for a laser of this type containing 2.5 torr N₂ and 1 torr CO₂. The gain resulting from the addition of water vapour can clearly be seen.

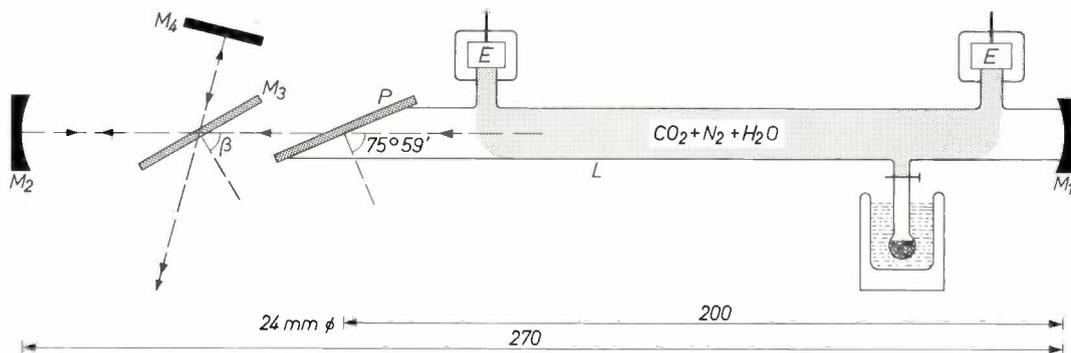


Fig. 5. A sealed-off experimental CO₂ laser with variable coupling-out (diagram). L laser tube. M_1 and M_2 concave mirrors ($R = 240$ cm) which form the ends of the resonant cavity. P plane-parallel germanium plate; the angle between the normal of the plate and the axis of the tube is equal to the Brewster angle. M_3 planar germanium plate, acting as coupling-out mirror; the reflection varies with the angle β . M_4 plane mirror. E cylindrical platinum electrodes in side tubes. Between these electrodes a gas discharge is maintained (shaded). — This tube did not contain any helium; the beneficial effect of helium on the radiation emission had not been discovered when this tube was developed.

angles is then gradually increased, the coupled-out beam increases in intensity. Eventually, however, the intensity decreases again, because once the coupled-out fraction of radiation exceeds a certain limit, the radiation inside the resonant cavity is no longer able to induce transitions in all the excited CO₂ molecules. Part of the excitation energy present in the gas is then released by spontaneous emission and makes no contribution to the laser beam. The laser beam is found to be most intense when β is roughly 68° ; the reflection is then about 10%.

The discharge tube of the laser shown in fig. 5 is filled with 2.5 torr nitrogen, 1 torr CO₂ and roughly 0.2 torr water vapour. The reflectors are very highly reflecting gold-surfaced mirrors each with a radius of

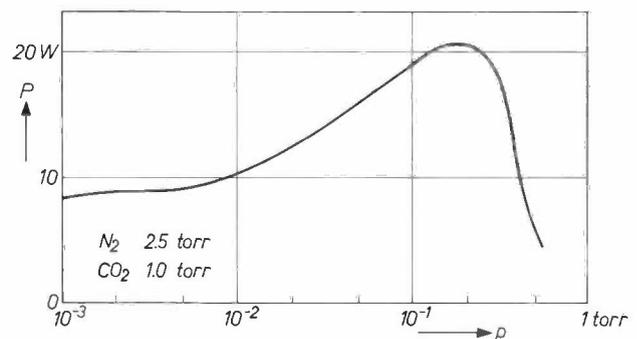


Fig. 6. Maximum output power P of a CO₂ laser of the type shown in fig. 5 as a function of the partial pressure p of the water vapour at the carbon dioxide and nitrogen pressures indicated.

Laser with plane-parallel coupling-out mirror

The variable, loss-free coupling-out system just described, which depends on the fact that very pure germanium absorbs hardly any radiation at a wavelength of 10 μm , operates very well at beam intensities lower than 20 W. At greater intensities, however, absorption becomes troublesome since it increases with rising temperature. The slight initial heating effect caused by an

[12] G. Moeller and J. Dane Rigden, *Appl. Phys. Letters* **7**, 274, 1965.

[13] C. K. N. Patel, *Appl. Phys. Letters* **7**, 15, 1965.

[14] W. J. Wittman and G. van der Goot, *J. appl. Phys.* **37**, 2919, 1966 (No. 7). This coupling-out system has also been used in a laser for plasma research; see: W. J. Wittman, A sealed-off Michelson type CO₂ laser for diagnostic studies of gaseous plasmas, *Appl. Phys. Letters* **10**, 347-349, 1967 (No. 12).

[15] See reference [4].

intense beam gives rise to increased absorption, so that the temperature rises still further, and so on.

An arrangement suitable for lasers of somewhat higher power (about 100 W max.) is illustrated in *fig. 7*. In this laser the tube is sealed off at one end (on the left in the figure) simply by a plane-parallel germanium plate about 2 mm thick ^[16]. This plate is not

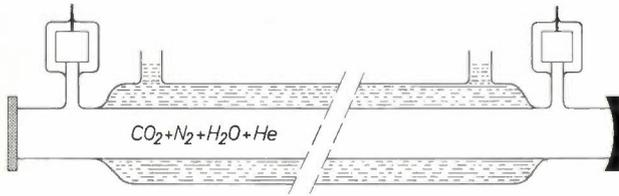


Fig. 7. Semi-confocal laser using a plane-parallel germanium plate (left) as exit window. With a laser of this type, 240 cm long, a continuous radiant output of 103 W has been obtained, with a beam divergence of only 2×10^{-3} radians. There is a water-cooling jacket around the laser tube.

Since the optical resonant cavity of the laser also operates selectively, it quite often happens that radiation of only one wavelength is found in the laser beam. This wavelength is determined by the thickness of the plane-parallel crystal and by the length of the laser. As we have seen in the previous section, operation at a single transition has no adverse effect on the intensity of the beam; all energy stored in the rotational levels ultimately becomes available for stimulated emission.

Coupling-out by means of a plane-parallel germanium plate is not only simple and inexpensive but also results in a parallel laser beam with very little divergence. For a laser as shown in *fig. 7*, with a length of 240 cm, an inside diameter of 22 mm and a gold-coated mirror of a radius of curvature of about 480 cm (a semi-confocal system), a divergence of only 2×10^{-3} radians was measured. The continuous radiant output of this laser, whose photograph is shown in *fig. 8*, is 103 W.

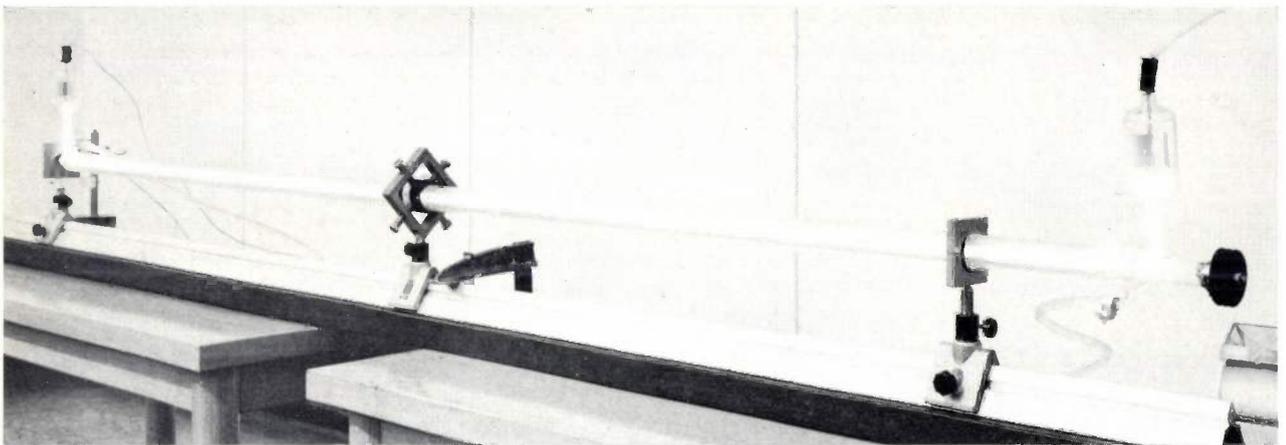


Fig. 8. Operating CO₂ laser of the type shown in *fig. 7*; see also the title photograph. The plane-parallel mirror-exit window of germanium is at the left and the curved gold mirror at the right.

coated with a gold film, and both the front and back faces contribute to the reflection. The reflected beam is therefore the resultant of the radiation reflected from these two faces, which means that the reflection is dependent on the wavelength (or frequency) of the radiation. In the least favourable case the reflection is zero, and in the most favourable 78%. The frequencies at which maximum reflection is found with a plate of the thickness mentioned lie about 7.5×10^{10} Hz apart.

The frequency interval between two laser transitions with successive rotational quantum numbers is 5.5×10^{10} Hz. This means that among the many frequencies occurring in the vibration-rotation spectrum there are only a few for which the germanium plate is highly reflecting and where at the same time the overpopulation of the upper level is considerable (*fig. 3*).

With this laser arrangement, in which the mirrors are cemented straight on to the tube, it is obvious that the two end faces of the tube must be accurately parallel and ground at right angles to the axis. A deviation of about 10^{-3} radians, which can be corrected by bending the laser tube slightly in the middle, is however permissible. The direct cementing of the mirrors to the tube results in a particularly stable assembly.

For lasers with a tube longer than about 2 metres, there is a disadvantage in using a plane coupling-out mirror. The region in which the laser pattern is formed then coincides only partly with the region where the active medium is situated. In other words, the "filling factor" is far from 100%. Part of the active medium cannot therefore contribute to the radiation emission, which means that the maximum possible efficiency of a

long laser cannot be obtained with a planar coupling-out mirror.

Lasers with two concave mirrors

A much higher filling factor can be achieved by using instead of one planar coupling-out mirror, *two* concave mirrors (*fig. 9*). Long lasers with two concave

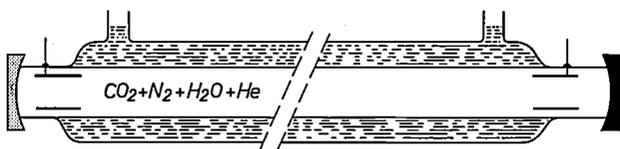


Fig. 9. High-power laser with a zinc selenide exit window, which also acts as a mirror, and with a coaxial electrode arrangement in the tube. The absorption coefficient of ZnSe is low and not very dependent on temperature. The reflectivity of the ZnSe mirror is about 60%. A 300 cm long laser of this type has given a continuous radiant output of 160 W.

mirrors therefore generally have a higher beam intensity and a higher efficiency; for short lasers the difference is negligible.

Clearly a simple concave germanium mirror cannot be used without doing something to improve its reflectivity; there is only one reflecting surface, so that the reflection is too small (no greater than 36%). The reflection has to be improved by depositing dielectric layers on the concave surface.

This is not however the optimum solution in all cases. A difficulty found with lasers that deliver a power higher than 100 W — the very types in which planar mirrors are not entirely satisfactory — is the marked temperature dependence of the infra-red absorption of germanium, as described above. For lasers of this type we have used an exit window of zinc selenide. Although zinc selenide has a higher absorption coefficient than germanium at room temperature, its temperature dependence is weaker.

With a laser as shown in *fig. 9* fitted with a zinc selenide exit window (tube length 300 cm, diameter 2 cm, radius of curvature of mirrors 480 cm) we have obtained an output of 160 W. The reflectivity of the ZnSe mirror in this laser was about 60%. The other mirror was of gold-plated aluminium and had a reflectivity of about 98%.

Further design data and other details relating to the lasers

The *electrodes* for sealed-off CO₂ lasers have to satisfy some very difficult requirements. Hot cathodes, successfully used in lasers filled with inert gases, are ruled out here because the gas mixture in an operating CO₂ laser contains components that are chemically highly reactive (e.g. free nitrogen and oxygen atoms).

It is therefore necessary to use cold cathodes and even then extra care has to be spent on them. We have had particularly good experience with cylindrical electrodes of platinum. These electrodes were at first placed in side tubes (see *figs. 5 and 7*). The disadvantage of this method is that the voltage drop across the side tubes adversely affects the efficiency. Another disadvantage is that the gas discharge is deflected near the side tube openings. This means that the rays experience a change in refractive index when they leave the gas discharge in the neighbourhood of the electrodes, and the resultant deflection of the rays operates as a loss factor.

These drawbacks are overcome if the electrode cylinders are arranged coaxially in the laser tube. We have already seen a laser with this electrode arrangement in *fig. 9*. With a laser of this type, the tube being 150 cm long, the internal diameter 21 mm and the radius of curvature of the mirrors 240 cm, we have obtained a maximum output of 65 W.

The power of the laser beam depends not only on the gas composition, the mirror configuration and the electrodes, but also on the dimensions of the tube. Roughly speaking, the radiated power is proportional to the length of the tube. The radiation emission per unit volume does not therefore change very much if the length of the tube is varied. Closer examination shows that the efficiency increases somewhat more than proportionally with the tube length. The reason for this is that most of the losses become relatively somewhat smaller as the length of the tube is increased.

The diameter is found not to have so much effect on the output. A change in diameter affects some processes favourably and others unfavourably. For example, a large diameter favours strong inversion of the populations of the levels, since a vibrationally excited nitrogen molecule is readily de-excited when it comes into contact with the wall of the laser tube. The result is that the inversion density is lower at the wall of the laser than along the axis of the tube. This effect is relatively less serious in wider tubes. Unfortunately however, the electron temperature, and hence the efficiency of nitrogen excitation, decreases with increasing tube diameter. If the diameter is made too wide, moreover, there is a risk of the gas discharge becoming unstable. We have also found that if the diameter is small the maximum power is reached at greater partial pressures of the gas components. The radiation density consequently increases as the diameter decreases. In practice it has been found that the optimum diameter is between 20 and 22 mm.

The best discharge conditions are found, in our

^[16] W. J. Witteman, *IEEE J. of quantum electronics* QE-2, 375, 1966 (No. 9).

experience, in a d.c. discharge. If the radiant power is varied by varying the current, a higher efficiency is found at lower currents. Fig. 10 shows the radiant power and the corresponding efficiency as functions of the discharge current for a laser tube of the type shown in fig. 9 with the dimensions given above. The theoretical maximum efficiency (i.e. the ratio of the energy of a radiation quantum to that of an excitation quantum) is 41%. Our experiments gave a maximum of 21%, as can be seen from fig. 10. This means that in the most favourable case more than half the power supplied to the discharge is used for the excitation of the upper level. In view of the fact that part of the input power has to be available for maintaining the gas discharge (ionization, etc.) this is an exceptionally good efficiency.

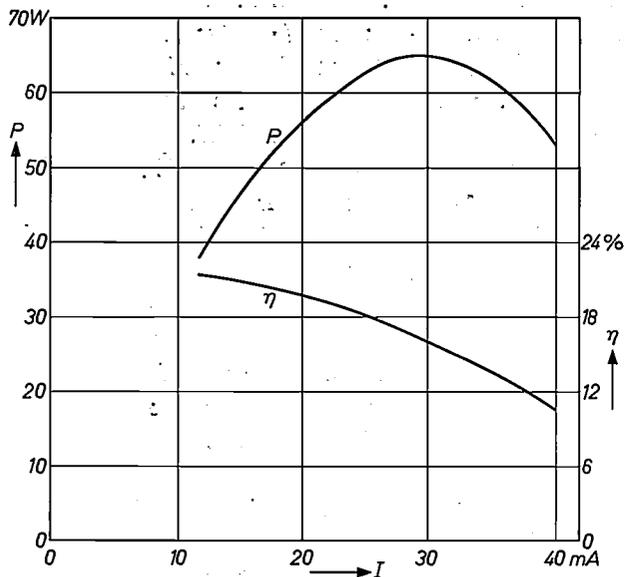


Fig. 10. Output power P and efficiency η as functions of the current I in the gas discharge for a laser 150 cm long, of the type shown in fig. 9.

The excitation mechanism used in our laser is therefore one of rather special interest.

Summarizing the above, it can be said that the most outstanding features of a CO₂ laser are the power per cm³ and the efficiency. In our lasers the efficiency is in fact fairly close to the theoretical maximum. The intensity of the laser beam can be stepped up as required by lengthening the tube. This soon results, however, in instruments that can only be moved with difficulty: a laser beam of 1000 W calls for a tube length of between 15 and 20 metres. For practical reasons we have therefore as yet gone no further than 200 W.

Summary. Infra-red lasers (wavelength 10.6 μm) that operate at a high efficiency and deliver a very intense beam can be produced by making use of the vibrational-rotational transitions of CO₂. The upper level (00⁰1) is excited by means of a gas discharge, and the excitation is greatly intensified by the presence of nitrogen whose molecules can easily be excited into a vibrational state and readily lose their vibrational energy to CO₂. The lower level is the 10⁰ vibrational state. Transitions to 02⁰ are also possible, but are strongly suppressed owing to the ease with which the two states are interchangeable. It has been found that the depopulation of the lower level can be considerably improved by the addition of water vapour; an H₂O molecule readily takes over the vibration energy of the 10⁰ state of CO₂ and relaxes very quickly. When the gas mixture contains water vapour there is no problem in sealing off the laser tube, as water vapour can in addition increase the life of a sealed-off laser to more than 1000 hours. This rules out the necessity for continuous replenishment of the gas. The addition of helium improves the gas discharge conditions. The gas mixture thus has four components: CO₂, N₂, H₂O and He.

Experimental sealed-off CO₂ lasers of various types have been made:

- 1) With variable coupling-out by means of a germanium mirror *outside* the laser tube; particularly useful for research.
- 2) Using a plane-parallel germanium plate as exit window perpendicular to the axis of the tube, and a single concave mirror.
- 3) Using two concave mirrors, the one for coupling-out being of zinc selenide; specially designed for very high outputs.

The beam intensity is roughly proportional to the length of the tube. An output of 160 W has been obtained from a tube 3 m long and containing two concave mirrors. The optimum tube diameter is between 20 and 22 mm. The highest efficiency obtained is 21% (the absolute maximum is 41%).

“Sawing” crystals for gramophone pick-up heads



Rochelle salt (potassium sodium tartrate) gives a strong piezoelectric effect. The effect is put to use in pick-up heads, which convert mechanical vibrations into electrical signals. The element in the pick-up head which carries out the conversion consists of two thin plates of Rochelle salt which are cemented together and fitted with electrodes; this element is mechanically connected to the stylus.

Crystal plates for these elements are made at the Philips Radio, Gramophone and Television Works at Hasselt, Belgium. Rochelle salt crystals weighing about 1.5 kg are grown on a seed crystal by allowing a saturated aqueous solution to cool for about 25 days.

Slices 18 mm thick are then sawn from the crystal, and four of these can be seen in the photograph. Next, the slices are sawn into wafers 3 mm thick. This is not done by sawing in the ordinary sense, but by localized dissolving. The crystal slices are lightly pressed against an array of taut parallel wires which describe a sawing motion, while at the same time they are constantly cleaned by felt strips and wetted with fresh water. This simple procedure prevents the fractures that can result from local temperature differences when mechanical sawing is used. Finally, the wafers are sawn by the same process into pieces of the required dimensions.

Energy paper

There is today a keen and increasing demand for cordless appliances. In principle, two types of current source are available for power supply in these appliances: dry cells, and storage batteries. The choice between these two types is dictated by a variety of factors.

First of all there is the question of the ease of use which is always sought after for cordless appliances. In this respect, the dry cell is the more attractive proposition. "Recharging" is simply a matter of replacement. The storage battery sooner or later requires the use of the mains supply voltage and a charger, requires occasional maintenance, and can only be fully used again after a fairly long charging interval. A draw-back with both the dry cell and the storage battery is that it is not easy to tell at any given moment how much energy is still available.

Technically the storage battery is superior to the dry cell. The power it can deliver for a given volume or weight is many times greater, and the discharge curve is much flatter.

From an economic point of view an important question is whether the relatively high purchase price of the storage battery — not forgetting the cost of the charger — is offset by the advantage of being able to use the storage battery time and time again after recharging.

All these factors are involved in the practical choice between dry cell and storage battery; there is no one decisive criterion. Where the power consumed is extremely low (a battery-operated clock, for example, takes about 0.001 W) a dry cell is the only possible choice. For apparatus requiring higher power (e.g. 12 W for a portable television set) a storage battery is needed. If the power requirements are in the region of 1 W, either dry cells or storage batteries may be used (as, for example, in the "Philishave" types SC 7970 and SC 8020).

Summarizing, we can say that such desirable features as easy recharging, visible energy reserve, high power density, flat discharge curve and a reasonably low price, cannot be combined either in the normal dry cell or in the storage battery.

The type of cell to be described here, which we shall refer to as an "energy-paper" cell, does in fact combine these virtues to quite a considerable extent.

In every chemical cell we find a system consisting of two or more substances that can react with one another. The working principle is that electrons involved in the reaction are not transferred directly from one substance to the other (which would result almost

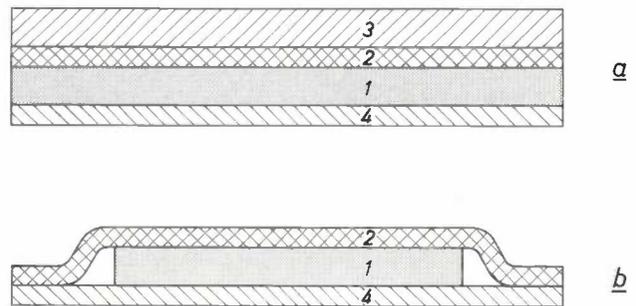


Fig. 1. a) Construction of an energy-paper cell. 1 energy paper proper, consisting of a sheet containing potassium persulphate, powdered carbon and paper fibre. 2 dry sheet of paper containing crystals of common salt. 3 zinc or magnesium sheet. 4 conducting foil.

b) 1, 2 and 4 can be combined to form a "sandwich".

exclusively in the generation of heat) but are forced to "make a detour" around an external circuit and thus to perform useful work in the load.

In the conventional dry cell the zinc can is the electron donor. The electrons travel round the external circuit to the manganese dioxide paste, which acts as the acceptor. The manganese dioxide paste contains carbon to improve the conductivity. Zinc and manganese dioxide are separated by an electrolyte solution, in which the collective ions carry a current of the magnitude of the electron current.

The donor/acceptor/electrolyte combination used in the energy-paper cell consists of zinc or magnesium, potassium persulphate (mixed with powdered carbon) and a solution of common salt, respectively.

The design of the energy-paper cell is however quite different from that of a dry cell, as can be seen in *fig. 1*. The figure shows the following components:

- 1) The energy paper proper, which is a dry sheet consisting of paper fibre impregnated with potassium persulphate and finely-powdered carbon. The paper fibre makes the sheet slightly flexible.
- 2) A dry sheet of paper, e.g. filter paper, impregnated with crystals of common salt, which is moistened before use.
- 3) A sheet of zinc or magnesium.

The components 1 and 2 can be combined with a conducting foil to make an easily manageable sandwich. The zinc or magnesium sheet can be built into the appliance which is to be powered by the energy paper. An experimental version ready for use in an electric shaver is shown in *figs. 2* and *3*. A sandwich whose charge capacity is just sufficient for one shave can be fitted in the battery holder of the shaver (reloading is thus a simple matter). From the stack of sand-

wiches left, the user can see exactly how many shaves are left. The small dimensions of the sandwich (45×45 mm, about 1 mm thick, weight about 2 grammes demonstrate the high power density, when it is remembered that the shaver requires about 1 W. With magnesium as donor the greater part of the charge is supplied at a terminal voltage between 1.9 and 1.5 V (flat discharge curve). The cost of the basic materials of an energy-paper cell, and the capacity per cubic cm, are about the same as for a conventional dry cell.

In this experimental model the magnesium or zinc layer has to be renewed from time to time: magnesium is fairly susceptible to corrosion. In another experimental model this material is fastened to the sandwich in sheet form, and is thus changed every time a new sandwich is used.

Energy paper is moistened only at the moment it is put to use. Premature contact with water will reduce its shelf life: it is more susceptible in this respect than ordinary "dry" batteries. It has however been found that dry energy paper, after being stored air-tight for 8 to 10 months at a temperature of 45°C , had lost only 10% of its capacity.

The high power density, which is higher than that of a normal dry cell by a factor of at least 5, is partly due

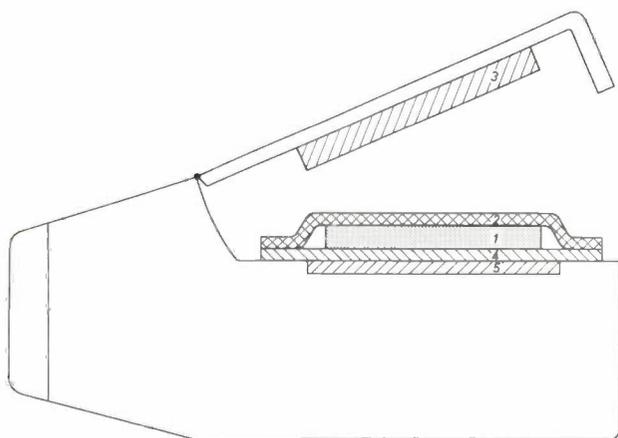


Fig. 2. Diagram of an electric shaver powered by energy paper. The zinc or magnesium sheet 3 is built in here as well as an electrode 5 which connects to the other pole of the battery. For each shave an energy-paper sandwich is inserted in the holder, layer 2 is moistened through the opening at the centre, and then the user only has to snap the case shut to use the shaver for 6 or 7 minutes.

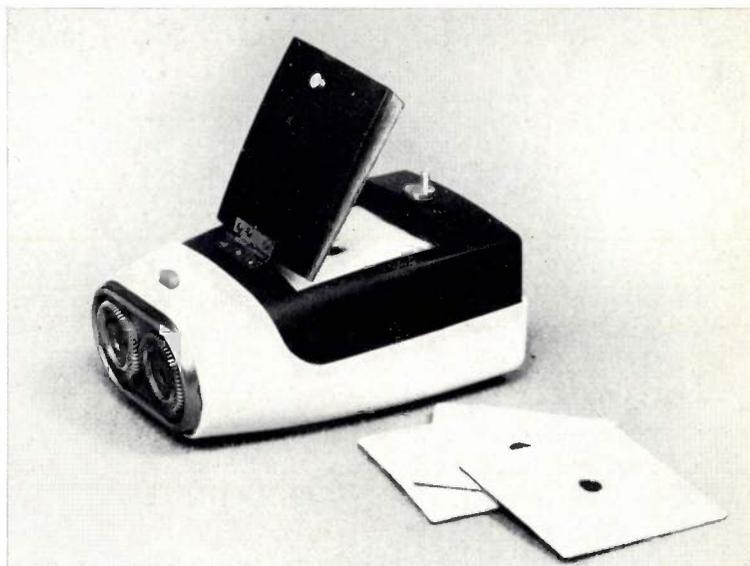


Fig. 3. Experimental version of an electric shaver of the type shown in fig. 2.

to the cell being in the form of a thin sheet, thus minimizing the internal resistance. Another contributory factor is that potassium persulphate gives soluble reaction products, thus preventing isolation of the carbon particles.

The high power density of energy paper makes it generally suitable for applications where relatively high power is needed for a short time. This need may arise, for example, when a car battery is run down. It has been found that the engine can be started again using a suitable quantity of energy paper. The battery does not have to be removed; all that need be done is to connect the energy-paper cell in parallel with the car battery for a few minutes. To start a car under these circumstances it is of course essential that the electrical system of the car as a whole should be in good condition.

Other possible applications are in toys. As a rule these require little energy, and the high power density of energy paper can be exploited here by making a very light power pack. Using small power packs of this kind it is possible to limit the wastage of power inherent in playing with toys, thus reducing running costs.

P. A. Boter
M. D. Wijnen

P. A. Boter and Dr. M. D. Wijnen are with Philips Research Laboratories, Eindhoven.

The skin effect

II. The skin effect at high frequencies

H. B. G. Casimir and J. Ubbink

In Part I of this article [1] the skin effect was considered for various configurations of current-carrying conductors without laying any restrictions on the frequency. It was taken for granted that audio or radio frequencies were involved. However, as the frequency or the conductivity of the metal increases, complications arise in the theory of the skin effect, and these will be the subject of this second part.

To study these complications it is useful to take a simple situation, such as emerges from the following considerations. As the frequency increases, the electromagnetic fields and currents in the metal become concentrated in an increasingly thinner layer under the surface, while outside the metal the finite value of the velocity of light becomes apparent: the wave character of the electromagnetic field emerges more clearly and the wavelength decreases. Now if the skin depth in the material and the wavelength outside are small compared with the smallest radius of curvature of the surface, a typical "optical" situation arises, which is basically that of an electromagnetic wave incident on a surface. It may be remarked that microwaves represent an intermediate case: in microwave devices the wavelength is of the same order of magnitude as the device itself, whereas the skin depth is relatively very small (as it is even at much lower frequencies). The simplest situation is that of a plane electromagnetic wave incident normally on the plane surface of a metal. For a linearly-polarized plane wave in free space, E , H and the direction of propagation are mutually perpendicular. In the following we shall use a co-ordinate system as shown in *fig. 1*, with the z -axis normal to the surface (metal for $z > 0$), E and J along the x -axis, and H along the y -axis. E , J and H then depend only on z (and on t , as $\exp j\omega t$).

The classical skin effect in this situation may be described very simply. Writing down Maxwell's equations (I,1) and (I,2), with $J = \sigma E$, and omitting the displacement current, we find the following relations between the fields in the metal:

$$-\partial H / \partial z = \sigma E, \dots \dots \dots (1)$$

$$\partial E / \partial z = -j\omega\mu H. \dots \dots \dots (2)$$

Prof. Dr. H. B. G. Casimir is a member of the Board of Management of N.V. Philips' Gloeilampenfabrieken; Dr. J. Ubbink is with Philips Research Laboratories, Eindhoven.

The solution is again (cf. I,13):

$$E/E_0 = J/J_0 = H/H_0 = \exp [-(1 + j)z/\delta_k], \dots (3)$$

where the subscript 0 refers to the value at the surface, and where δ_k again represents a length, namely the classical skin depth as defined in (I,10):

$$\delta_k^2 = 2/\omega\mu\sigma. \dots \dots \dots (4)$$

In our treatment of the low frequency case in Part I, had we wished to, we could have pictured the "current-carrying substance" as a continuous inertialess fluid. The complications arising at high frequencies are a direct consequence of the inadequacy of this picture: the *inertia* of the electrons leads to relaxation effects, while the *particle character* of the electrons manifests itself in the anomalous skin effect: the skin depth becomes smaller than the mean free path of the electrons.

The nature of the complications becomes evident if the conduction process is imagined as follows. The electrons move about with randomly oriented velocities of magnitude v . The current is carried by a relatively small directed component. The resistance arises because the electron motion is interrupted by collisions; the

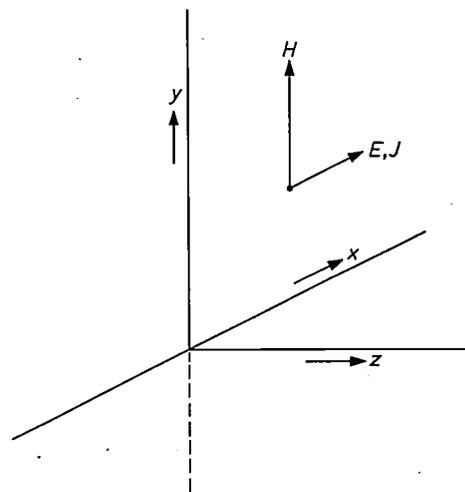


Fig. 1. The co-ordinate system. The (x,y) -plane is the boundary plane between free space (on the left, $z < 0$) and metal (on the right, $z > 0$). An electromagnetic wave is incident on the surface from the left. E and J are directed along the x -axis, H along the y -axis.

time between two collisions is characterized by a relaxation time τ , and the mean free path of the electrons is $l = v\tau$. Let us now compare three lengths: the mean free path l , the skin depth δ and the distance v/ω which an electron travels in $1/2\pi$ of the period of the alternating field. (It may be noted that comparing l with v/ω is equivalent to comparing $\omega\tau$ with 1.)

The various limiting cases that may occur are shown in *fig. 2*. At low frequencies and not too high a conductivity we have case A: $l \ll \delta$ and $l \ll v/\omega$, which

alternating field between two collisions. The collisions are now of little significance to the alternating field, which "sees" a layer of virtually free electrons; these electrons almost fully reflect the incident wave. This is the reflection which is characteristic of many metals in the optical region. The skin depth here is independent both of ω and of τ . We can describe B as a case of *extreme relaxation*, since the current, which in A is in phase with the field, lags more and more behind the field as the frequency increases, owing to the inertia of

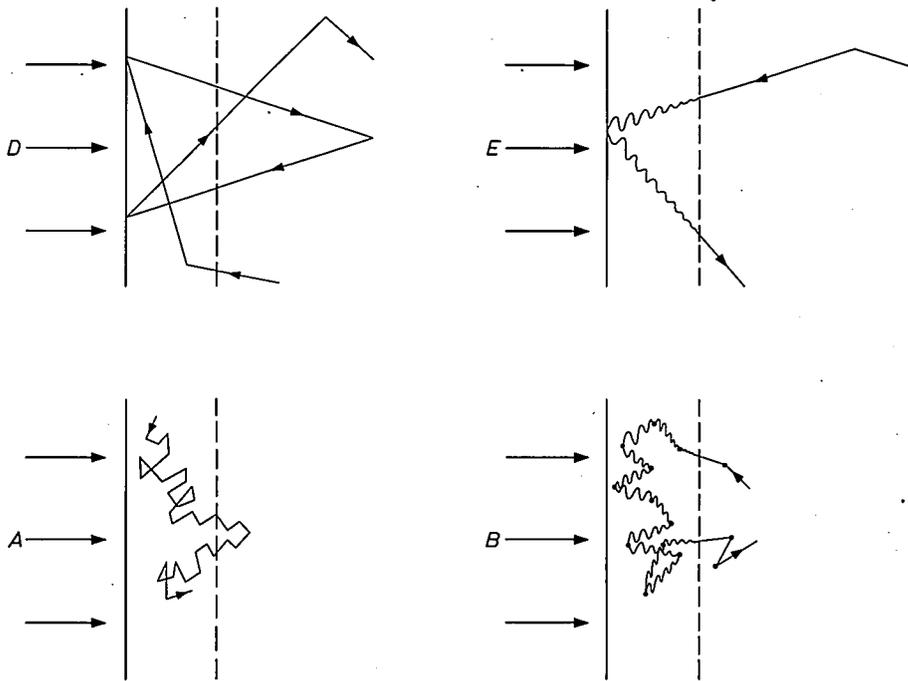


Fig. 2. Diagram of electron paths in a metal in and near the skin layer, in four extreme cases. The dashed line marks the skin layer, whose thickness may of course assume widely different values. The zig-zag line represents a colliding electron, the wavy lines represent the oscillatory motion of the electron due to the alternating field, and the arrows indicate the incident electromagnetic waves.

- A) $l \ll \delta$ and $l \ll v/\omega$ (classical skin effect)
- B) $v/\omega \ll l \ll \delta$ (relaxation)
- D) $\delta \ll l$ and $\delta \ll v/\omega$ (anomalous skin effect)
- E) $v/\omega \ll \delta \ll l$ (anomalous reflection)

l is the mean free path, δ the skin depth and v/ω the distance covered by an electron in $(1/2\pi) \times$ one period of the alternating field.

corresponds to the classical skin effect. The electron experiences many collisions during the time it spends in the skin layer and within one period of the alternating field. If the frequency is increased δ and v/ω become smaller, and if the conductivity is increased, l becomes larger and δ smaller. In both cases the region of the classical skin effect is ultimately left behind; we then come either to case B, where $v/\omega \ll l \ll \delta$, or to case D, where $\delta \ll l$ and $\delta \ll v/\omega$. In B the electrons still undergo many collisions during their time in the skin layer, but there are now many periods of the

the electrons, till in B it lags almost 90° behind the field. In D we have the *anomalous skin effect*: l has become greater than δ . The incident radiation is more strongly absorbed than would be calculated from the theory of the classical skin effect; this effect is to some extent related to the higher resistivity in films which are

[1] H. B. G. Casimir and J. Ubbink, The skin effect, I. Introduction; the current distribution for various configurations, Philips tech. Rev. 28, 271-283, 1967 (No. 9). This article is referred to in the following as I. The equations in I are referred to as (I,1), (I,2), etc.

thinner than the mean free path. In the extreme case events in the skin layer, and hence both the absorption and the reflection, become independent of l and hence independent of the low-frequency conductivity of the metal. If we increase the frequency in case D, or the mean free path in case B, we arrive at situation E, where $v/\omega \ll \delta \ll l$. This case is closely related to B: here again the field "sees" a layer of virtually free electrons, and the reflection is almost 100%. E is referred to as the case of *anomalous reflection*; the residual absorption is a consequence of the collisions at the surface and not, as in B, of the normal collisions within the material.

In these four cases the displacement current can be disregarded. If, starting from B or E, we increase the frequency still further, the displacement current in the Maxwell equation (I,1) ceases to be negligible compared to \mathbf{J} . At the "plasma frequency" \mathbf{D} and \mathbf{J} are equal in magnitude. At still higher frequencies the power not reflected at the surface is not dissipated in a thin layer but is propagated forwards as travelling waves just as light waves in free space: *transmission* takes place. A skin effect no longer exists.

In the following we shall deal in more detail with the effects outlined above. At this point it is useful to refer to figs. 11 and 12 which show the combinations of ω and τ at which the limiting cases mentioned above occur. In these figures C is the transmission region. At the end of this article we shall discuss, with reference to fig. 12, which cases fall into the category of communications engineering and which into optics.

Before examining the effects referred to, we shall attend to two preliminaries. In the first place, in the optical treatment and also in the microwave range, certain concepts are useful for expressing the properties of the metal surface in relation to incident electromagnetic waves. These concepts, such as the complex refractive index and the complex skin depth, will be defined in some detail and we shall deal briefly with their inter-relationship. Secondly, in order to introduce some terms occurring in the electron theory of conduction, we shall present a very concise summary of this theory, confining ourselves to the free electron model.

Skin depth, refractive index and surface impedance

There are various concepts which serve to characterize a metal surface in relation to incident electromagnetic waves. In optics one uses the *optical constants* n and k , to describe the refraction and the absorption respectively of light in the material. These can be combined to form a *complex refractive index* N . At lower frequencies, particularly in the microwave region and in discussions of the anomalous skin effect, use is made of the *complex skin depth* δ . At still lower frequencies

the *surface impedance* Z is a suitable quantity.

The definitions of these quantities and the relationship between them follow most conveniently by considering a plane wave incident normally upon a plane surface and assuming that the electromagnetic fields and currents in the material itself are exponential functions of place and time.

In optics the plane wave in the direction $+z$ is described by

$$E = E_0 \exp(-2\pi kz/\lambda) \exp j\omega(t - nz/c) = E_0 \exp j\omega(t - Nz/c), \dots (5)$$

where

$$N = n - jk. \dots (6)$$

The velocity of electromagnetic waves in the material is c/n , and the amplitude decreases over a distance λ by a factor $\exp 2\pi k$. In (5) $c = 1/\sqrt{\epsilon_0\mu_0}$ is the velocity and λ the wavelength in free space; $c = \omega\lambda/2\pi$.

The *complex skin depth*

$$\delta = \delta' + j\delta'' \dots (7)$$

is now defined such that the behaviour of the fields in the material is given by:

$$E = E_0 \exp j(\omega t - z/\delta), \dots (8)$$

and comparing (5) and (8) we see that:

$$1/\delta = \omega N/c. \dots (9)$$

The concept of *surface impedance* is arrived at in the following way. It is "easy" for currents to exist in metallic material but "difficult" for electric fields. The electric fields at the surface of a metal are weak but the magnetic fields may be strong. The magnetic fields induce currents in the metal, and these (because the conductivity is not infinite) give rise to weak electric fields and hence absorption. Assuming that the currents and fields vanish as z goes to infinity, then for the induced current:

$$\int J dz = H_0, \dots (10)$$

where H_0 is the magnetic field at the surface. This is found by applying Stokes's theorem to (I,1) along a contour that first runs parallel to \mathbf{H} at the surface and then penetrates deep into the metal to where the fields are zero. The displacement current is taken to be negligible. The surface impedance $Z = R + jX$ is now by definition:

$$Z = E_0 / \int J dz = E_0/H_0, \dots (11)$$

where E_0 is the electric field component along the surface and perpendicular to \mathbf{H}_0 . If the displacement current is not negligible, the second expression of (11) is the general definition of the surface impedance.

From (2), (8) and (11) we have at once:

$$Z = \omega\mu\delta. \quad (12)$$

Since only the fields at the surface appear in the definition (11) of the surface impedance, we can use this definition also for situations where the fields in the bulk of the material are *not* exponential. Formally we are at liberty to retain for such situations the definitions (9) and (12) for the complex refractive index and the complex skin depth.

Quantities directly measurable are the *absorptivity* α and the *reflectivity* ρ ; ρ is the reflected fraction of the incident power and α the absorbed fraction. We assume that the material extends to infinity in the direction of the $+z$ -axis, so that there are no complications due to reflection from the other side. We then have $\alpha + \rho = 1$. The term "transmission", already used in the introduction for the effect found above the plasma frequency, means then that the radiation penetrates a considerable distance, i.e. many wavelengths, into the material. At such frequencies a finite thickness of the material transmits a finite fraction of the radiation. The absorptivity α should not be confused with the absorption coefficient, i.e. the fraction of the power dissipated per metre in a travelling wave; this is equal to $2\omega k/c$.

The absorptivity α and the reflectivity ρ may be expressed in terms of δ , Z or N , as follows. Outside the metal, i.e. for $z < 0$, the electromagnetic field is a superposition of the incident plane wave and a reflected wave:

$$\begin{aligned} E &= Z_0 P \exp j\omega(t - z/c) + Z_0 Q \exp j\omega(t + z/c), \\ H &= P \exp j\omega(t - z/c) - Q \exp j\omega(t + z/c), \end{aligned}$$

where P and Q are complex coefficients and $Z_0 = \sqrt{(\mu_0/\epsilon_0)} = 377 \Omega$, the impedance of free space. At the surface $z = 0$ the fields are:

$$\begin{aligned} E_0 &= Z_0(P + Q) \exp j\omega t, \\ H_0 &= (P - Q) \exp j\omega t. \end{aligned}$$

E and H are continuous at the surface. It follows, using (11), that:

$$Z = Z_0 \frac{P + Q}{P - Q},$$

and hence
$$\frac{Q}{P} = \frac{Z - Z_0}{Z + Z_0},$$

so that

$$\rho = \left| \frac{Q}{P} \right|^2 = \left| \frac{Z - Z_0}{Z + Z_0} \right|^2 = \frac{(R - Z_0)^2 + X^2}{(R + Z_0)^2 + X^2}.$$

and
$$\alpha = \frac{4RZ_0}{(R + Z_0)^2 + X^2}. \quad (13)$$

In metals, for a very wide range of values of ω and τ , $|Z| \ll Z_0$, or $c/\omega \gg |\delta|\mu_r$; in other words, provided that μ_r does not differ too much from unity, the free-space wavelength is much greater than the modulus of the complex skin depth, and

$$\alpha \approx 4R/Z_0 = 4(\omega/c)\mu_r\delta'. \quad (14)$$

How far this range extends will be seen later. At optical frequencies we can put $\mu_r = 1$.

If we compare (3) with (8) we see that in the case of the *classical skin effect* (A in fig. 2):

$$(1 + j)/\delta_k = j/\delta,$$

hence

$$\delta = \frac{1}{2}(1 + j)\delta_k, \quad \delta' = \delta'' = \frac{1}{2}\delta_k, \quad (15)$$

where δ_k is the real quantity defined by (4).

The free-electron model

As a background to the matter to be treated, we shall now present a concise summary of the electron theory of metallic conduction, in its simplest form — the free-electron model.

As early as the beginning of this century Drude^[2] put forward a theory which took account of the relaxation effects. Drude's model runs as follows. If v_e is the mean velocity of the electrons, then:

$$m\dot{v}_e = eE - mv_e/\tau; \quad (16)$$

m is the mass and e the charge of an electron. The term on the left-hand side is the inertia term, the first term on the right the external force, and the second term on the right is a frictional term characterized by a relaxation time τ . The relaxation time is the characteristic time needed to reach a static equilibrium, that is to say the time the current takes to reach a steady state after a change in the external field. In the simplest form of electron theory τ is of the order of the average time between two collisions of an electron with the lattice. If n_e is the concentration of the electrons, then the current density $J = n_e e v_e$, from which, using (16), it follows that:

$$\dot{J} = (n_e e^2/m)E - J/\tau. \quad (17)$$

In the d.c. case, i.e. when the field is constant ($\omega = 0$), the left-hand side of the equation is zero, and we have:

$$J = \sigma E,$$

with

$$\sigma = n_e e^2 \tau / m. \quad (18)$$

For sinusoidal fields and currents $\dot{J} = j\omega J$ and we find from (17) and (18):

$$J = \frac{\sigma E}{1 + j\omega\tau}. \quad (19)$$

[2] P. Drude, Ann. Physik 14, 677 and 936, 1904.

By means of this simple model Drude established a relationship between the electrical properties of metals at low frequencies and their optical properties. This will be discussed in the next section.

The result of a quantum-mechanical treatment of the free-electron problem is that the Drude model remains formally applicable with the restriction that we must assign to the electrons a velocity v , the Fermi velocity, to be distinguished from the drift velocity v_e discussed above. The Fermi velocity is a randomly oriented velocity of substantially the same value for all conduction electrons; for an ordinary metal, it is of the order of 10^6 m/s, whereas the drift velocity in an ordinary metal at a current density of, say, 1 A/mm² is of the order of 10^{-4} m/s. The fact that the electrons move in a lattice of positive ions can be taken into account by assigning to them an *effective mass* m instead of the free electron mass. The Fermi velocity is given by n_e and m :

$$(mv/\hbar)^3 = 3\pi^2 n_e. \quad (20)$$

The electrons are subject to two kinds of scattering: by phonons (the quanta of thermal lattice vibrations), giving a relaxation time τ dependent on the temperature; and by lattice imperfections, which means that τ also depends on the purity of the metal. The constancy of the Fermi velocity implies that the relaxation time τ is directly related to a *mean free path* l ,

$$l = v\tau. \quad (21)$$

In Drude's theory n_e , m and e only occur combined in the form nee^2/m . We now introduce the *plasma* (angular) *frequency* ω_p and a corresponding length λ_p as follows:

$$\omega_p^2 = nee^2/\epsilon m, \quad (22)$$

$$\lambda_p^2 = m/\mu nee^2, \quad (23)$$

$$\omega_p \lambda_p = 1/\sqrt{\epsilon \mu}. \quad (24)$$

$2\pi\lambda_p$ is the wavelength corresponding to ω_p for electromagnetic waves propagating in a dielectric material characterized by ϵ and μ (the "medium", see I, Introduction). The quantities ω_p and λ_p will be used for the present only as a measure of the electron concentration. Later they will acquire a more direct physical significance as a characteristic frequency and a characteristic length.

The metal is now characterized by ω_p or λ_p (electron concentration) and τ (purity and temperature). We then have, from (18) and (22), (23), and (4), for example:

$$\sigma = \epsilon \omega_p^2 \tau = \tau/\mu \lambda_p^2, \quad (25)$$

$$\delta_k^2 = 2\lambda_p^2/\omega \tau, \quad (26)$$

and the complex skin depth in the *classical region*

(A in fig. 2) is given by (cf. eq. 15):

$$\delta^2 = j/\omega \mu \sigma = j\lambda_p^2/\omega \tau. \quad (27)$$

The following considerations apply to metallic conductors. If we exclude semimetals and semiconductors we can treat the parameters n_e , v , ω_p and λ_p (and also m in so far as the effective mass differs from the electron mass) as "insensitive" material parameters. They are virtually constants for a given metal, and are about the same from one metal to another. In the free-electron theory the values of v , ω_p and λ_p are determined entirely by n_e . Against these are the "sensitive" parameters: σ , l , τ . These may vary by many factors of 10, depending on temperature and purity. They are proportional to one another, the constant of proportionality being determined by the insensitive parameters.

To give some idea of what may be regarded as typical values for the quantities we have introduced, *Table I* gives values for the insensitive parameters of a "standard metal", calculated from a reasonable choice of n_e , m , ϵ_r and μ_r . For the same metal τ and l have been calculated for three representative values of σ , corresponding to very pure copper at low temperature, copper at room temperature, and constantan.

Non-instantaneous and non-local relations between J and E

We now have the groundwork for making a closer study of the effects mentioned in the introduction. Of these, the relaxation effects and the transmission can be understood on the basis of Drude's theory, as we shall

Table I. Selected parameters defining a "standard metal" and other parameters calculated from them. The sensitive parameters l and τ have been calculated for three representative values of σ . The following constants were used in the calculations: $e = 1.60 \times 10^{-19}$ C, $\hbar = 1.05 \times 10^{-34}$ Js, $\epsilon_0 = 8.85 \times 10^{-12}$ F/m, $\mu_0 = 1.26 \times 10^{-6}$ H/m.

Selected parameters	$n_e = 6.0 \times 10^{28} \text{ m}^{-3}$ $m = 9.1 \times 10^{-31} \text{ kg}$ $\epsilon_r = 1$ $\mu_r = 1$	
Calculated insensitive parameters	$v = 1.40 \times 10^6 \text{ m/s}$ $\omega_p = 1.38 \times 10^{16} \text{ s}^{-1}$ $\lambda_p = 2.18 \times 10^{-8} \text{ m}$ $\tau/\sigma = 6.0 \times 10^{22} \text{ } \Omega\text{ms}$ $l/\sigma = 8.3 \times 10^{-16} \text{ } \Omega\text{m}^2$	
Sensitive parameters	$\sigma \approx 10^{12} (\Omega\text{m})^{-1}$ (pure copper, 4 °K)	$\tau \approx 5 \times 10^{-10} \text{ s}$ $l \approx 10^{-3} \text{ m}$
	$\sigma = 6 \times 10^7 (\Omega\text{m})^{-1}$ (copper, 300 °K)	$\tau = 4 \times 10^{-14} \text{ s}$ $l = 5 \times 10^{-8} \text{ m}$
	$\sigma = 2 \times 10^6 (\Omega\text{m})^{-1}$ (constantan)	$\tau = 1 \times 10^{-15} \text{ s}$ $l = 2 \times 10^{-9} \text{ m}$

see in the next section. On the other hand, the phenomena associated with the anomalous skin effect and anomalous reflection were first investigated only comparatively recently (London 1940 [3], Pippard 1947 [4]).

Between the relaxation effects and the anomalous skin effect there exists, in a certain sense, a kind of symmetry. This is to be seen as follows. In describing the skin effect it is necessary to solve Maxwell's equations for given boundary conditions and for this it is essential to know the relationship between J and E in the material. In the theory of the classical skin effect we simply write $J = \sigma E$ in the metal. This is a *local, instantaneous* relationship: the field at a given place and at a given instant determines the current density at the same place and at the same instant. In Drude's theory the current at a given place at the time t_0 is determined by the values of the field at the same place, but at times of the order of τ or less previous to t_0 ; in other words, there is a *local, non-instantaneous* relation. In the theory of the anomalous skin effect we encounter a *non-local* relation: since the field varies considerably over distances of the order of a mean free path, the velocity of an electron at a given place is determined by fields in an environment of the order of magnitude of l . The symmetry referred to between the relaxation effects and the anomalous skin effect is thus the symmetry between "non-instantaneous" and "non-local".

The relaxation effects in Drude's theory are so much simpler than the non-local theory because it is possible and meaningful to restrict the treatment to a single frequency, whereas in the non-local theory a restriction to a single wavelength in the material is meaningless; moreover, the presence of the surface is a complicating factor in the non-local theory. Because we may restrict ourselves to a single frequency in Drude's theory the time factor may be eliminated and we thus come to the simple local relationship between J and E given by (19). The non-instantaneous origin of (19) is still reflected in the imaginary term in the denominator.

It should be noted, incidentally, that the explicit occurrence of the mean free path l in the anomalous (non-local) theory — in addition to the parameter τ already encountered in Drude's theory — means that the Fermi velocity v , a typical quantum-mechanical quantity, will now be of significance.

Relaxation effects and transmission

We shall now consider the modifications to be made to the theory of the skin effect when Drude's theory is applied. We now have the relaxation time τ as a new

parameter along with σ , the conductivity at zero frequency; τ and σ are related by the expression (18) or (25). We shall still keep to the simple configuration given in fig. 1. Two modifications now arise in eq. (1). In the first place we take the displacement current into account; in the second place we do not substitute σE for J but $\sigma E / (1 + j\omega\tau)$ as given by (19). If the fields and currents have an exponential time dependence as given in (8), Maxwell's equations may then be written:

$$jH/\delta = j\omega\epsilon E + \frac{\sigma E}{1 + j\omega\tau} \quad (28)$$

$$-jE/\delta = -j\omega\mu H \quad (29)$$

Multiplying these equations we obtain:

$$1/\delta^2 = \epsilon\mu\omega^2 - \frac{j\omega\mu\sigma}{1 + j\omega\tau} \quad (30)$$

Together with (9), this expression establishes a connection between the complex refractive index $N = n - jk$ on the one hand and σ and τ on the other. For our further discussion we write (30) with the aid of (22) to (25) in the following form:

$$(\lambda_p/\delta)^2 = (\omega/\omega_p)^2 - 1 + \frac{1}{1 + j\omega\tau} \quad (31)$$

giving δ as a function of ω and τ . The ω - τ -plane (fig. 3), in which ω is plotted horizontally and τ vertically, both on a logarithmic scale, can now be divided into three

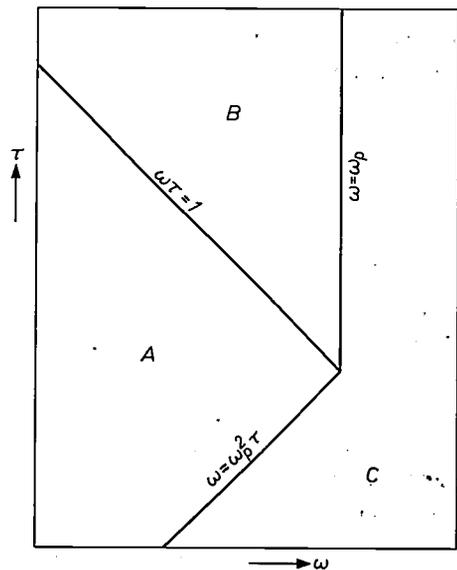


Fig. 3. The ω - τ -diagram, in which τ is plotted vertically and ω horizontally, both on a logarithmic scale. With Drude's theory three regions A, B and C can be distinguished in which, well away from the boundaries, the behaviour of the electromagnetic waves is very simple. A is the region of the classical skin effect (see fig. 2), B the relaxation region (see fig. 2) and C the transmission region. In A, δ is given approximately by (32), in B by (34) and (35), and in C by (36). The equations for the boundaries are indicated in the figure.

[3] H. London, Proc. Roy. Soc. A 176, 522, 1940.
 [4] A. B. Pippard, Proc. Roy. Soc. A 191, 385 and 399, 1947.

regions. In each of these regions (well away from the boundaries) the behaviour of δ is particularly simple.

A) $\omega\tau$ is so small that the second and third terms on the right-hand side of (31) can be combined to $-j\omega\tau$, and the first term is much smaller than this new term if $(\omega/\omega_p)^2 \ll \omega\tau \ll 1$, i.e. $\omega\tau \ll 1$ and $\omega \ll \omega_p^2\tau$. We then have approximately $(\lambda_p/\delta)^2 = -j\omega\tau$, or:

$$\delta^2 = j\lambda_p^2/\omega\tau. \quad \dots \dots (32)$$

This is the expression for the classical skin effect (27). The region A is therefore the region of validity for the *classical skin effect*. From (9), (6) and (25) it follows that (27) is equivalent to:

$$n = k = \sqrt{(\sigma/2\varepsilon\omega)}. \quad \dots \dots (33)$$

This is the *Hagen and Rubens relation*^[5], the classical skin effect expressed in optical terms.

B) The middle right-hand term of (31) is the most significant when $\omega\tau \gg 1$ and $\omega \ll \omega_p$. Then approximately $\delta^2 = -\lambda_p^2$, or $\delta' = 0$ and $\delta'' = \lambda_p$. The electromagnetic field is now an oscillation without wave character, which decays along the z -axis as $\exp(-z/\lambda_p)$ (sometimes called an "evanescent wave"). The penetration depth is independent of frequency and of d.c. conductivity (i.e. of relaxation time). Between two collisions the electrons undergo many cycles of the alternating field ($\tau \gg 1/\omega$) so that what the electromagnetic field primarily "sees" is a layer of entirely free electrons which screen but do not absorb because the mean velocity of the electrons, and hence the current, is always 90° out of phase with the field. Given $\omega\tau \gg 1$ this 90° phase lag follows immediately from (19). B is the *relaxation region*. The incident wave is fully reflected. The absorption, however, is zero, and the reflection complete, only to the zero order approximation. The first order approximation is found by also taking into account terms of the first order in

$1/\omega\tau$ in (31). In that case we find, for $\omega \ll \omega_p$:

$$\delta'' = \lambda_p, \quad \dots \dots (34)$$

$$\delta' = \lambda_p/2\omega\tau. \quad \dots \dots (35)$$

C) Only the first right-hand term in (31) is significant. This is the case if $\omega \gg \omega_p$ and also if $\omega \gg \omega_p^2\tau$. Then approximately $1/\delta^2 = \varepsilon\mu\omega^2$, and hence

$$\delta' = 1/\omega\sqrt{(\varepsilon\mu)}, \quad \delta'' = 0. \quad \dots \dots (36)$$

The solutions (8) now represent undamped travelling waves of phase velocity $1/\sqrt{(\varepsilon\mu)}$ and wavelength $2\pi\delta'$. The fields are therefore in principle no longer limited to a thin layer at the surface, and *transmission* occurs. We are then really outside the region with which we are concerned: a skin effect no longer occurs. The parameters that define the behaviour of the electrons have vanished. The electrons have no longer any effect on the electromagnetic field (again, of course, only in the zero order approximation).

In a previous section we gave an approximate expression (equation 14) for the relation between the complex skin depth and the absorptivity; we can now state the region in which this expression is valid. The condition for the validity of (14) was $c/\omega \gg |\delta|$, or $\omega^2/c^2 \ll 1/|\delta|^2$ (putting $\mu_r = 1$). This means that the first term on the right-hand side of (30), and thus the first term on the right-hand side of (31), must be small compared with the other terms; in other words, coming from A or B we must not approach the boundary with C too closely.

The relation between these limiting cases and the transition from the one wave behaviour to the other is found by constructing the curve representing δ in the complex δ -plane, as ω goes from zero to infinity (*fig. 4*); and further by considering the character of the wave field corresponding to various points in the complex δ -plane (*fig. 5*). *Fig. 4a* shows the locus of δ in the

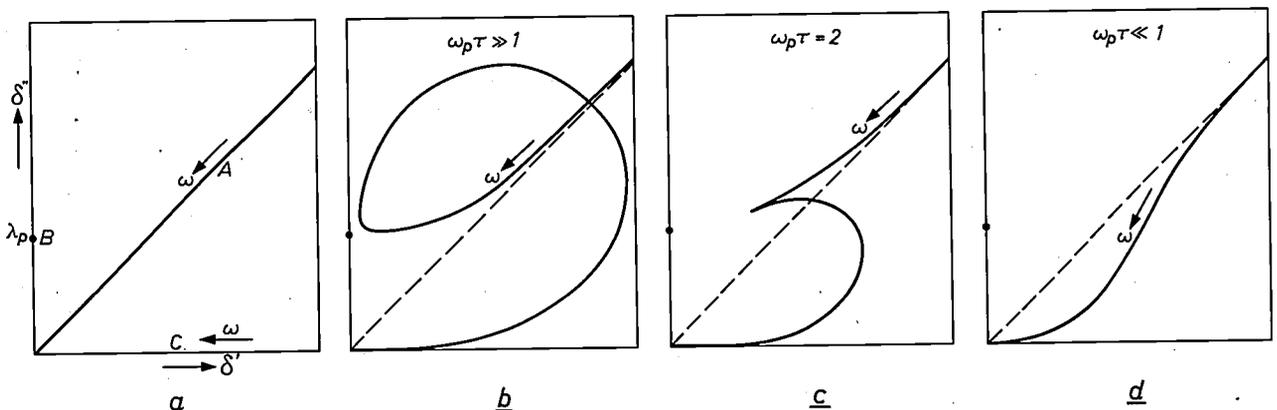


Fig. 4. The path followed by δ in the complex plane, according to Drude's theory, as ω goes from $0 \rightarrow \infty$, a) in the three limiting cases A, B and C of *fig. 3*, b) for $\omega_p\tau \gg 1$, c) for $\omega_p\tau = 2$, and d) for $\omega_p\tau \ll 1$. The arrow indicates the direction of increasing ω .

limits A, B and C as ω varies (increasing in the direction of the arrow). If we examine qualitatively how δ varies in accordance with the complete expression (31), we arrive at fig. 4b for $\omega_p\tau \gg 1$, fig. 4c for $\omega_p\tau = 2$ and fig. 4d for $\omega_p\tau \ll 1$. For $\omega_p\tau \gg 1$, δ describes a loop

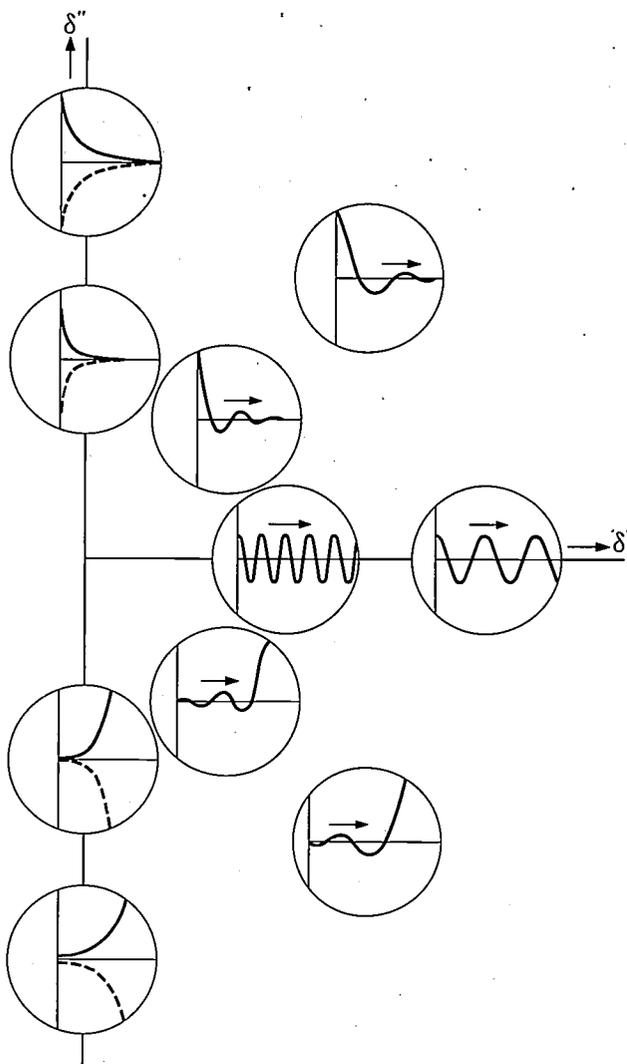


Fig. 5. The wave character of the field in the metal appropriate to various regions of the complex δ -plane. For waves travelling to the right (i.e. along the positive z -axis) δ lies in the first or the fourth quadrant ($\delta' > 0$). We have attenuated waves in the first quadrant, waves of increasing amplitude in the fourth quadrant, unattenuated waves on the δ' -axis ($\delta'' = 0$), and on the δ'' -axis ($\delta' = 0$) non-propagating oscillations having no wave character ("evanescent waves"), either decaying ($\delta'' > 0$) or augmenting ($\delta'' < 0$) along the z -axis.

between B and C; this loop disappears for small $\omega_p\tau$. The intermediate case with a cusp occurs for $\omega_p\tau = 2$ (this cusp has no obvious physical significance). For very large $\omega_p\tau$ the loop is traversed in a very short ω -interval.

Fig. 5 shows the wave fields corresponding to certain points in the δ -plane. These fields follow from the space

and time dependence of the field as given in (8):

$$\exp j(\omega t - z/\delta) = \exp(-\delta''z/|\delta|^2) \exp j(\omega t - \delta'z/|\delta|^2).$$

We shall confine ourselves to waves travelling to the right (or to the evanescent case), $\delta' \geq 0$, i.e. to the first and fourth quadrants of the complex δ -plane. In the first quadrant we have attenuated waves ($\delta'' > 0$), in the fourth quadrant waves of increasing amplitude ($\delta'' < 0$), along the real axis unattenuated waves and along the imaginary axis the evanescent case (non-propagating oscillations) either decaying exponentially ($\delta'' > 0$) or augmenting exponentially ($\delta'' < 0$) with z .

In these figures there are two further points to note. a) In fig. 4 we see that δ is always in the first quadrant; we therefore always have attenuated travelling waves, and energy is always dissipated in the material. In the limiting cases B and C the dissipation is zero: in C the waves are unattenuated, and in B no waves are propagated and there is therefore no energy transport. b) In the "plasma transition" B \rightarrow C at large values of $\omega_p\tau$ (fig. 4b) δ first goes off along the imaginary axis ($\delta' \approx 0$, δ'' increases) and later returns along the real axis ($\delta'' \approx 0$, δ' decreases). At first, therefore, the field penetrates further and further, while the fields at different points have virtually the same phase; later phase differences occur over large distances, i.e. we have long travelling waves, which become shorter and shorter and are virtually not attenuated.

To what extent does Drude's theory cover the experimental facts? In general terms we may say the following.

For our standard metal the "plasma boundary" $\omega = \omega_p$ is at a wavelength of $2\pi\lambda_p = 2\pi \times 2.18 \times 10^{-8} \text{ m} = 0.137 \mu\text{m}$, i.e. in the far ultra-violet (see fig. 6). The value of τ at the point where the three regions meet is $1/\omega_p = 0.725 \times 10^{-16} \text{ s}$; this is a factor of 500 lower than the τ for copper at room temperature: $\tau_k = 3.6 \times 10^{-14} \text{ s}$. The wavelength at which copper, at room temperature, passes the boundary $\omega\tau = 1$ is $2\pi c\tau_k = 68 \mu\text{m}$.

Confining ourselves at first to "optical" frequencies (the visible range and the near infra-red and ultra-violet) we find from these rough estimates that only metals which are poor conductors will be in region A; for metals which are good conductors optical frequencies lie in region B.

Constantan is an example of a *poor conductor*, where the Hagen-Rubens relation, which is appropriate to region A, is still reasonably valid. For the *good conductors* (Cu, Ag, Au, Al) the expressions (34) and (35) for region B give a good *qualitative* description of the experimental behaviour of the optical constants as

[5] E. Hagen and H. Rubens, Ann. Physik 11, 873, 1903.

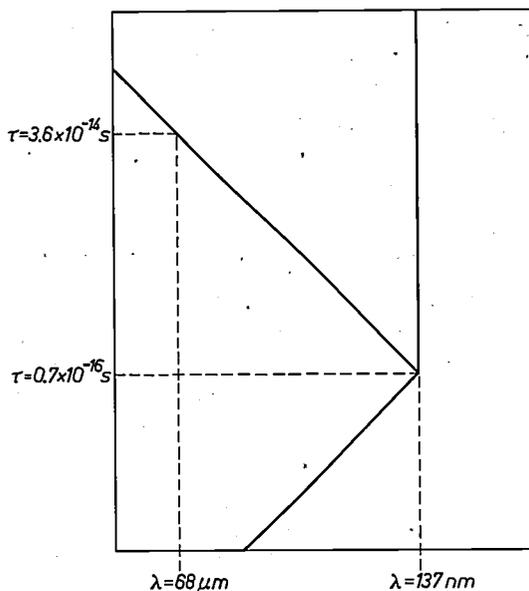


Fig. 6. Some characteristic values of τ and λ in the ω - τ -diagram (see fig. 3) for the standard metal. At the point where the three regions meet ($\omega = \omega_p$, $\tau = 1/\omega_p$), $\tau = 0.7 \times 10^{-16}$ s, and $\lambda = 137$ nm. At $\tau = 3.6 \times 10^{-14}$ s (corresponding to the value of σ for copper at 300 °K) the boundary line $\omega\tau = 1$ corresponds to a wavelength $\lambda = 68$ μ m.

functions of ω , provided ω is not too large [6]. To obtain numerical agreement however it is necessary to insert a value for τ which is appreciably smaller than the value corresponding to σ , the d.c. conductivity (see Table II). The results of measurements in the optical region would thus be in agreement with the theory if these metals were worse conductors than they really are. A possible explanation for this is that the concentration of impurities in the thin layer under the surface, in which the optical effects take place, is greater than in the metal as a whole. This also explains the disparities in the results of various workers and explains why the results depend to such a marked extent on the method of preparing the surface [6].

The values of ω at which there ceases to be qualitative agreement are in general much lower than the

Table II. Comparison of optical and low-frequency electrical properties of copper, silver and gold, after Försterling and Freedericksz [7]. $\tau(\text{opt})$ is the relaxation time necessary to describe the results of the optical measurements by Drude's theory; $\tau(\sigma)$ is the relaxation time derived from σ . We have converted the σ -values given by the authors to τ by means of eq. (25). The factor $\mu\lambda_p^2$ which occurs in this conversion was derived for the free-electron model from the density and the atomic weight, assuming one electron per atom. τ is expressed in units of 10^{-14} s and $\mu\lambda_p^2$ in units of 10^{-22} Hm.

	Cu	Ag	Au
$\tau(\text{opt})$	0.46	0.95	1.7
$\tau(\sigma)$	2.5	3.9	2.6
$\mu\lambda_p^2$	4.2	6.1	6.1

expected plasma frequency. At these frequencies absorption bands are found which are attributed to quantum jumps of electrons induced by the radiation. These are not accounted for in Drude's theory.

In metals such as gold, copper and silver the plasma boundary is thus obscured by these absorption bands. This does not apply to the alkali metals. In 1933 Wood [8] discovered that these become transparent in the ultra-violet. Zener [9] interpreted the wavelength at which this occurs as the plasma boundary. The theoretical and experimental transition wavelengths are compared in Table III.

Table III. The plasma boundary, i.e. the transition wavelength at which a metal ceases to reflect and becomes transparent: experimental values, after Wood [8] (λ_{exp}) and theoretical values ($2\pi\lambda_p$) after Zener [9]. n_{ext}/n_a is the number of electrons per atom at which the theoretical plasma wavelength would coincide with the experimental value.

	λ_{exp}	$2\pi\lambda_p$	n_{ext}/n_a
Li	205 nm	155 nm	0.54
Na	210	210	1.00
K	315	290	0.85
Rb	360	320	0.79
Cs	440	360	0.67

In the low-frequency and radio-frequency regions the classical skin effect theory provides in general a good description of the phenomena. In his measurements of the high-frequency resistance of superconducting tin in 1940 H. London [3] found, however, as an incidental result, that at low temperature the apparent resistance of the normally conducting metal was greater at microwave frequencies than the d.c. resistance. He then put forward a suggestion that this was due to the fact that the skin depth becomes smaller than the mean free path of the electrons, and in so doing laid the foundation for the theory of the anomalous skin effect.

The anomalous skin effect

In calculating the classical skin effect we implicitly assumed that we could define the current density \mathbf{J} and the field strength \mathbf{E} at any point in the metal, and that these quantities were always connected by the relationship $\mathbf{J} = \sigma\mathbf{E}$. Now if the skin depth — a distance over which the fields vary considerably — becomes smaller than the mean free path (see D in fig. 2) then the velocity of the electron (and hence its contribution to the current density) is determined not only by the local field but also by all the fields to which the electron has been subjected since its last collision.

A simple picture of the situation — which can also be used to obtain a reasonable estimate of the effect — has been given by Pippard [4] with his "ineffectiveness con-

cept". An electron moving approximately at right angles to the surface will be subject to the field of the skin layer only for a small part of its transit time between two collisions. Its interaction with the field is therefore very slight and it is therefore not effective in the screening and absorption of electromagnetic waves. Screening and absorption, and therefore the formation of a skin layer, are mainly due to the few electrons which graze the surface at small angles, since these spend a large part of their free transit time in the skin layer. The fraction of the electrons in the skin layer that are effective depends on the skin depth whose value, however, calculated in the classical way with a reduced number of effective electrons, is in turn dependent on this fraction. If δ is the skin depth, then the ratio of the number of effective electrons (the electrons whose path between two collisions is almost entirely inside the skin layer) to the total number in the skin layer is roughly δ/l , where l is the mean free path of the electrons. In the skin layer there is therefore an effective conductivity $\sigma_{eff} = b\omega\delta/l$, where b is a number approximately equal to 1, or somewhat greater because the non-grazing electrons do still make a slight contribution. With this effective conductivity we again calculate the skin depth, using equation (4), and find:

$$\delta^2 = 2/\omega\mu\sigma_{eff} = (2/\omega\mu\sigma)l/b\delta = \delta_k^2/lb\delta,$$

hence

$$\delta^3 = (1/b)\delta_k^2 l. \quad \dots \quad (37)$$

This expression is instructive inasmuch as it shows that, apart from a factor $1/b^{1/3}$, the skin depth is the geometric mean of the classical skin depth δ_k (doubly weighted) and the mean free path l . But at the same time it is perhaps misleading, as it suggests that δ depends on the sensitive parameter l . That this is *not* the case follows if we substitute for δ_k^2 the expression (26); with $l = v\tau$ we obtain:

$$\delta^3 = (2/b)(v/\omega)\lambda_p^2. \quad \dots \quad (38)$$

We can also arrive at this result by a line of thought which we shall use again at a later stage. This is based on the introduction of an "effective relaxation time" τ_{eff} . This is the time during which the field and the electron are in *uninterrupted* interaction with one another. It determines the effect of the electron on the field (screening, absorption). In the classical case the field remains the same between two collisions, so that

τ_{eff} is equal to the time τ between two collisions. In the case we are considering, however, the period of uninterrupted interaction is marked by the electron entering or leaving the skin layer and a collision with the surface: τ_{eff} is now approximately equal to the time needed to pass through the skin layer. If we put $\tau_{eff} = b\delta/v$, and, by analogy with (26), $\delta^2 = 2\lambda_p^2/\omega\tau_{eff}$, we arrive immediately at (38). The reason for the fact that δ is independent of the sensitive parameters might therefore be summarized as follows: the thickness of the skin layer is determined by the effective interaction between the electrons and the field, and this interaction, provided $l \gg \delta$, is in turn determined by the thickness of the skin layer; the sensitive parameters have thus been completely eliminated.

A rigorous theory of the anomalous skin effect has been given by Reuter and Sondheimer [10], and a somewhat less rigorous theory by Chambers and Pippard [11]. Where a comparison is possible, both treatments give the same result. In the theory three steps can be distinguished.

1) The main problem is to find the current density occurring at a point in the material as a result of the field, not only at the point itself but also within a radius of the order of magnitude of l . Chambers and Pippard postulate that an electron "remembers" the fields that it has traversed in the past, but that its memory fades in the course of time as $\exp(-t/\tau)$. A calculation along these lines leads to the following non-local relation between \mathbf{J} and \mathbf{E} (see fig. 7):

$$\mathbf{J} = \frac{3\sigma}{4\pi l} \int \frac{\mathbf{r}(\mathbf{E} \cdot \mathbf{r}) \exp(-r/l)}{r^4} dV. \quad \dots \quad (39)$$

The current density \mathbf{J} is a volume integral; \mathbf{r} is the radius vector between a volume element dV where a field \mathbf{E} prevails and the point where \mathbf{J} is calculated.

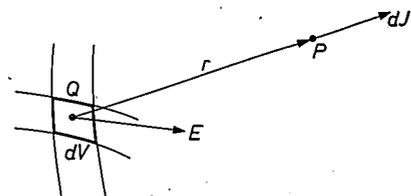


Fig. 7. Referring to eq. (39) giving the current density at P . A contribution $d\mathbf{J}$ comes from electrons from the element dV at Q that reach P without colliding, and which have been subject to the field \mathbf{E} in Q ; \mathbf{r} is the radius vector from Q to P .

The result (39) can be obtained as follows. The field in the direct environment of Q (fig. 8a) affects the current in P because P is traversed by a number of electrons which have come from Q without colliding on the way. We note in the first place that a field in Q perpendicular to \mathbf{r} has no effect on the current density in P : electron paths leading from Q to the environment of P are twisted slightly by such a field; in the presence of such a field

[6] M. Parker Givens, *Solid State Physics* 6, 313, 1958.
 [7] K. Försterling and V. Freedericksz, *Ann. Physik* 40, 200, 1913.
 [8] R. W. Wood, *Phys. Rev.* 44, 353, 1933.
 [9] C. Zener, *Nature* 132, 968, 1933.
 [10] G. E. H. Reuter and E. H. Sondheimer, *Proc. Roy. Soc. A* 195, 336, 1948.
 [11] A. B. Pippard, *Advances in Electronics and Electron Physics* 6, 1, 1954; *Reports on Progress in Physics* 23, 194, 218, 1960.

the electrons that pass through P are not the same as those when there is no field, but their number and mean velocity is identical in both cases (fig. 8b, c). Let us now consider all the paths in a solid angle $d\Omega$ of the electrons that pass P from the Q -direction in a given short time interval; their number is proportional to $d\Omega$ (fig. 8d). If we follow these paths backwards in time, that is to say, in the direction of Q , then every now and then one will, as it were, be "knocked out of the cone". The chance of this happening in an element dr is equal for each path and proportional to dr . The number of "undisturbed" paths thus decreases exponentially in the direction of Q and the number of relevant

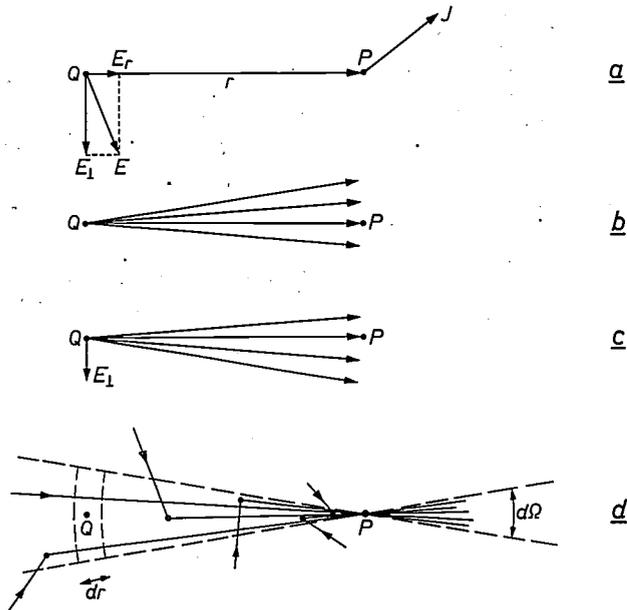


Fig. 8. Calculation of the effect of the field at Q on the current in P via electrons originating directly from Q .

paths in the element dr around Q is proportional to $\exp(-r/l)dr$. If we now assume that the average contribution to the current intensity in P due to the field in Q is proportional to the parallel component of the field in Q (in doing so allowance is made for the collisions), then the contribution dJ to the current density in P is proportional to $(r/r)E_r \exp(-r/l)dr d\Omega$, where E_r is the field component in Q in the direction of r . The factor r/r indicates that dJ has the direction of r , since E_{\perp} has no effect. Substitution of $E_r = (\mathbf{E} \cdot \mathbf{r})/r$ and $dr d\Omega = dV/r^2$ gives the expression (39), apart from the proportionality factor. The integration is easily carried out for a homogeneous field \mathbf{E} ; if in this case we put $\mathbf{J} = \sigma\mathbf{E}$, we find immediately the value of the factor of proportionality.

In (39) the relaxation effects are neglected. These can be taken into account by substituting for \mathbf{E} the "retarded field" $[\mathbf{E}]$, i.e. the field in dV , at the instant $t_0 - r/v$ at which the electrons from dV contributing to \mathbf{J} at t_0 passed through dV .

2). Next, the effect of the surface has to be taken into account. It is necessary to make an assumption about the manner in which the electron is reflected at the surface; the cases which have been calculated completely are those of *specular reflection* and *diffuse reflection*. In the latter case the "memory" of the electrons is

completely wiped out upon reflection. The integral in (39) then has to be taken only over the volume of the metal. For specular reflection the calculation proceeds as if the electrons were able to pass through the surface undisturbed, and as if outside the metal the electric field were a mirror image of that inside the metal; the integral is then taken over the entire volume. In its effect on a reflected electron prior to the reflection, the field in the metal is simulated by the image field.

3) Substitution of the integral expression found for \mathbf{J} in Maxwell's equations (omitting the displacement current) yields, after elimination of \mathbf{H} , an equation for \mathbf{E} which has been solved by the above-mentioned authors [10] [11], for the cases where all the electrons undergo either specular or diffuse reflection at the surface.

The result (38), based on simple considerations, agrees surprisingly well with the rigorous theory. According to the latter, equation (38) — after the correct value for b has been inserted — is a valid expression for the complex skin depth in the anomalous limit without relaxation ($\delta \ll l$, $\delta \ll v/\omega$, D in fig. 2), except for one detail. The considerations that lead to (38) are too oversimplified for a proper analysis of the complex character of δ . It is found, in fact, that the rigorous theory simply adds a minus sign to equation (38). In other words, in the *anomalous limit* (without relaxation) the complex skin depth is given by:

$$\delta^3 = -(2/b)(v/\omega)\lambda_p^2. \dots (40)$$

δ must lie in the first quadrant of the complex δ -plane (passive material, waves travelling to the right). The minus sign then implies that [12]

$$\delta'' = \delta'/3. \dots (41)$$

The value of b depends on the manner in which the electrons are reflected at the surface. For diffuse reflection $b = 4\pi/3 \approx 7.3$; for specular reflection $b = 3^{5/2}\pi/3 \approx 10.3$. Here, therefore, the character of the reflection has only a minor effect. As we shall see later, this is no longer the case if $v/\omega \ll \delta$ (\mathbf{E} in fig. 2).

The rigorous theory, of course, provides more information. In the first place the complex skin depth is also calculated in the transition region where l is not much larger or smaller than δ_k . Secondly, the field configuration in the metal is calculated completely. The fields are no longer exponential functions of place. For a given relative decrease at the surface the fields extend much deeper into the metal than would be the case with an exponential function. (Even so, the complex skin depth is still a useful concept for describing the surface in relation to incident waves; see the remark following equation 12.)

In order to find the boundary in the ω - τ -plane (cf.

fig. 3) between the regions of validity of the classical skin effect (A) and of the anomalous skin effect (D) we put $l = |\delta|$. Calculating δ in the classical limit as given by (27), we then find:

$$\omega\tau^3 = (\lambda_p/v)^2. \quad (42)$$

This new boundary is shown in fig. 9 together with the boundaries indicated in fig. 3.

For our standard metal (Table I) at the new boundary we have $\omega\tau^3 = 2.4 \times 10^{-28} \text{ s}^2$, which means that for the values $\tau = 5 \times 10^{-10} \text{ s}$ and $\tau = 4 \times 10^{-14} \text{ s}$ from

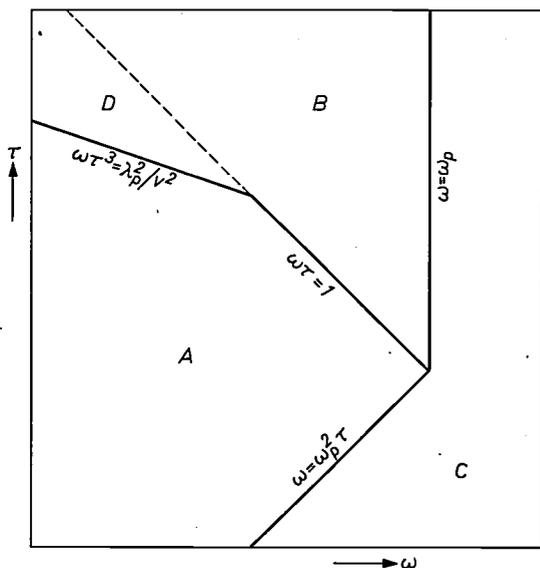


Fig. 9. The ω - τ diagram as in fig. 3, showing the new boundary between the classical skin-effect region (A) and the region of the extreme anomalous skin effect (D). In the latter region the skin depth is given approximately by eq. (40).

Table I the boundary is reached at the values $\omega \approx 2 \text{ s}^{-1}$ and $\omega \approx 4 \times 10^{12} \text{ s}^{-1}$ respectively. We must therefore expect that the "best" metal, e.g. pure copper at 4 °K, will have already become anomalous at very low frequencies and that at millimetre waves ($\omega \approx 10^{12} \text{ s}^{-1}$) even a very "moderate" metal, like copper at 300 °K, will show anomalous behaviour. At the point where the new boundary runs into the old one, $\omega\tau = 1$, we have $\tau = \lambda_p/v$ which, for our standard metal, is $1.6 \times 10^{-14} \text{ s}$. A metal which is a poor conductor, such as constantan ($\tau \approx 10^{-15} \text{ s}$) will therefore not enter the anomalous region D as the frequency is increased, but the relaxation region B.

As an elegant confirmation of the theory of the anomalous skin effect we reproduce in fig. 10 some experimental results obtained by Chambers [13]. This

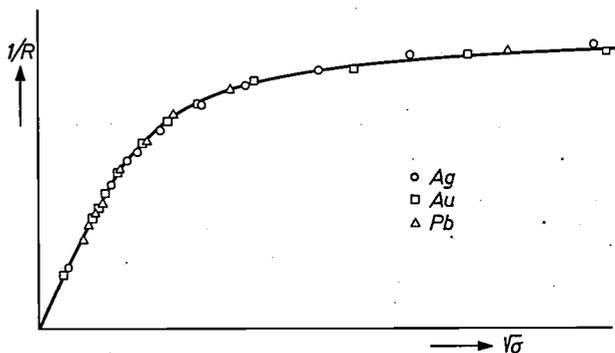


Fig. 10. Surface conductivity $1/R$ of silver, gold and lead as a function of $\sqrt{\sigma}$ at 3600 MHz, after Chambers [13]. The points represent measured values obtained at various temperatures. The drawn curve is calculated from the theory of the anomalous skin effect with diffuse reflection of the electrons at the surface. For each material the scale is chosen so as to give the best fit between experimental values and theoretical curve. (On the basis of specular reflection the best fit that can be obtained is not nearly so good.) This matching process gives σ/l , a quantity which is independent of the sensitive parameters (in our notation $\sigma/l = 1/\mu\nu\lambda_p^2$). The initial portion corresponds to the classical limit, where the relationship is a linear one; the right-hand portion corresponds to the extreme anomalous limit, where R is independent of σ .

figure shows the surface conductivity ($1/R \propto 1/\delta'$) for a given frequency, plotted against the root of the conductivity σ (on arbitrary scales). According to the classical theory these quantities are proportional to one another: $1/R = \sqrt{2\sigma/\mu\omega}$. The most striking feature of δ' in the anomalous limit is its independence of the sensitive parameters, e.g. σ (see equations (40) and (41)). This is neatly demonstrated in fig. 10.

According to (40) and (41) in the anomalous limit and at a given frequency δ' is entirely determined by n_e (through v and λ_p). Rigorous application of the free-electron theory gives in this limit:

$$n_e^2 = 3\pi^2(n/4b\mu e^2\omega\delta'^3)^3. \quad (43)$$

Chambers used this to find the number of free electrons per atom for various metals; see Table IV [14].

The fact that values near unity are found for the simple metals may be seen not only as a confirmation in general terms of the theory outlined above, but also as an indication that the surfaces used by Chambers were treated with great care. If the surface is not smooth and free from stresses the values found for δ' can easily be on the large side, resulting (as $n_e \propto \delta'^{-3/2}$) in values for n_e which are far too low. Moreover, the free-electron theory can be expected to be a reasonable approximation only for the simplest metals. For these reasons these results should be seen rather as a confirmation of the theory than as an accurate determination of n_e .

Table IV.

Cu	1.0
Ag	0.68
Au	0.60
Sn	1.10
Hg	0.23
Al	0.37

[12] This relation can also be derived directly from the Kramers-Kronig relations, if it is assumed as given that $\delta \propto \omega^{-1/3}$ (see Pippard [11]).

[13] R. G. Chambers, Proc. Roy. Soc. A 215, 481, 1952.

[14] E. H. Sondheimer, Phil. Mag. Suppl. 1, 1, 1952.

Anomalous reflection

The question that immediately arises on considering fig. 9 is whether relaxation effects arise in metals that are in the anomalous region if the frequency is increased still further. Conversely, we may ask whether anomalies comparable with the anomalous skin effect occur if, at a given frequency, the mean free path increases (e.g. because the temperature is reduced) in a metal in the relaxation region B. And in either case, what may we expect to happen to the penetration depth and the surface impedance?

Let us first increase the frequency, starting from the anomalous region D. The characteristic difference between regions A and B was connected with the number of periods of the alternating field which the electron experiences between two collisions: this was much smaller than 1 in A ($\tau \ll 1/\omega$) and much greater than 1 in B ($\tau \gg 1/\omega$). In the extreme anomalous region, however, the relaxation time τ is no longer significant and therefore no effect can be expected when the boundary $\tau = 1/\omega$ is passed. Instead, what now matters is whether the electron undergoes many periods of the alternating field or only a fraction of a period *during the time it spends in the skin layer*. This time is about $|\delta|/v$, and the new boundary is therefore $|\delta|/v = 1/\omega$. Obtaining $|\delta|$ from (40) and disregarding the factor $2/b$ this is equivalent to:

$$\omega/\omega_p = v/c. \quad \dots \dots \dots (44)$$

Coming from D and passing this boundary we enter the region E (fig. 11) which we shall call the region of *anomalous reflection* (see also E in fig. 2). Here, as in B, the field sees a layer of virtually free electrons, and the penetration depth, as in B, is equal to λ_p . The time spent in the skin layer is now approximately λ_p/v ; putting this equal to $1/\omega$ we again find the boundary (44) between D and E.

What difference now remains between regions E and B? To the zero order of approximation there is no difference: in both cases the layer of free electrons has a screening effect (penetration depth λ_p) but it does not absorb, since the velocity of an entirely free electron in an alternating field of *constant* amplitude is always 90° out of phase with the field. The difference is of the first order. In B some absorption takes place because the free motion of the electrons is interrupted occasionally by a collision. In E the reflection of the electron at the surface must be held responsible for the absorption (at least if the reflection is not specular); a small contribution (becoming smaller with increasing frequency) to the absorption in E arises because the electron is moving through the skin layer and thus sees an alternating field of *varying* amplitude (Holstein [15]).

The boundary between B and E (see fig. 11) lies

approximately where the mean free path is equal to the penetration depth: $l = \lambda_p$, or:

$$\tau = \lambda_p/v. \quad \dots \dots \dots (45)$$

In region E — in contrast to D — it makes a great difference whether the reflection is specular (and therefore imperceptible to the alternating field since the tangential velocity is unchanged) or diffuse (and therefore equivalent to a collision). To make this plausible, let us consider how energy is transferred from the alternating field to an electron. Suppose that an electron is subject, between collisions, to a homogeneous alternating field. For cases A and B this is a reasonable approximation ($l \ll \delta$). The velocity of the electron immediately after a collision is completely arbitrary and on average makes no contribution to the transfer of energy from the field to the electron. The additional velocity imparted to the electron as a result of the

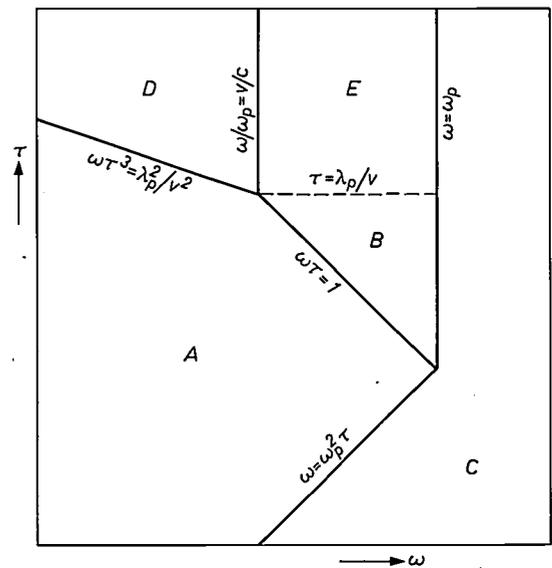


Fig. 11. The ω - τ -diagram as in fig. 9, with the new region E, the region of "anomalous reflection". Both in E and B eq. (34) is valid to the zero order approximation. To a first approximation, the absorption is given in B by eqs. (35) and (14) and in E by (46).

acceleration by the alternating field always results initially in a positive transfer of energy from the field to the electron. If $\omega\tau \ll 1$, then at a given field amplitude this energy transfer between two collisions increases as τ increases. On the other hand, if $\omega\tau \gg 1$, the total energy transfer between two collisions will be independent of τ , because, after the initial transfer of energy during about one period, energy will be transferred alternately from the electron to the field and vice versa, with zero net transfer of energy. The energy transfer *per second* therefore decreases as τ increases.

Going now from A to D, we may make use of an

“effective relaxation time”, for in case D the electrons enter the skin layer suddenly. They begin by taking up energy as they enter (disregarding their random velocity upon entry), and long before the field has completed one period the energy exchange is interrupted, either at the surface (diffuse reflection) or upon emerging from the skin layer (specular reflection). The ratio of the absorptions for diffuse and specular reflection would be expected to be about 2 at the most.

In going from B to E the situation is different. In case E the electron enters the skin layer very “gradually” with respect to the alternating field: the electron experiences many periods in the time that it sees the field grow from zero to its value at the surface. There is no distinct initial energy transfer, and positive and negative contributions continue approximately to compensate one another. If at the surface the reflection is specular the situation remains unchanged, and the total absorption is much smaller than in B. If however the reflection is diffuse, then immediately after a reflection, just as after a collision in B, there is a distinct transfer of energy from field to electron.

We can now make an estimate [16] of the absorptivity for region E with diffuse reflection by using the absorptivity for region B but with an *effective* τ , this being the time spent in the skin layer per collision: $\tau_{\text{eff}} = 2\lambda_p/v$. Using (35) and (14) we find that the absorptivity in B is given by $a = 2/\omega_p\tau$. It then follows for region E that $a = 2/\omega_p\tau_{\text{eff}} = v/c$. This differs only by a factor of 4/3 from the absorptivity found by Holstein [15] in an exact calculation for diffuse reflection:

$$a = \frac{3}{4} v/c. \quad (46)$$

Holstein also made the calculation for specular reflection, but since all experimental results indicate that the reflection is entirely diffuse, we shall not consider this here. Holstein’s results are implicit as a limiting case in the Reuter and Sondheimer theory [17].

For our standard metal the boundary (45) lies at 1.6×10^{-14} s, so that the measurements by Försterling and Freedericksz (see Table II) lie more or less in the boundary region. In Table V we have added the values for $\tau_E = 8\lambda_p/3v$ to the data of Table II: τ_E is the effective τ which, with (35), gives the absorption in E. It is clear that the theory for the E-region does not explain these optical results: in the first place τ_E in the E-region ought to be lower than the actual τ ; moreover τ_E is even further away from τ than $\tau(\text{opt})$. For low values of $\tau(\text{opt})$ therefore, we still have to resort to an explanation like that mentioned on page 308.

Table V.

τ in 10^{-14} s	Cu	Ag	Au
$\tau(\text{opt})$	0.46	0.95	1.7
$\tau(\sigma)$	2.5	3.9	2.9
τ_E	3.1	4.2	4.2

To verify the theory in the E-region, and in particular equation (46), it is necessary to carry out absorption measurements at high values of τ , that is to say on pure metals at low temperatures, paying particular attention to keeping the surface clean. A detailed comparison of published experimental material with the Reuter and Sondheimer theory has been given by Dingle [18]. By way of illustration we shall just mention Ramanathan’s results [19] and those found by Biondi [20].

For copper at $1.4 \mu\text{m}$ and 4.2°K Ramanathan measured an absorptivity of 0.006. For our standard metal, equation (46) gives an absorptivity of 0.0035. The agreement is very reasonable, particularly compared with the value of $a \approx 0.00003$ predicted for this case by the classical theory, from (35).

Biondi has measured the absorptivity of copper and silver, in the wavelength region 0.3 to $3.3 \mu\text{m}$ and at about 4.2°K , with τ values (derived from the d.c. resistance) of about 10^{-11} s. In addition to a few expected absorption bands (cf. page 308) at the shorter wavelengths, he found for $\lambda > 1.5 \mu\text{m}$ that, in qualitative agreement with (46), the absorptivity was independent of the wavelength (and also independent of temperature, although the temperature interval covered was small, being 3.4 - 4.2°K). Numerically, Biondi’s conclusion is that, even with a virtually ideal surface and far from the absorption bands, equation (46) accounts to a considerable extent but not *fully* for the absorption measured in the extreme E-region. From a further theoretical study of the processes possible in the E-region, a “volume-absorption process” in the skin layer has been put forward (as opposed to the surface process of (46)) and this is held to be responsible for the remaining absorption.

Table VI presents the theoretical and experimental values given by Biondi.

Table VI. The absorption factor of copper and silver, after Biondi [20].

	Cu	Ag
Surface-effect (eq. 46)	0.0029	0.0036
Volume-effect	20	09
Total	49	45
Experimental	50	44

[15] T. Holstein, Phys. Rev. 88, 1427, 1952.

[16] See also Holstein’s estimate, given by Biondi [20].

[17] R. B. Dingle, Physica 19, 311, 1953.

[18] R. B. Dingle, Physica 19, 348, 1953.

[19] K. Ramanathan, Proc. Phys. Soc. A 65, 532, 1952.

[20] M. A. Biondi, Phys. Rev. 102, 964, 1956.

Briefly, the volume-absorption process may be described as follows. In pure metals the absorption correlated with the "normal" relaxation time τ (in A and B) is due to electrons colliding with phonons (the quanta of the thermal lattice vibrations). At very low temperatures no phonons are present ($\tau \rightarrow \infty$). However, an absorption process of the following kind is still possible: the electron absorbs a photon (quantum of the alternating field) and simultaneously generates a phonon as a result of its interaction with the crystal lattice.

Concluding remarks

Fig. 12 shows once more the ω - τ -diagram, now with a scale in seconds along the τ -axis and a scale in radians/second along the ω -axis. A scale for the wavelength λ is also added. The boundaries between the regions A, B, C, D and E are indicated in the diagram

for a number of monovalent metals. These have been calculated from the expressions from the free electron theory, in which n_e is determined from the atomic weight and the density with the assumption of one conduction electron per atom.

It can be seen from this figure that the limiting cases of the free electron model that interest the communications engineer differ from those that interest the optical physicist. In communications we are not as yet concerned with waves shorter than 1 mm, and are therefore interested only in the normal skin effect (A) and the anomalous skin effect (D). In optical physics, in the visible spectrum (hatched area on λ -scale), or in the near infra-red, we are interested only in the relaxation region, with normal reflection B or anomalous reflection

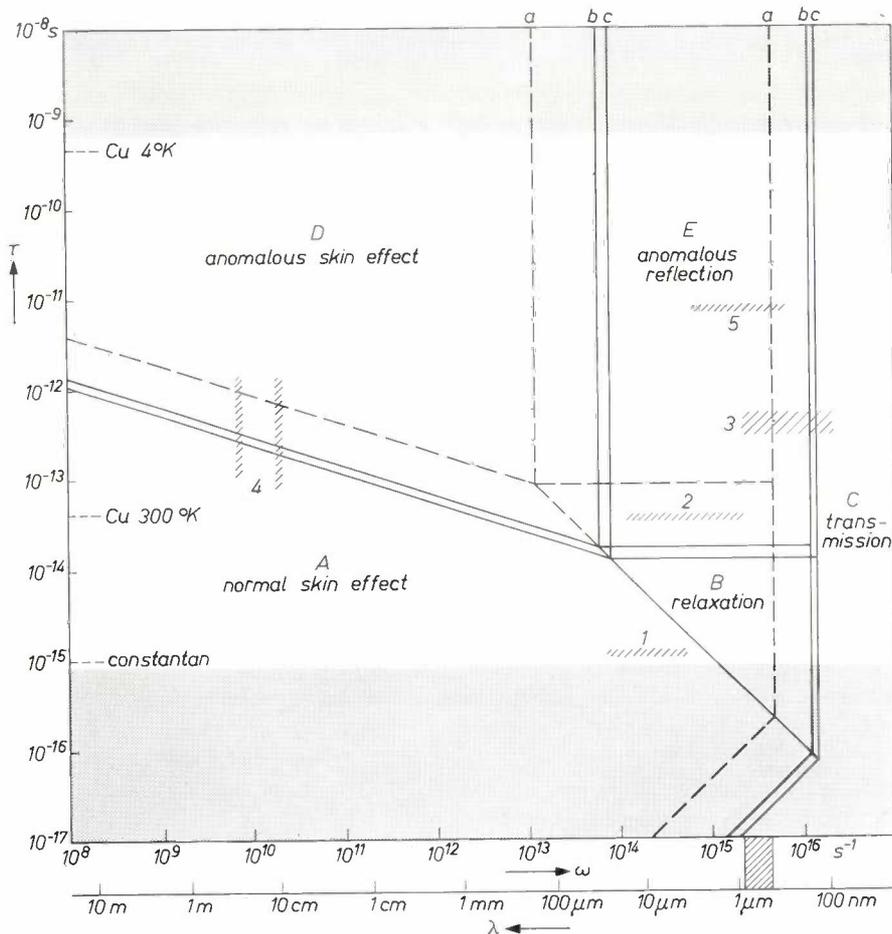


Fig. 12. The ω - τ -diagram for various monovalent metals. The electron concentration n_e has been obtained from density and atomic weight on the assumption that there is one electron per atom; the Fermi velocity v and plasma frequency ω_p have been derived from this using the expressions for the free-electron model. *a*) Caesium, *b*) the standard metal, *c*) copper. The values of n_e for the other alkali metals and for silver and gold lie between those for caesium and copper, and thus correspond to intermediate diagrams. A classical skin effect region, B relaxation region with normal reflection, C transmission region, D anomalous skin effect region, E relaxation region with anomalous reflection.

The finely-shaded strips indicate the ω - τ -region where measurements have been carried out by: 1 Hagen-Rubens [5], 2 Försterling-Fredericksz [7], 3 Wood [8], 4 Chambers [13], 5 Biondi [20]. (The τ -value of 3 is uncertain; Wood gave no values for conductivity.)

A wavelength scale is also given along the horizontal axis; the hatched area corresponds to the visible spectrum.

tion E. In optics we are concerned with the A-region only for the infra-red (see for example the hatched area *I*, corresponding to the experiments of Hagen and Rubens mentioned on page 307). The region D is right outside the domain of optics. E differs from B in that the absorption is considerably higher than would be calculated with the expression appropriate to B, the reason being that the collisions at the surface begin to become relatively more significant than the collisions with the lattice.

The difference between D and A is of a similar nature to that between E and B: the absorption in D is greater than would be derived from the expressions for A. We shall illustrate this by taking a microwave cavity resonator whose Q (quality factor) is to be improved by using a material of higher conductivity. Let the cavity be cylindrical, with equal height and diameter ($2a$) and let it be excited in the TE_{011} mode (the wavemeter mode). In this case $a = 0.66 \lambda$ (where λ is the wavelength in free space) and $Q = a/2\delta'$. For $\lambda = 3$ cm and copper at room temperature ($\sigma = 6 \times 10^7 \Omega^{-1}\text{m}^{-1}$ and $\tau = 4 \times 10^{-14}$ s) the classical skin effect theory is still valid (see fig. 12), so that according to eq. (15) $\delta' = \frac{1}{2} \delta_K$. Using $\delta_K^2 = 2\lambda_p^2/\omega\tau$ (eq. 26) we find $\delta' = 0.3 \mu\text{m}$ and $Q = 30\,000$ (for λ_p , and for v below, we take the value for the standard metal, given in Table I). Now for a metal where τ is 10^4 times greater (e.g. very pure copper at 4 °K), we would expect, on the basis of the classical expressions, δ' to be 100 times smaller, and Q therefore 100 times greater. Fig. 12 shows, however, that the anomalous limit should be a better approximation than the classical limit for this value of τ . Using $\delta' = \frac{1}{2}|\delta|$ and $|\delta|^3 = (2/b)(v/\omega)\lambda_p^2$ — see eqs. (40) and (41) — we find $\delta' \approx 0.07 \mu\text{m}$, which is smaller by a factor 4. Thus, the Q of the cavity is only 4 times, rather than 100 times, greater than it was for copper at room temperature.

In optical physics it is usually the case that the free electron effects in B and E, and in particular the absorp-

tion mechanism we have described, are swamped by the absorption bands, which are of different origin. This happens to an increasing extent at higher frequencies, where an increasing number of processes can be excited by the radiation. Consequently, the transmission region C, which on the basis of the free-electron model for metals would be expected in the ultra-violet or X-ray regions, is in practice of no importance for metals. Nevertheless, the fairly sharp transition from B or E to C, the plasma boundary, has been convincingly demonstrated by Wood's experiments [8].

The simplified theory outlined in the foregoing is of course generally too approximate for quantitative and detailed study of the phenomena. Let us quote two of the shortcomings that very soon appear upon closer examination. In the first place it is too naïve to assume that only one relaxation time is sufficient for the whole range of frequencies. Secondly, in nearly all metals the velocity of the electron (the Fermi velocity) depends upon the direction in which it moves in the crystal structure.

In spite of these and other shortcomings, we hope that the present article gives some idea of the relationships that exist between a variety of phenomena.

Summary. The skin effect at high frequencies is examined for the case of an electromagnetic wave at normal incidence upon the plane surface of a metal. In this case simple relations can be given between the concepts "skin depth", "refractive index" and "surface impedance". The complications which arise at high frequencies are discussed with the aid of the free-electron model for a metal. In Drude's classical model it is found that at increasing frequencies the frequency range of the classical skin effect is followed first by a "relaxation range" (with total reflection); at still higher frequencies the metal becomes transparent. In pure metals, taking account of the mean free path of the electrons, and in particular of the fact that this can be greater than the skin depth, one finds an "anomalous skin effect" at frequencies between the classical range and the relaxation range. Moreover, in the relaxation range the absorption becomes anomalous. In both cases the absorption is greater than would be calculated from the classical theory.

Crystal growth by temperature-alternating methods

H. Scholz

The growth of single crystals by means of chemical transport reactions has been under investigation for some time at the Philips Aachen laboratories [1]. A chemical transport reaction is brought about by means of a temperature gradient in a system consisting of a solid phase and a transporting liquid or gaseous phase; the reaction is based on the temperature dependence of the equilibrium constant of the system. If the transport reaction is endothermic the transport takes place from the hot to the cold zone, the transporting phase being unsaturated in the hot zone and supersaturated in the cold zone, with a saturation boundary in between. If the reaction is exothermic the transport goes in the opposite direction.

We have been able to show that for growing single crystals it is particularly advantageous to reverse the direction of the temperature gradient periodically instead of keeping it constant, which is the usual practice. With an alternating temperature gradient the material is transported towards the place with the lowest *average* temperature (in endothermic reactions). As a result of the alternation of the temperature gradient, intervals of crystal growth alternate with intervals of etching, making the crystal growth highly selective. In systems where the crystal growth in a constant temperature gradient is associated with high seed rates, so that only small crystals are formed, the use of an alternating temperature gradient can considerably reduce the number of seeds and thus lead to the formation of larger crystals.

If the time of growth is sufficiently long, crystals will finally remain only at the place which has the lowest average temperature (see *fig. 1*, which shows the transport of GaP for constant and alternating temperature gradients). It has been found that good results are achieved if, right at the beginning of growth, a specially selected seed crystal is placed at this position; the spontaneously formed seed crystals do not always have the perfection that might be desired.

An interesting feature of our method is that at any point of the system the temperature during certain intervals during the growth will be lower than at the place where crystal growth is greatest (the place with the lowest *average* temperature). Crystal growth does not depend on the greatest temperature difference

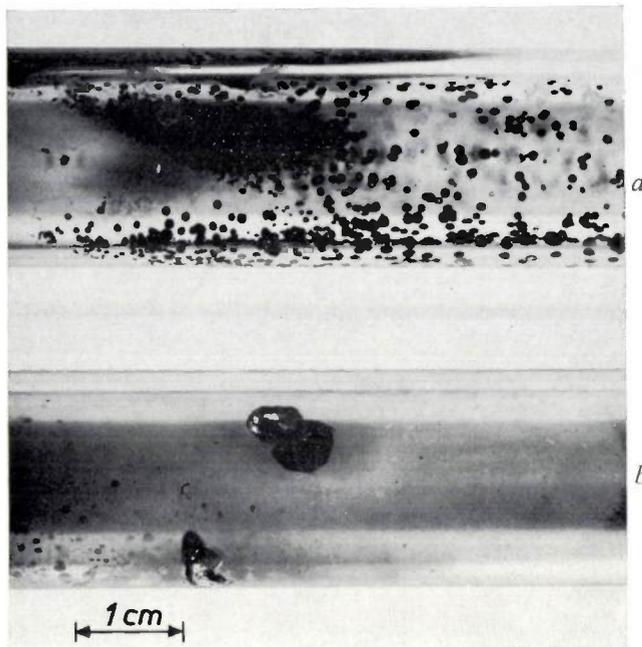


Fig. 1. Crystal growth of GaP using bromine as transport medium, with *a*) constant temperature gradient and *b*) alternating temperature gradient. The pressure and temperature conditions in these experiments were otherwise comparable. In (*b*) the crystals have formed in the middle of the tube, as this is the place with the lowest average temperature.

present in the system but on the difference in *average* temperature. With this method it is possible to obtain very small effective temperature differences by means of differences in average temperature.

Both *linear* and *axially symmetrical* experimental arrangements have been developed [2]. In this article we shall be concerned solely with the axially symmetrical methods, of which two types are possible in principle: I. An axially symmetrical reaction vessel is rotated in a space where the temperature gradient is perpendicular to the axis. The period of the temperature gradient variation at each point in the vessel is equal to the time taken for the reaction vessel to complete one revolution. II. The temperature of the whole wall is periodically raised and lowered with respect to the temperature at the centre of the reaction vessel. The only object of rotating the vessel in this case is to smooth out temperature irregularities which arise because the arrangement is not perfectly symmetrical.

Fig. 2 shows a cross-section of the experimental arrangement. A rotatable reaction vessel 1 is enclosed by two glass domes 4 and 5; the first is of quartz glass

Dr. H. Scholz is with the Aachen laboratory of Philips Zentral-laboratorium GmbH.

and can be heated by a heater coil, the second is of a heat resistant glass, and serves to minimize heat losses by convection. The source material 2 is placed along the wall of the reaction vessel, and the seed crystal 3 in the centre. In method I the required temperature gradient is produced with the aid of a heat-reflecting screen 6. In method II this screen is omitted.

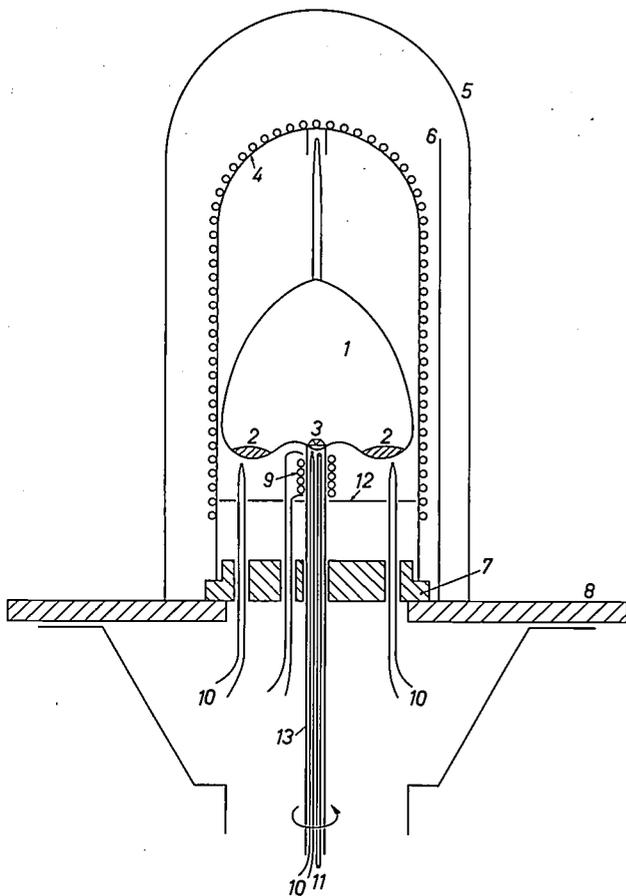


Fig. 2. Experimental arrangement for axially symmetrical methods. 1 reaction vessel, which is slowly rotated. 2 source material. 3 seed crystal. 4 quartz glass dome, whose top and cylindrical parts can be heated by two separate heater coils. 5 heat-resistant glass dome to minimize heat losses due to convection. 6 heat-reflecting screen which helps to set up a temperature gradient perpendicular to the axis. 7 ceramic support for quartz glass dome. 8 baseplate of ceramic material. 9 heating for base. 10 three thermo-couples. 11 air-cooling tube. 12 radiation screen. 13 hollow drive shaft for reaction vessel; this shaft is rotated by a motor.

The arrangement described here can be used to obtain temperatures up to 700 °C. If the inside of the outer dome is coated with a layer of gold a temperature of 900 °C can be achieved. (This applies only to method II; the gold-coated glass would in fact not permit the temperature gradient required in method I.)

Let us now take a closer look at method I, starting with the cross-section of the reaction vessel shown in *fig. 3*. Below the cross-section can be seen the temperature gradient along the line where the temperature drop is greatest, shown for two temperature distributions.

First of all we want to know the temperature conditions under which there will be transport of material from the wall of the reaction vessel to the centre. As the vessel has axial symmetry it is sufficient to use the average temperature \bar{T}_p of one point of the wall and the temperature T_c at the centre.

If the temperature drop along the line under consideration is linear, then $\bar{T}_p = T_c$ and no material transport will take place. In this case the saturation

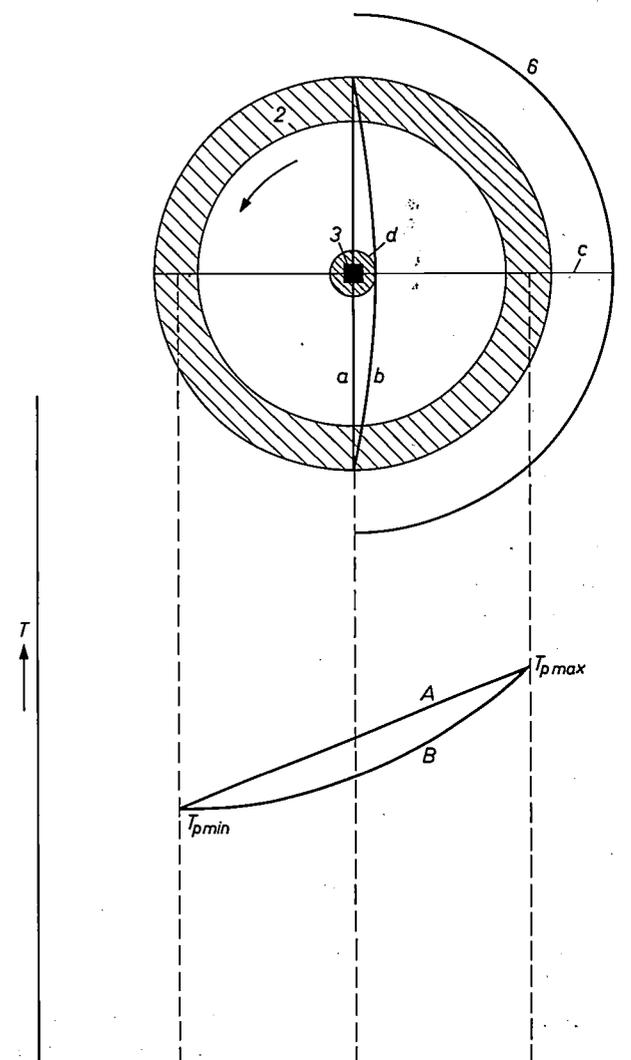


Fig. 3. At the top: transverse cross-section through the reaction vessel. 2 source material. 3 seed crystal. 6 reflecting screen. *c* line along which the temperature gradient is greatest. The lines *a* and *b* indicate the saturation boundaries that occur when the temperature along line *c* has the distribution *A* or *B*, as shown at the bottom of the figure.

In distribution *A* no crystal growth is possible. *d* is the limit up to which crystal growth is possible with a temperature distribution *B*.

[1] See A. Rabenau, Philips tech. Rev. 26, 117, 1965.

[2] For more details about this method see H. Scholz and R. Kluckow, Chemie-Ing.-Technik 37, 1173, 1965, and Crystal Growth (suppl. J. Phys. Chem. Solids), Proc. int. Conf. Boston 1966, p. 475.

boundary of the system passes through the centre at right angles to the direction of the maximum temperature drop (curve *a*). If the temperature at the centre is reduced to bring about the transport of material from the wall to the centre the saturation boundary is displaced, with the result that the seed crystal enters the supersaturated region (curve *b*).

When the reaction vessel is in rotation, all the material not too close to the centre passes alternately through a supersaturated and an unsaturated region. It is here that the required crystal selection takes place. No spontaneously formed seed has any chance of survival in this region.

The small zone around the centre on the other hand is continuously supersaturated. Because of this steps have to be taken to ensure that only the seed crystal introduced into the system is present in this zone. This is easily accomplished by keeping the centre for some time at a temperature higher than T_p before the growth process begins. It is of course still possible for seeds to form in the supersaturated zone during the growth period, but this has been only a very rare occurrence in our experiments, which is not surprising bearing in mind how very small the relevant zone is.

Fig. 4 shows a single crystal of $\alpha\text{-Fe}_2\text{O}_3$ grown by method I. The experimental data for this crystal are given in Table I^[3].

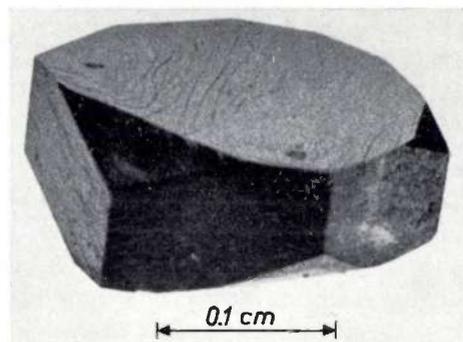


Fig. 4. An $\alpha\text{-Fe}_2\text{O}_3$ crystal grown by method I. The experimental data are given in Table I.

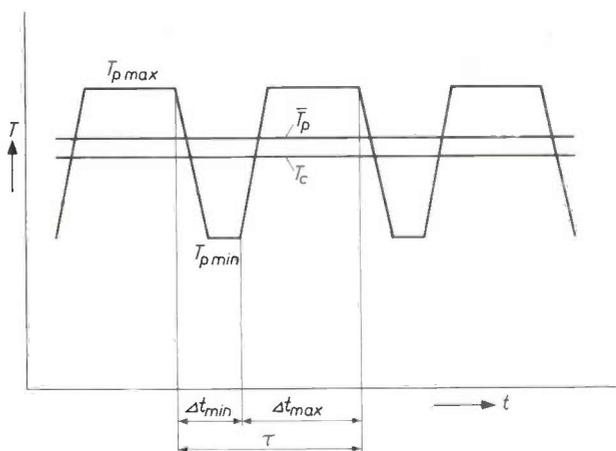


Fig. 5. Temperature-time curves for the axially symmetrical method II in which rotation of the reaction vessel is only necessary for smoothing out small irregularities in temperature.

Table I. Experimental data for the growth of an $\alpha\text{-Fe}_2\text{O}_3$ crystal by method I.

Weight of seed crystal	6.8 mg
Transport medium at 20 °C	100 torr HCl and 118 torr Cl ₂
$T_{p \max}$	583 °C
$T_{p \min}$	533 °C
T_c	545 °C
Period of rotation	100 min
Growth time	600 h
Weight of crystal	19 mg

Fig. 5 gives the temperature-time curves of an experiment carried out by method II. The temperature at the centre was kept constant while the wall temperature was periodically raised and lowered. This method also results in crystal selection at the centre, because in the time during which the temperature is lowered the centre is hotter than the whole of the wall. The centre also therefore alternates between an unsaturated and a supersaturated state.

Table II presents the experimental data for two crystals grown by method II. Fig. 6 shows one of these crystals measuring about 1 cm. Fig. 7 shows the reaction vessel with the crystal and the remainder of the source material.

A comparison of methods I and II leads to the following conclusions. Method II is generally preferable. The fact that selection also takes place in the

Table II. Experimental data for two experiments with $\alpha\text{-Fe}_2\text{O}_3$ by method II.

	Experiment 1	Experiment 2
Weight of seed crystal	1.5 mg	1.1 mg
Transport medium at 20 °C	100 torr HCl	112 torr HCl
$T_{p \max}$	592 °C	594 °C
$T_{p \min}$	542 °C	546 °C
T_p	575 °C	578 °C
T_c	556 °C	558 °C
Δt_{\max} [*]	1 h	1 h
Δt_{\min} [*]	0.5 h	0.5 h
Growth time	2400 h	428 h
Weight of crystal	398 mg	approx. 100 mg

[*] $\Delta t_{\max} + \Delta t_{\min} = \tau$, see fig. 5.

[3] The growth rate in this experiment is low compared with that in the experiments carried out by method II, discussed in the following. This is explained by the fact that the actual temperature in the centre of the reaction vessel differs considerably from the value determined by measurement from outside. The actual values of T_c and T_p lie much closer together than indicated here. This was found from measurements carried out simultaneously inside and outside the vessel.

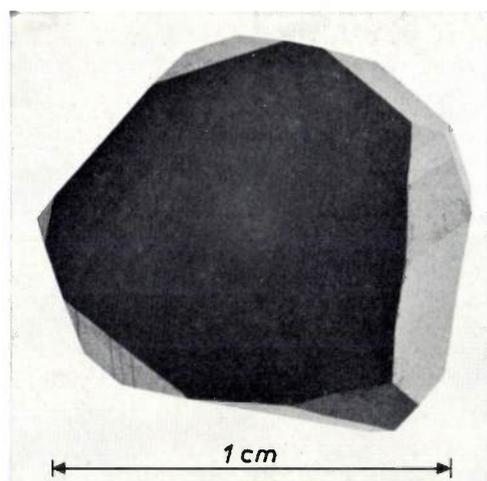


Fig. 6. An $\alpha\text{-Fe}_2\text{O}_3$ crystal grown by method II. Experimental data are given in Table II.

centre makes the method basically much simpler. Moreover, the intervals of etching of the growing seed crystal have the useful effect of eliminating imperfections in the crystal. Method II cannot however be used if liquid is formed in the system during the growth process, since condensed drops of liquid can adversely affect crystal growth. Using method I it is possible to ensure (provided the reaction vessel is rotated slowly enough) that the formation of liquid remains limited to the place where the wall temperature is lowest.

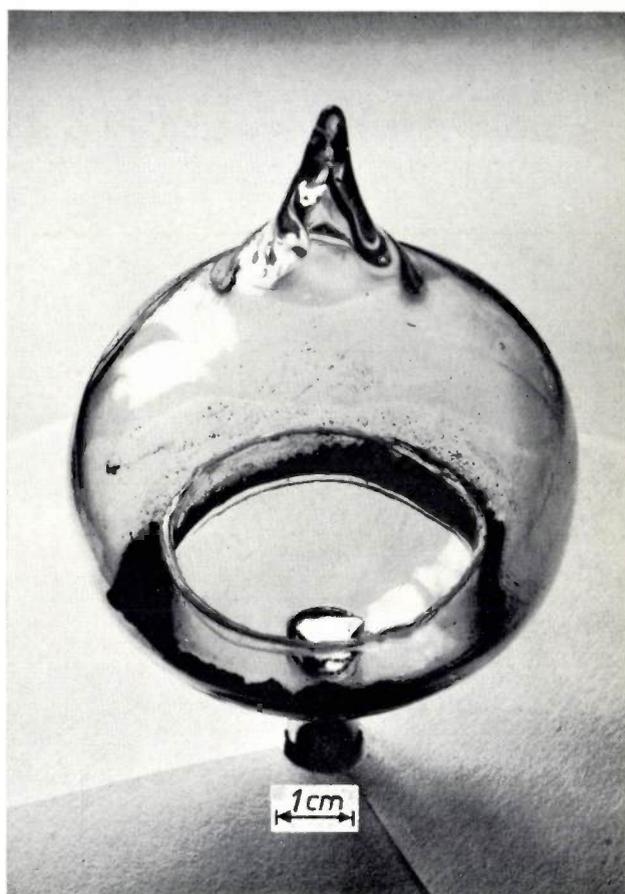


Fig. 7. The reaction vessel with a crystal grown by method II. The remainder of the source material can also be seen.

Summary. Experiments on crystal growth by means of chemical transport reactions have been in progress for some time at the Aachen laboratories of Philips Zentrallaboratorium GmbH. A constant temperature gradient has normally been used in these experiments, and with some systems this can have the result that only very small crystals are formed, because the seed rate is too high. In such cases the use of an alternating temperature gradient can considerably reduce the number of seeds formed, so that larger crystals are formed. In the methods based on this principle, crystal growth is governed by the

average temperature in the system during crystal growth. The experiments described are carried out with an axially symmetrical arrangement. In one such method an axially symmetrical reaction vessel is rotated in a space where the temperature gradient is perpendicular to the axis. The source material is placed along the wall of the vessel and a seed crystal is placed at the centre. In another method the temperature of the whole wall of the reaction vessel is periodically raised and lowered with respect to the temperature of the seed crystal at the centre. Single crystals of $\alpha\text{-Fe}_2\text{O}_3$ of about 1 cm have been grown in this way.

A miniature electric motor for experiments on vision

Physiological experiments have shown that when an object is seen by a human observer the image projected on to the retina of the eye continuously alters its position on the retina, even if the observer tries to fix his vision on the object. This is due to small random oscillations and movements of the eyeball which are not consciously experienced by the observer. If movement of the eyeball with respect to the object is prevented by a mechanical connection between eyeball and object (as described below), a normal observer finds that the object "disappears" after a few seconds, i.e. he can no longer see the object.

This indicates that an image on the retina is transmitted to the brain through transitional phenomena alone.

be kept down to a few grammes, and in addition there were difficult requirements for its geometry and performance (smoothness of running). In the following short description we attempt to show how current techniques of miniaturization have been used to achieve the desired results. The motor weighs less than 4 grammes.

The essentials of the arrangement used in the experiments are shown in *fig. 1* (this arrangement was developed at Nijmegen, apart from the motor). The observer lies on his back, and an aluminium tube about 2.5 cm long is placed on one eyeball, at the iris, and held in place by gentle suction. The object for viewing is at the end of the tube; this is a black dot or other shape drawn off-centre on a transparent disc (2)

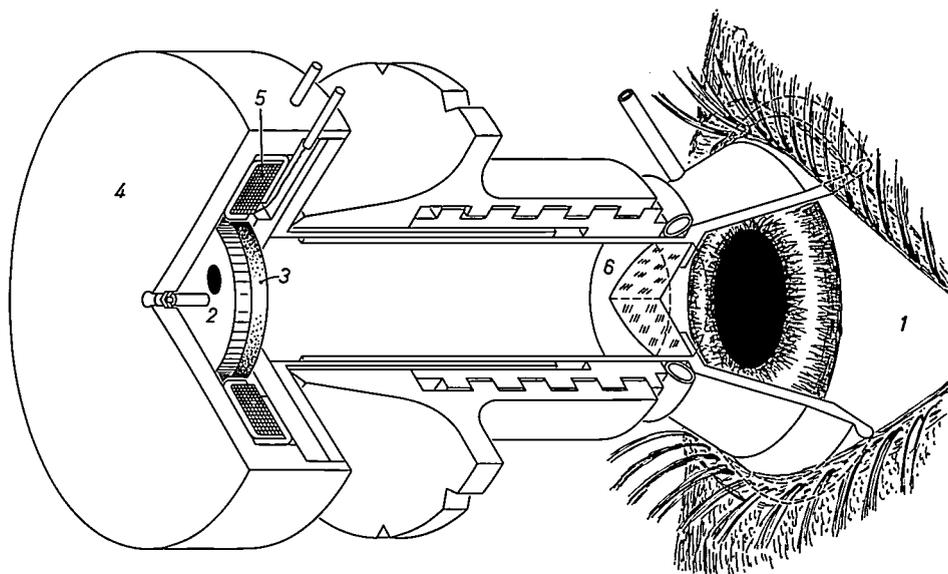


Fig. 1. Cut-away view of the device mounted on the eyeball. 1 eyeball. 2 transparent disc with object to be viewed (a black dot) and annular rotor 3. 4 transparent stator housing with stator coil 5. 6 lens.

Extensive experiments on this subject have been performed by Dr. H. J. M. Gerrits at the University of Nijmegen^[1]. For these experiments it was not only desirable to be able to hold the object for viewing fixed with respect to the eyeball, it was also desirable to be able to give the object a movement of variable speed with respect to the eye. An electric motor to meet these requirements has been designed in these Laboratories at Dr. Gerrits's request. The weight of the motor had to

A lens (6) placed in the tube a few millimetres away from the eye enables the eye to see the object in sharp focus. The disc has to be able to rotate, and the motor for this is arranged around it: a spindle attached to the disc is carried in two ruby bearings mounted in a

[1] H. J. M. Gerrits, thesis, University of Nijmegen, 1967.

[2] W. L. L. Lenders, The orthocyclic method of coil winding, Philips tech. Rev. 23, 365-379, 1961/62.

transparent stator housing (4). At its circumference the disc carries an annular ferroxdure rotor (3) with 16 regularly spaced radial poles, alternately N and S. The stator housing is attached to an aluminium sleeve which fits over the tube; both tube and sleeve are fitted with a plastic screw thread so that the motor and object can be brought to the proper distance for viewing.

The stator consists of two iron annuli which fit together like the two halves of a shoe-polish tin; see *fig. 2*.

large or a 2-phase motor is used. However, neither of these alternatives could be attempted here on account of weight restrictions, and in order to present to the eye a smoothly rotating object we have turned to a method using stroboscopic illumination. The motor is run at the top speed (50 r.p.s.), where the angular velocity is very steady, and a stroboscope lamp is used as the light source: the lamp is made to run at a little more or a little less than 50 flashes per second by means of a

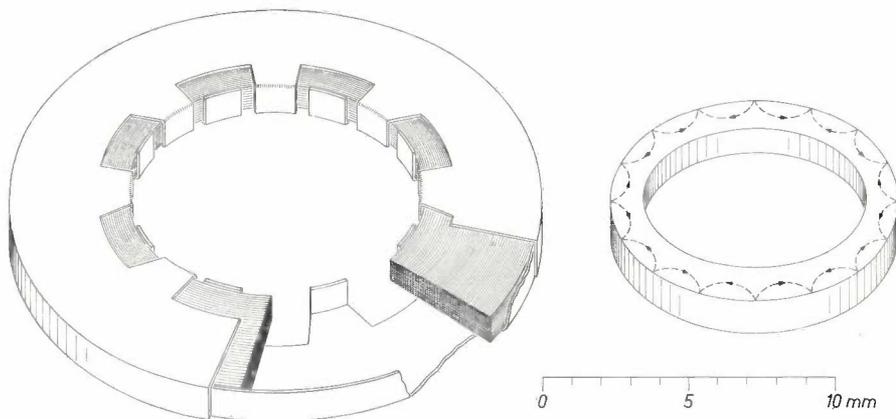


Fig. 2. Left: stator, consisting of two annular iron "lids" fitting together with the coil between. Right: annular ferroxdure rotor, with 16 poles. The dotted lines indicate the magnetization. The air gap between rotor and stator is less than 0.1 mm.

Each half has eight poles and the complete assembly encloses a coil of 950 turns of 50 μm dia. copper wire. A manageable coil is obtained, at the same time making the best use of the available space, by the use of orthocyclic winding [2]. The coil is supplied at an a.c. voltage of 5 V from an audio generator whose frequency can be varied from 1 to 400 Hz; the power taken is 50 mW.

The motor operates in essentially the same way as a single-phase stepping motor. At about 16 Hz the rotor becomes resonant and starts into synchronous rotation. If the supply frequency is then varied the motor remains in synchronous rotation (from 1/8 to 50 r.p.s.), and the observer sees the object rotating about the centre of the field of view when the object is illuminated by a source behind the transparent stator housing.

The type of motor described must necessarily deliver its torque as a series of pulses. At low speed there is therefore a noticeable ripple in the angular velocity, unless the moment of inertia of the rotor is made

suitable circuit. The observer then sees the object rotating at the difference frequency.

In addition to the motor supply leads, flexible connections for maintaining the suction in the cylinder and also for water feed and return have to be provided: water is circulated through an annular water-jacket around the lens (6 in *fig. 1*) and prevents misting of the lens by maintaining it at 37 °C.

All these requirements result in a device, which although miniaturized, does not have negligible bulk and weight. A mild local anaesthetic is used to prevent discomfort when the device is applied to the eyeball.

For a more detailed description of the device the reader is referred to the thesis quoted above [1].

B. Bollée

Ir. B. Bollée is with Philips Research Laboratories, Eindhoven.

Recent scientific publications

These publications are contributed by staff of laboratories and plants which form part of or co-operate with enterprises of the Philips group of companies, particularly by staff of the following research laboratories:

Philips Research Laboratories, Eindhoven, Netherlands	<i>E</i>
Mullard Research Laboratories, Redhill (Surrey), England	<i>M</i>
Laboratoires d'Electronique et de Physique Appliquée, Limeil-Brévannes (S.O.), France	<i>L</i>
Philips Zentrallaboratorium GmbH, Aachen laboratory, Weisshausstrasse, 51 Aachen, Germany	<i>A</i>
Philips Zentrallaboratorium GmbH, Hamburg laboratory, Vogt-Kölln-Strasse 30, 2 Hamburg-Stellingen, Germany	<i>H</i>
MBLE Laboratoire de Recherche, 2 avenue Van Becelaere, Brussels 17 (Boitsfort), Belgium.	<i>B</i>

Reprints of most of these publications will be available in the near future. Requests for reprints should be addressed to the respective laboratories (see the code letter) or to Philips Research Laboratories, Eindhoven, Netherlands.

- P. Beekenkamp:** Colour centres in borate, phosphate and borophosphate glasses. Thesis, Eindhoven, 1965. *E*
- H. J. Benseck:** YIG-tuned reflector for 3 cm waves. Proc. IEEE **54**, 2003-2004, 1966 (No. 12). *H*
- M. Berth, F. Desvignes & R. Petit:** Tube analyseur d'images pour l'infrarouge. Onde électr. **46**, 1321-1329, 1966 (No. 477). *L*
- G. Blasse & A. Bril:** On the Eu^{3+} fluorescence in mixed metal oxides, V. The Eu^{3+} fluorescence in the rocksalt lattice. J. chem. Phys. **45**, 3327-3332, 1966 (No. 9). *E*
- A. J. Burggraaf:** The mechanical strength of alkali-aluminosilicate glasses after ion exchange. Thesis, Eindhoven, 1965.
- A. Claassen & L. Bastings:** The determination of nickel with dimethylglyoxime in iron and steel containing cobalt and copper. Analyst **91**, 725-731, 1966 (No. 1088). *E*
- B. J. Curtis & H. Graffenberger:** The floating zone crystal growth of lanthanum hexaboride. Mat. Res. Bull. **1**, 27-31, 1966 (No. 1). *A*
- H. J. van Daal & A. J. Bosman:** Influence of native defects on transport properties of Li-doped NiO. Physics Letters **23**, 525-526, 1966 (No. 9). *E*
- J. Dieleman, J. W. de Jong & T. Meijer:** Acceptor action of alkali metals in II-VI compounds as detected by electron spin resonance techniques. J. chem. Phys. **45**, 3178-3184, 1966 (No. 9). *E*
- W. F. Druyvesteyn, C. A. A. J. Greebe & A. J. Smets:** Helicon resonances in metal boxes filled with a ferromagnetic material. Solid State Comm. **4**, 643-644, 1966 (No. 12). *E*
- R. F. Hall:** Field emission characteristics and surface charging of Au-Cs-O. Proc. 7th Int. Conf. on phenomena in ionized gases, Beograd 1965, Vol. I, p. 229-232; Gradevinska Knjiga Publ. House, Beograd 1966. *M*
- M. Jung:** Pseudo-abrupte legierte *p-n*-Übergänge in GaP. Phys. Stat. sol. **18**, 743-748, 1966 (No. 2). *A*
- J. E. Knowles & K. Tweedale:** The origin of delta noise in square-loop ferrites. IEEE Trans. on magnetics **MAG-2**, 593-597, 1966 (No. 3). *M*
- J. Liebertz:** Der Einfluß von Lösungsgenossen auf die Einstellung des Gleichgewichts $\text{InOOH}/\text{In}_2\text{O}_3$ unter hydrothermalen Bedingungen. Berichte Bunsenges. phys. Chemie **70**, 1051-1052, 1966 (No. 9/10). *A*
- M. Loty:** Facteurs de qualité des tubes à rayons cathodiques à réponse rapide. Onde électr. **46**, 1314-1316, 1966 (No. 477). *L*
- A. G. J. van Oostrom:** Validity of the Fowler-Nordheim model for field electron emission. Thesis, Amsterdam, 1965. *E*
- D. R. Tilley, J. P. Baldwin & G. Robinson:** Parallel critical fields of type I tin films. Proc. Phys. Soc. **89**, 645-659, 1966 (No. 3). *M*
- B. Tuck:** Photoluminescence of $\text{GaAs}_{0.7}\text{P}_{0.3}$. Phys. Stat. sol. **18**, 541-545, 1966 (No. 2). *M*
- F. W. de Vrijer, A. L. Tan & A. G. van Doorn:** Advanced techniques for "Plumbicon" cameras. J. SMPTE **75**, 1080-1082, 1966 (No. 11). *E*
- J. S. C. Wessels:** Isolation of a chloroplast fragment fraction with NADP^+ -photoreducing activity dependent on plastocyanin and independent of cytochrome *f*. Biochim. biophys. Acta **126**, 581-583, 1966 (No. 3). *E*

Frequency-agile radar

N. Backmark, J. E. V. Krim and F. Sellberg

Radar systems are used for detecting the presence of certain objects — usually ships or aircraft — and determining their position. These observations are beset by various inaccuracies because of the complexity of the surfaces of these targets, whose attitude relative to the radar aerial considerably affects the strength of the received echo signal. These inaccuracies can be eliminated by making the frequency of the radar transmitter jump in an irregular way from one pulse to another ("frequency agility"). By using the Philips spin-tuned magnetron a light and compact frequency-agile radar has been achieved.

Introduction

The purpose of a radar system is to detect the presence of certain objects of interest — usually aircraft or ships — and then to determine their co-ordinates. Until recently it was the general practice to assign to each radar system its own fixed frequency. As we shall show in this article, however, there is much to be gained if instead the frequency of the radar transmitter is made to jump irregularly from pulse to pulse.

This method of operation has been chosen — even though it requires a more complicated equipment — on account of certain variations which can occur in radar echoes. These variations, which may occur even with targets that change their position relatively little, can both reduce the probability of detection and affect adversely the accuracy with which the angular co-ordinate is measured^[1]. Variations of this kind arise because structures such as aircraft and ships do not have simple shapes, but present to the radar beam a large number of reflecting surfaces of different size and at varying angles to the beam.

Owing to interference between the echoes from the different surfaces of the target the effective reflecting surface, and hence the amplitude of the echo reaching the radar receiver, depends critically on the attitude of the target. It was found in certain cases, for

example, that a rotation of the target by only $1/3^\circ$ caused a variation in signal amplitude of 15 dB. This effect, referred to as "scintillation", results therefore in a kind of fading, making a target which is otherwise within the range of the radar temporarily "invisible". Since with a radar system it is generally desirable to keep the target under continuous observation, the effect can obviously be troublesome.

Another phenomenon, "glint", is due to interference between the various component echoes from a target and gives an effect in which the apparent direction of the received signal depends on the attitude of the target. We know from practical experience with radars that the point from which the echo seems to be coming need not even lie within the "optical" confines of the target.

This point from which the echo seems to be coming is known as the scattering centre of the target. When a target is continuously illuminated at constant frequency by a radar system, the changing attitude of the target will cause the centre of the scattering cross-section to wander over its surface. As the attitude of the target only changes slowly, so too does the scattering centre. This slow displacement can easily be followed by automatic tracking radars, such as those used in fire con-

Civilingenjör N. Backmark and Teknologie Licentiat F. Sellberg are with Svenska Philips A.B., Stockholm; Ir. J. E. V. Krim is with N.V. Hollandse Signaalapparaten, Hengelo, Netherlands.

^[1] See: J. I. Marcum, A statistical theory of target detection by pulsed radar, IRE Trans. on information theory IT-6, 59-267, 1960, and P. Swerling, Probability of detection for fluctuating targets, IRE Trans. on information theory IT-6, 269-308, 1960.

control systems on warships, so that an apparent movement of the target is observed. Although it is not of first importance which part of the target is followed, the drift of the scattering centre over the target surface can cause serious errors in the measurement of target speed. The drift speed, which is added vectorially to the true speed of the target, causes inaccuracy in the prediction of target location, and in fire control, this prediction is of course the very thing that has to be correct.

As both of the phenomena described — “scintillation” and “glint” — result from interference between component echoes, they are very strongly affected by the radar frequency. One may therefore expect that a particular target will return a strong echo at one frequency and a weak one at another, even if its attitude does not change. Moreover, the apparent direction of the echo can deviate one way at one frequency, and the opposite way at another.

Since the radar beam has a certain width, a number of pulses — say 20 — will arrive at the target every time the beam sweeps over it. If the frequency of the transmitter is now made different at every pulse, then in the first place we may expect that interference between the echoes will not result in cancellation of the signal at all the frequencies. In the second place we may expect that the apparent directions of the echoes from the different pulses will be statistically distributed about the true bearing. In fact, with the bearing, the result of the frequency changes is that the scattering centre shifts to and fro at great speed over the target surface. An automatic tracking system is *not* able to follow such *fast* displacements, but takes the mean observation over a period depending on the time constant of the tracking system. The use of frequency agility thus increases the probability of detection and also improves the accuracy of direction finding and the prediction of target location.

Another advantage of using frequency agility is that the chance of jamming or interference from other transmitters is reduced. In an area where there are several radar stations with fixed frequencies it is necessary to allocate these frequencies carefully so as to avoid inadvertent interference between them. If, however, the transmitter frequency jumps irregularly from pulse to pulse, there is very little chance that the receiver frequency, which is tied to that of the transmitter, will in consecutive receiving intervals be equal to, or close to, the frequency of a pulse received from some other transmitter. This holds for both accidental and deliberate interference.

Finally, the use of frequency agility does away with the difficulty of “second-trace” echoes. These are echoes reflected from targets so far distant that the echoes reach the receiver only after the next pulse has been transmitted; they combine with the echoes of the

second pulse reflected from nearby targets, producing an incorrect and confusing indication on the radar screen. Clearly, if transmitter and receiver jump to an entirely different frequency after a pulse has been sent out, the receiver will be insensitive to these “second-trace” echoes.

These considerations, which have been confirmed by experiment [2], prompted Philips to design a frequency-agile radar system, some particulars of which are given in this article. The main problems were in the generation of the transmitter frequency and the design of the receiver.

Generation of the transmitter frequency

In microwave radar systems magnetrons are very often used in the transmitter as the source of high-frequency energy, particularly when the equipment has to be transportable and thus as simple and light as possible. This is because a magnetron possesses a unique combination of features: though small and light it has high output power and high efficiency, it is very sturdy in construction and it generates the output power at a sufficiently constant frequency to make external frequency control superfluous.

The frequency at which a magnetron oscillates is chiefly determined by the resonant frequency of its cavities. A simple magnetron has no elements which can be used to tune it and therefore oscillates at one unalterable frequency. Nowadays, however, owing to the need to be able to avoid jamming or unintentional interference from other radar transmitters by changing the operating frequency, nearly all radar transmitters have tunable magnetrons. The means of changing the magnetron frequency generally consists in the insertion of adjustable rods into the resonant cavities. To maintain a high efficiency there has to be a rod in each cavity and all rods must be moved simultaneously and by the same amount. This is done by means of a screw mechanism and a gear transmission system. Although this method of tuning is very effective and enables a wide range of frequencies to be covered, it is far too slow for a frequency-agile radar. With this method a complete tuning cycle takes a few seconds, whereas in order to be able to change the operating frequency to the required extent between two successive pulses a tuning speed about a thousand times greater is needed.

This explains why up till now magnetrons have not been used in frequency-agile radars. One solution of the problem that has been found is to employ a frequency control system comprising a number of crystal oscillators, frequency dividers and fre-

[2] See: B. G. Gustafson and B.-O. Ås, System properties of jumping-frequency radars, Philips Telecomm. Rev. 25, 70-76, 1964.

quency multipliers, followed by amplifiers and terminated by a high-power aperiodic output stage. The design of a light and compact frequency-agile radar unit was however made possible by the use of the *spin-tuned magnetron* developed for this work in the Valve Laboratory of Svenska Philips A.B., Stockholm.

The spin-tuned magnetron

This magnetron, whose construction is shown in simplified form in *fig. 1*, contains a rotating element. This consists of a disc which is situated above the anode block and whose axis coincides with that of the magnetron.

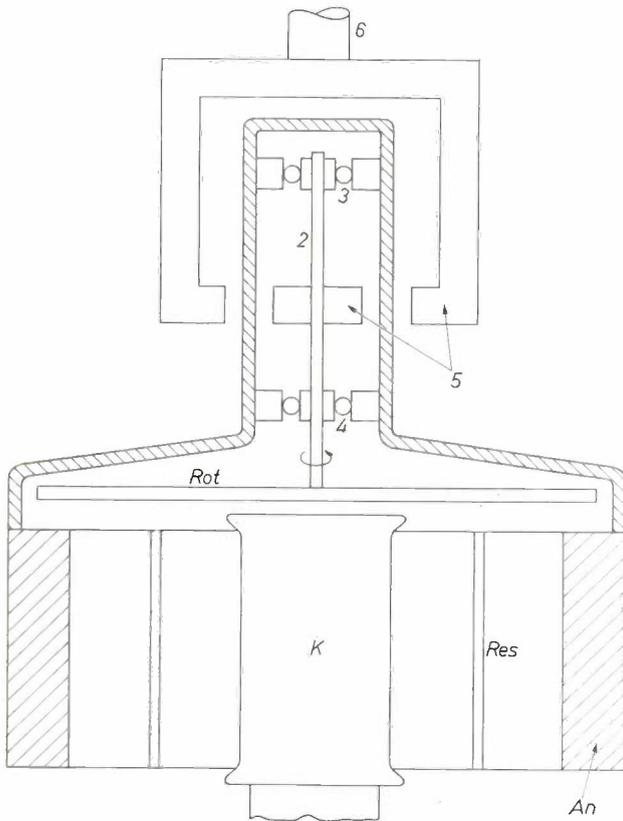


Fig. 1. Simplified cross-section and horizontal projection of the spin-tuned magnetron. *K* cathode, *An* anode block and *Res* one of the resonant cavities of the magnetron. *Rot* rotating disc with teeth *1*, driven by shaft *2*, mounted in ball-bearings *3*, *4* and driven by a magnetic coupling *5*, whose external yoke is connected to the shaft *6* of an electric motor.

The disc, which spins above the resonant cavities, has a toothed edge, and there are as many teeth as there are resonant cavities. The shaft of the disc runs on ball-bearings, and is driven by means of a rotating magnetized yoke which is situated outside the vacuum and connected to the shaft of an electric motor.

When a tooth of the disc moves over a cavity, it causes a change in the resonant frequency. The two effects are additive, resulting in a wide frequency swing. Since the teeth alter the tuning of all the resonant cavities identically, optimum efficiency is obtained. In every revolution of the disc as many tuning cycles are completed as there are cavities in the anode block, and

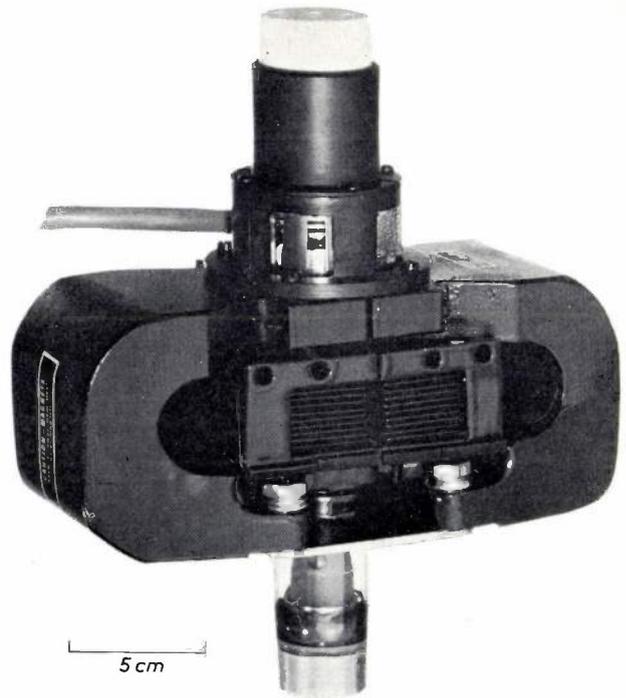


Fig. 2. Spin-tuned magnetron type YJ 1181. The motor which drives the tuning disc is located on the upper side.

therefore the tuning speed is proportional to the number of cavities as well as to the speed of revolution of the disc.

In one type of spin-tuned magnetron, type YJ 1181 (see *fig. 2*), which operates in the frequency band 8.5 to 9.6 GHz and has 16 resonant cavities, the frequency varies almost sinusoidally with time and, at a speed of 4500 revolutions per minute, one tuning cycle is completed in 900 microseconds. The frequency swing is 450 MHz and the maximum tuning speed is 1.4 MHz per microsecond. No simple relation will exist between the speed of rotation of the disc and the pulse repetition frequency — especially if the latter is made subject to small random fluctuations — and therefore the transmitter frequency jumps in an irregular manner

from one pulse to another (see *fig. 3*).

As there are no reciprocating movements, no vacuum-tight bellows are required. This means that the life of the tuning mechanism is good and does not limit the life of the magnetron as a whole. Moreover, the simplicity of the construction guarantees a high degree of reliability.

The power required to rotate the tuning disc is very small: only 18 W for the YJ 1181 magnetron. If it is required to transmit a fixed frequency, the disc can be held at any desired position by means of an electro-mechanical device.

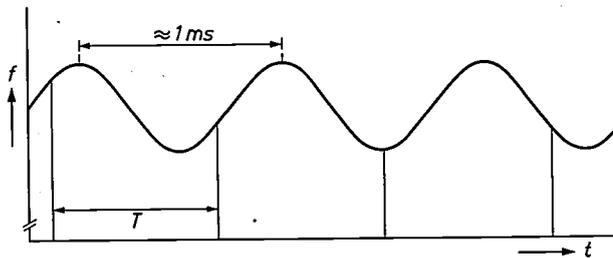


Fig. 3. Rotation of the tuning disc causes the magnetron frequency f to vary with time in a regular pattern which is approximately sinusoidal. Each magnetron pulse is so short that it appears on the time scale as a point. There is no simple relation between the number of revolutions per second of the tuning disc and the pulse repetition frequency (the pulse interval T is not necessarily constant from pulse to pulse) and therefore the transmitter frequency jumps in an irregular way from pulse to pulse.

The receiver

In the receiver the incoming echo signals, whose frequency is determined by the transmitter, are mixed with a local-oscillator signal and the resultant difference-frequency signal is amplified in an i.f. amplifier before being detected. For frequency-agile operation it is therefore necessary to solve the problem of how to keep the difference in frequency between the transmitter and the local oscillator exactly at the value to which the i.f. amplifier is tuned.

This requires that the local oscillator should be tunable over the same range of frequencies and that it should be capable of being tuned at the same speed as the magnetron. In fact, as we shall show later, it requires to be tuned at a much higher speed than the magnetron. Furthermore the output power of the oscillator should be high enough over the whole frequency band to guarantee a satisfactory i.f. signal. These requirements can be satisfied by a backward-wave oscillator (BWO), which has two advantages for such an application. Firstly, owing to its non-resonant character, it can deliver sufficient power over a very wide frequency range; and secondly, the frequency control is entirely electronic, so that there are no speed restrictions of mechanical origin.

To permit the reception of echoes from targets very close to the transmitter, the local oscillator must already be set to the correct frequency at the moment the pulse transmission has ended. So as to be able to detect long-range targets as well, this frequency setting should remain fixed, preferably until just before the moment at which the next pulse is generated. This means, therefore, that only a fraction of the time between two successive transmitted pulses is available for setting the local-oscillator frequency.

Since such fast and accurate control of the local oscillator frequency could not be achieved with conventional automatic frequency-control circuits, an entirely new control system was designed. The principle, which we shall now describe, is illustrated by the block diagram in *fig. 4*.

The output signal of the local oscillator *LO* is fed to the mixer stage of the receiver *R* and also to a circulator *Circ*; the signal emerging from the next port of the circulator is applied to the output of the magnetron and thence injected into the magnetron resonant structure.

At the beginning of each control period (this is equal to the time interval between two successive pulses) the resonant frequency of the magnetron will differ from the oscillator frequency, so that the local-oscillator signal is reflected back to the circulator. It then continues through the circulator, following the same path as the high-power pulse which appears at a certain instant from the magnetron, and arrives in the transmitter aerial waveguide. Incorporated in this waveguide is an automatic, instantly acting switch that conducts high-power signals, i.e. the outgoing pulse, to the aerial, and

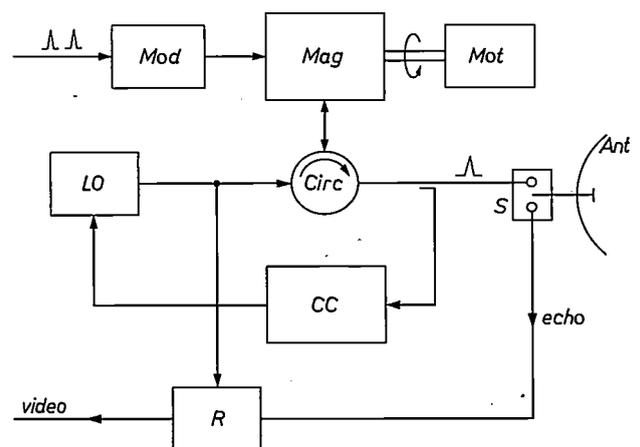


Fig. 4. Simplified block diagram of the frequency control circuit. *Mag* spin-tuned magnetron, tuned by electric motor *Mot*. The magnetron is driven by modulator *Mod*. *LO* local oscillator (backward-wave oscillator). *Circ* circulator. *S* transmit-receive switch (duplexer). *Ant* transmitting-receiving aerial. *CC* control circuit. *R* receiver. The control circuit controls the local oscillator in such a way that the oscillator frequency during each receiving interval differs from the frequency at which the last radar pulse was transmitted by an amount equal to the (fixed) intermediate frequency *IF*.

low-power signals, i.e. the reflected BWO signal, to the control circuit *CC*, where they are detected. During the presence of the d.c. voltage (obtained by detection) the control circuit delivers a signal that causes the local-oscillator frequency to drop rapidly along line *b* (see *fig. 5*) from the value *a* it had at the beginning of the control period. The value *a* is greater than the maximum tuning frequency of the magnetron: the magnetron frequency follows the curve *e*.

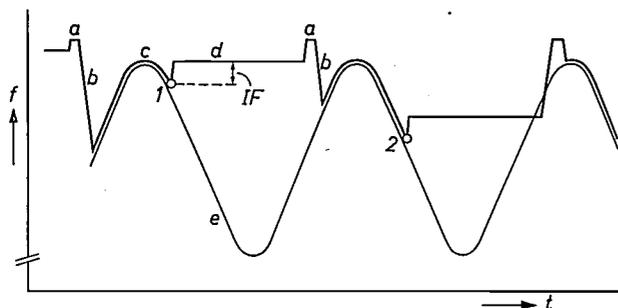


Fig. 5. Curve of local-oscillator frequency (*a-b-c-d*) and of the variable magnetron frequency (curve *e*). The instants of time *1* and *2* mark the magnetron pulses. During the receiving interval (*d*) which follows *1*, *2*, etc., the local-oscillator frequency remains constant and higher by a fixed amount *IF* than the frequency of the transmitted pulse. For clarity the length of the control period *a-b-c* is somewhat exaggerated here in relation to the receiving interval, and *c* and *e* are separated.

At the instant that line *b* meets curve *e*, in other words as soon as the frequency of the local oscillator becomes equal to the tuned frequency of the magnetron structure, the latter absorbs part of the oscillator energy. There is then very little reflection and the detected signal drops abruptly to almost zero. At this abrupt drop in the detector signal the control circuit stops reducing the oscillator frequency.

After switching off the "search" along line *b*, the next step is to switch to "following" the magnetron frequency, as after termination of the search the magnetron frequency continues to follow curve *e*.

At the same time as the search is switched off an auxiliary oscillator is therefore switched on, which is used to modulate the frequency of the local oscillator. Since this frequency is very close to that of the magnetron, the effect of modulating the oscillator frequency is a perceptible amplitude modulation of the residual reflected and detected signal. Moreover, every time the oscillator frequency passes the magnetron frequency there is a phase shift in the modulation of the detected signal. By comparing the phase of the modulation component of the detected signal with that of a signal derived straight from the auxiliary oscillator it can be established whether the oscillator frequency is higher or lower than the tuned frequency of the magnetron. A control signal can thus be derived which adjusts the oscillator frequency in the right direction. With this

method the oscillator frequency is kept close to the magnetron frequency and therefore follows it along curve *c* which practically coincides with curve *e*.

At a given moment, determined by the modulator (*Mod* in *fig. 4*), the magnetron begins to oscillate, e.g. at the moment its tuned frequency passes the value at point *1* in *fig. 5*. The control circuit, being linked to the transmitter waveguide, reacts to this oscillation by altering the value which the voltage controlling the local oscillator has at that moment. The magnitude of this step in voltage is arranged to make the local-oscillator frequency differ from the magnetron oscillating frequency by an amount equal to the intermediate frequency (*IF*) of the receiver. The control circuit then freezes the control voltage, and hence the local-oscillator frequency remains at this changed value (*d*) for a period (the "receiving interval") which is somewhat shorter than the pulse interval of the modulator. At the end of the receiving interval the control circuit returns the oscillator frequency to the value *a* and a moment later switches on the search *b* again, thus starting the next control cycle.

As with this method of adjusting the local-oscillator frequency there is still a chance of slight errors at the moment the magnetron starts oscillating, the i.f. section of the receiver contains a modified conventional AFC circuit for fine adjustment of the local-oscillator frequency if it differs from the correct value. This adjustment takes place while the pulse is being transmitted.

Summarizing, the following phases may be distinguished in the tuning cycle:

- 1) Searching for the instantaneous tuned frequency of the magnetron along line *b*.
- 2) Following the changing magnetron frequency along curve *c*.
- 3) Stopping after reception of the transmitter pulse, immediately followed by a controlled frequency step *IF*.
- 4) Stopping at the frequency attained for the duration of the receiving interval, line *d*.
- 5) Return to the starting frequency *a*.

The control circuit has to ensure that the BWO completes the phases of this cycle correctly and in the right order. To enable it to do this, the control circuit receives information from various sources. Recapitulating, these are successively: the drop in the detected reflected high-frequency power, the output voltage of the phase-sensitive detector, the transmitter pulse and the output voltage of the frequency discriminator in the i.f. part of the receiver as a correction for the controlled step in BWO frequency.

The electronic functions for the return to the frequency *a*, the search, the step *IF* and for marking off the receiving interval are performed by appropriate circuits in the control system.

Future developments

The special features of the spin-tuned magnetron and of its associated control circuits make it suitable for other applications that lie outside the scope of conventional systems.

First, we may mention a system in which the transmitter frequency, instead of being changed in an arbitrary manner from one pulse to another, is varied to follow a predetermined programme.

Next, a frequency-agile transmitter can also be combined with a FRESCAN aerial^[3], i.e. an aerial where the direction of the principal axis of the beam is a function of the frequency. An aerial of this type can be used to scan a sector, without mechanically moving the aerial,

by electronically controlled variation of the frequency. Obviously, much higher scanning speeds are possible with this system, producing a much less cluttered picture on the screen and eliminating troublesome "tails".

Finally, the variable-frequency magnetron can be used in military systems for target analysis and gives good results in the determination of the optimum frequency for a reliable maximum return at any given moment.

^[3] See, for example: J. Croney, Doubly dispersive frequency scanning antenna (for two plane scanning), *Microwave J.* 6, No. 7, 76-80, 1963.

Summary. The article discusses certain inaccuracies that can occur in radar detection and tracking and explains how these can be eliminated by making the frequency of the radar transmitter jump in an irregular way from pulse to pulse. This procedure also offers protection against jamming or unintentional interference from other radar transmitters. The design of a fre-

quency-agile radar system was greatly facilitated by the development of the Philips spin-tuned magnetron, which is briefly described. The article then discusses the principles of a control circuit which, during each receiving interval, keeps the frequency difference between the transmitter and the local oscillator in the receiver equal to the intermediate frequency of the receiver.

The APT programming language for the numerical control of machine tools

J. Vlietstra

In the last few years there have been considerable advances in the numerical control of machine tools. However, on account of the large quantities of workpiece data which have to be fed to the control unit, efforts were soon made to bring in digital computers to handle the data. This article deals with a programming language in which the data can be coded in a simple manner. It also describes the way in which the data are automatically converted into the form required for a numerically controlled machine tool.

The numerical control of machine tools such as milling machines, jig boring machines, lathes, drawing machines, etc., was introduced about ten years ago. Initial difficulties were mainly mechanical and electronic, but these were overcome fairly quickly, and machine tools are now available which can be controlled reliably by a "control unit" [1]. This article will deal with a fresh problem that has arisen here: communication between the designer of the workpiece and the control unit. How can he most efficiently give to the unit the data for the item to be manufactured? Before going more deeply into this, we shall first briefly describe the data required by the unit and the form in which this data can be supplied.

First of all, information is given about the path which the tool (e.g. a milling cutter) must follow through the material to be machined. This path is generally described by a series of discrete points. The control unit guides the tool in a straight line from one predetermined point to the next. The tool will therefore not be able to follow a curve such as an arc of a circle perfectly, but it will be able to approach it as closely as desired provided that sufficient points are given. Furthermore, the control unit is so designed that it can itself calculate a number of points between the predetermined ones by linear or quadratic interpolation. Nevertheless, if a complicated path is to be followed a relatively large number of points have to be predetermined.

The control unit is also told what kind of tool is to be used, the speed at which the tool is to rotate, the cutting speed, etc. This information is referred to as "tool function" or "auxiliary function" data.

The path of the tool can thus be predetermined by a series of numbers, which are the co-ordinates of the successive locations of the tool. The data for the tool

functions can also be coded in the form of combinations of letters and figures. These are punched on paper tape, the "control tape", which is then fed into the control unit.

It is often necessary to make a large number of calculations to find the successive tool positions. Even though the actual calculations may be simple, it is clear that the use of a digital computer will considerably reduce the labour involved. Next, all the data must be recorded correctly and in the proper order on the punched tape: this operation is a source of error that should by no means be underestimated. Because of this difficulty, work was started some time ago on programming the computer itself to punch the control tape, without human intervention, from a few, simply coded data relating to the workpiece to be machined. In this work, every effort was made to produce a coding system that could be easily dealt with by the designer of the workpiece. Such a standardized notation or coding system is in general termed a *programming language*.

We have not yet dealt with all the implications of these problems. There are now many types of computer, and the variety of machine tool types is quite bewildering. Nevertheless, it is highly desirable to work towards a programming language that is *universal*; one that can be used for any computer that might be employed, and can provide control tapes for all kinds of machine tools.

This was recognized quite early in the day by a number of companies which were members of the Aerospace Industries Association in the United States.

[1] For the principles of numerical control see for example: T. J. Viersma, Some considerations on the numerical control of machine tools, Philips tech. Rev. 24, 171-179, 1962/63; J. A. Haringx, R. Ch. van Ommering, G. C. M. Schoemaker and T. J. Viersma, A numerically controlled contour milling machine, Philips tech. Rev. 24, 299-331, 1962/63.

In 1957 this group took over the work that had been begun at the Massachusetts Institute of Technology, and developed from this work the Automatically Programmed Tools programming language. Many other American companies entered into co-operation with the original group, and in 1961 the further development of the APT language and appropriate programmes for various computers was entrusted to the Research Institute of the Illinois Institute of Technology.

APT thus gradually became the most widely used programming language for numerical control. In 1964 European firms were also invited to co-operate, and Philips joined the organization a year later.

In this article we shall try to show clearly, with the aid of an example, the opportunities offered by the APT system to the designer and to the "part programmer", whose function is the coding of the data. A part programme written in APT language consists of a number of definitions of geometrical entities followed by a series of motion instructions. The appropriate APT programme processes these definitions and instructions, tracks down errors wherever possible and finally punches a control tape. If desired, it can also produce a list of all the quantities, both defined and calculated. This is an excellent aid in tracing errors in the part programme. An APT computer programme has also been found useful in designing a workpiece. The APT language is, in fact, very flexible, so that a large number of different drawings or models of a workpiece can be made by slight modifications of a few motion instructions or parameters at each repeat.

A part programme in APT language

We shall now show how a part programme is written for milling the cam shown in fig. 1. The cam is to be milled from a plate of metal or plastic. All the dimensions and the symbolic names of the geometric elements which appear are given in fig. 1. The names will be explained later. This workpiece is referred to as two-dimensional, since the centre of the cutter always moves in the same plane throughout the milling process. In our example the axis of the cutter remains perpendicular to this "base plane". To clarify this, we show in fig. 2 a section through the workpiece and through the body of a cylindrical flat-ended cutter. The cutter rotates about the axis AB, and M indicates the cutter centre, i.e. the point of intersection of the axis and the base of the cutter.

The part programme for punching the control tape is shown in fig. 3. It begins with some general data: the part programmer gives the number of the workpiece, describes the shape and dimensions of the cutter to be used and the tolerances on the dimensions of the work-

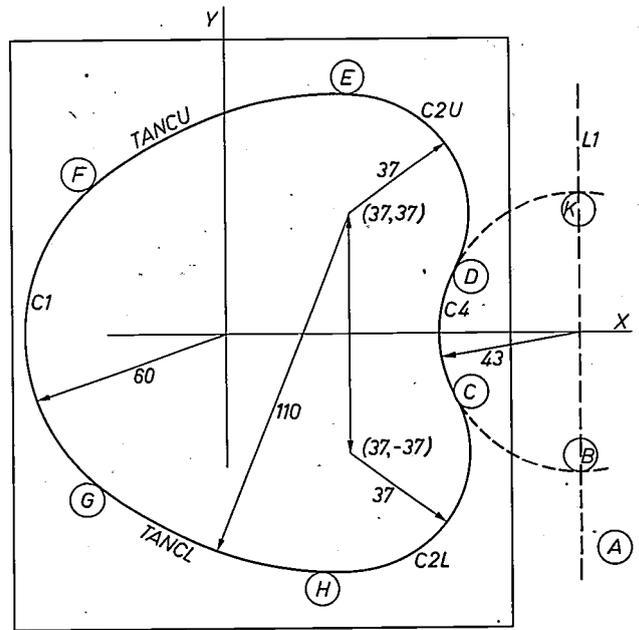


Fig. 1. Working drawing of a cam. The geometrical elements of the circumference are indicated by symbolic names, which are defined in the part programme. The letters in small circles represent successive positions of the cutter along its path.

piece, and specifies the lists he wants as secondary output, etc. Then follow the definitions of the geometrical elements making up the shape of the cam. Finally we have the motion instructions. Here the programmer describes the successive displacements of the cutter and indicates the desired tool functions.

We shall now discuss the part programme in detail where necessary. (In the programme all the definitions and instructions are numbered.)

a) General data.

PARTNO NOKSCHYF TEK. NR. 2AZ85636

Every part programme must begin with the instruction PARTNO followed by the description of the workpiece and its number. (NOKSCHYF TEK. NR. means cam drawing no.)

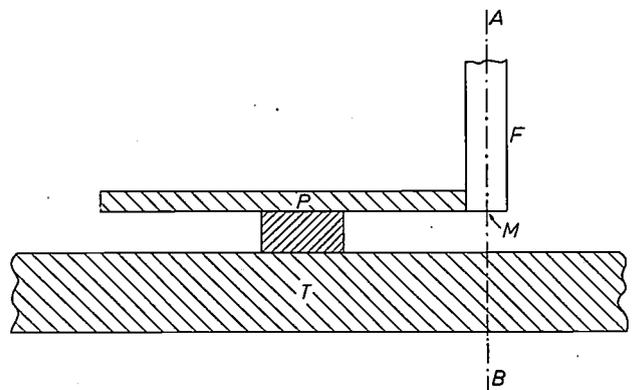


Fig. 2. Section through a cylindrical cutter F with a flat end, and through the plate P to be machined clamped to the work-table T, during the milling of the cam shown in fig. 1. The chain-dotted line AB is the axis of the cutter. The point M at which the axis intersects with the base of the cutter is called the cutter centre.

PARTNO	NOKSCHYF TEK, NR, 2AZB5636	00
	MACHIN/COBRA	10
	CLPRNT	20
	CUTTER/4	30
	INTOL/.01	40
	OUTTOL/.01	50
	C1=CIRCLE/0,0,60	60
	C2U=CIRCLE/37,37,37	70
	C2L=CIRCLE/37,-37,37	80
	TANCU=CIRCLE/YSMALL, IN, C1, IN, C2U, RADIUS, 110	90
	TANCL=CIRCLE/YLARGE, IN, C1, IN, C2L, RADIUS, 110	100
	C4=CIRCLE/XLARGE, OUT, C2U, OUT, C2L, RADIUS, 43	110
	PL1=PLANE/0,0,1,0	120
	L1=LINE/(POINT/CENTER, C4), ATANGL, 90	130
	SPINDL/ON	140
	FEDRAT/100	150
	FROM/120,-70,0	160
	GO/PAST, C4, TO, PL1, ON, L1	170
	GOLFT/C4, TANTO, C2U	180
	GDFWD/C2U, TANTO, TANCU	190
	GDFWD/TANCU, TANTO, C1	200
	GDFWD/C1, TANTO, TANCL	210
	GDFWD/TANCL, TANTO, C2L	220
	GDFWD/C2L, TANTO, C4	230
	GDFWD/C4, ON, L1	240
	RAPID	250
	GOTO/120,-70,0	260
	END	270
	FINI	280

Fig. 3. Part programme for machining the cam shown in fig. 1.

MACHIN/COBRA

The control tape must be suitable for the COBRA-controlled drawing machine [2] which, as well as making drawings and engravings also performs milling operations on certain types of plastics (methacrylates). The code COBRA indicates that the post-processor known by the symbolic name COBRA must be used. This post-processor will be discussed later.

CLPRNT

A list of all calculated locations of the cutter centre is requested (Cutter Locations Print).

CUTTER/4

A description of the shape of the cutter. In this case the simplest shape is described: a flat-ended cylindrical cutter with a diameter of 4 mm.

INTOL/.01

"Internal" tolerance. The cutter may not go more than 0.01 mm inside the desired circumference of the cam.

OUTTOL/.01

"External" tolerance. The distance from the cutter to the desired circumference may not exceed 0.01 mm.

b) Definitions of geometrical elements.

C1 = CIRCLE/0, 0, 60

Definition of a circle with the centre at (0, 0) and a radius of 60 mm. This circle is given the symbolic name C1. In the rest of the programme this circle is referred to by this name alone.

C2U = CIRCLE/37, 37, 37

Definition of circle C2U with the centre at (37, 37) and a radius of 37 mm.

C2L = CIRCLE/37, -37, 37

A definition similar to that of C2U.

TANCU = CIRCLE/YSMALL, IN, C1, IN, C2U, RADIUS, 110

Definition of the circle TANCU. It has a radius of 110 mm and touches circles C1 and C2U internally. Two circles satisfy these requirements: the required circle is the one whose centre has the smaller Y-coordinate (YSMALL).

TANCL = CIRCLE/YLARGE, IN, C1, IN, C2L, RADIUS, 110

Definition similar to that of TANCU.

C4 = CIRCLE/XLARGE, OUT, C2U, OUT, C2L, RADIUS, 43

Circle C4 has a radius of 43 mm and touches C2U and C2L externally. Two circles are possible: the one required is the one whose centre has the greater X-coordinate.

PL1 = PLANE/0, 0, 1, 0

Plane PL1 is the one described by the expression $0 \times X + 0 \times Y + 1 \times Z = 0$.

In this case PL1 is therefore the XY-plane.

L1 = LINE/(POINT/CENTER, C4), ATANGL, 90

Line L1 passes through the centre of C4 and makes an angle of 90° with the X-axis (ATANGL, 90).

c) Motion instructions.

SPINDL/ON

The cutter must begin to rotate. The number of revolutions per second does not have to be specified for the COBRA. This instruction remains in force until the next instruction about the speed of rotation.

[2] R. Ch. van Ommerring and G. C. M. Schoenaker, The COBRA, a small digital computer for numerical control of machine tools, Philips tech. Rev. 27, 285-297, 1966 (No. 11).

FEDRAT/100

The rate of displacement of the cutter, the feed rate, must, where possible, be 100 mm/s. This instruction remains in force until the next instruction about the feed.

FROM/120, -70, 0

Cutter movement starts with the centre of the cutter at the point $X = 120$, $Y = -70$, $Z = 0$ (position *A* in fig. 1).

GO/PAST, C4, TO, PL1, ON, L1

Take the cutter in a straight line from position *A* to *B*. Once the cutter has reached this position, it has passed circle *C4* and is now touching *C4* (both conditions are included in the code *PAST, C4*). The base of the cutter is touching plane *PL1* (coded as *TO, PL1*) and its centre now lies on line *L1*.

GOLFT/C4, TANTO, C2U

Move the cutter to the left along circle *C4* until it is tangent to circle *C2U* (position *D* in fig. 1).

GOFWD/C2U, TANTO, TANCU

Move the cutter further along *C2U* until it touches circle *TANCU* (position *E* in fig. 1).

The next motion instructions (200 to 230 in the programme of fig. 3) are now self-evident. Once the last of these instructions has been completed, the cutter is in position *C*. The programme then continues as follows.

GOFWD/C4, ON, L1

Move the cutter forward along *C4* until its centre is on line *L1* (position *K*).

RAPID

This instruction relates to cutter feed. From now on the cutter must travel at its maximum speed. Instruction 150 is cancelled.

GOTO/120, -70, 0

The cutter must now move back to its starting point.

END

Stop the rotation of the cutter.

FINI

The end of the part programme. All definitions and instructions must therefore be included between *PARTNO* and *FINI*.

Processing by the APT computer programme

By and large, an APT programme ensures that the computer performs the following processes:

- 1) The part programme is read in and the definitions and instructions are translated into another code. Meanwhile, the machine is able to trace a few of the syntactic and mathematical errors in the programme. An example of a syntactic error would be the occurrence in an instruction of a symbolic name not previously defined. Writing errors in geometric expressions like *CIRCLE*, *LINE*, *POINT*, etc.

are also regarded as syntactic errors. A mathematical error would be one like the definition of a point as the point of intersection of two parallel lines, or an ambiguous description of a circle. In this connection one need only think of the complicated requirements for the description of the circles *TANCU* and *TANCL*. Checks for this kind of error are by no means superfluous. Finally the machine calculates the standard or canonical form of the various geometrical elements. Thus, for each circle it calculates the co-ordinates of the centre and the radius from the conditions laid down for tangency, etc. The other parts of the computer programme can operate only with these standard forms:

- 2) The machine now calculates the successive positions of the cutter centre for the control unit of the machine tool. The factors taken into account here are the shape quoted for the tool, the tolerances specified, etc. The control unit will guide the cutter in a straight line from one position to the next. The closer the specified tolerances, the larger is the number of points to be calculated for the computer programme.
- 3) All these data relating to the cutter path must now be processed by another computer programme called the "post-processor". This part of the APT computer programme ensures that the data are converted into the required form for the machine tool to be used, in our case the *COBRA*-controlled drawing and engraving machine. All instructions relating to the tool functions are therefore passed on directly to the post-processor. An error check is once more carried out here. If, for example, the instruction *SPINDL/ON*, which starts the cutter rotation, has been omitted, a signal to this effect is given immediately.

In the development of APT, a distinction has always been drawn between the processes given in (1) and (2) which are of a linguistic and mathematical nature, and the processes in (3) which depend upon the machine tool to be controlled. This distinction is clearly visible in the APT computer programme. The sections dealing with the translation and calculation of the tool positions together form a complete programme. This, however, is written in such a way that the post-processors can easily be added to it. The advantages of this are obvious, and we think that it gives a particularly successful arrangement.

Several post-processors have been developed at the Philips laboratories in Eindhoven, one for the Giddings and Lewis jig boring machine, one for the Scharmann jig boring machine, one for the Milwaukee-Matic machine tools, one for a three-dimensional Numeric-

NOKSCHYF TEK, NR, 24285636				CARD NO,	00 TAPE NO,	2	
MACHIN/COBRA	=0			CARD NO,	10 TAPE NO,	4	
CUTTER/	4.0000			CARD NO,	30 TAPE NO,	6	
INTOLZ/	0.0100	0.0100	0.0100	CARD NO,	40 TAPE NO,	8	
OUTTOL/	0.0100	0.0100	0.0100	CARD NO,	50 TAPE NO,	10	
SPINDL /	ON			CARD NO,	140 TAPE NO,	12	
FEDRAT /	100,0000			CARD NO,	150 TAPE NO,	14	
FROM				CARD NO,	160 TAPE NO,	16	
	X	Y	Z				
	120,0000000	-70,0000000	0,0000000	CARD NO,	170 TAPE NO,	18	
DS IS/C4							
	X	Y	Z				
	107,9295425	-41,0080000	0,0000000	CARD NO,	180 TAPE NO,	20	
C4 (0) = CIRCLE/	107,9295	0,0000	0,0000	43,0000	CARD NO,	180 TAPE NO,	21
DS IS/C4							
	X	Y	Z				
	105,4351204	-40,9340684	0,0000000				
	102,8950453	-40,6998027	0,0000000				
	100,3744483	-40,3080718	0,0000000				
	97,8830815	-39,7603914	0,0000000				
	95,4305839	-39,0588804	0,0000000				
	93,0264441	-38,2062528	0,0000000				
	90,6799634	-37,2098076	0,0000000				
	88,4002203	-36,0614153	0,0000000				

Fig. 4. First part of the list of cutter centre locations, calculated with the APT computer programme to make the control tape for the cam shown in fig. 1.

Keller milling machine and one for the COBRA-controlled drawing and engraving machine. The last one is referred to in our example by its symbolic name, COBRA. The five post-processors form part of the APT computer programme for the Control Data 3600 computer at the Philips Computer Centre.

Results

The procedure that has just been described provides the following results:

- 1) The verification list for the part programme. This list repeats all the definitions and instructions in the original part programme. Any errors are printed out immediately after the instructions in which they are found.
- 2) The list of all the calculated positions for the centre of the tool. This is an important aid to the programmer in tracing any errors that could not be discovered by the computer programme. It is in fact possible for the part programme to be syntactically correct, even though the result does not fully correspond to the programmer's intention. The beginning of this list for our example is shown in fig. 4.
- 3) The control tape. This can be fed directly to the machine tool. Often, however, it is first used to make a drawing, from which the path travelled by the tool can be examined and measured. A further preparatory

step that can be taken is to use the tape to make a wooden model to check the shape of the workpiece before any expensive material is machined. Fig. 5 shows a check drawing of the cam. This drawing, made by the COBRA-controlled drawing machine, shows the outline of the cam and the path of the cutter centre.

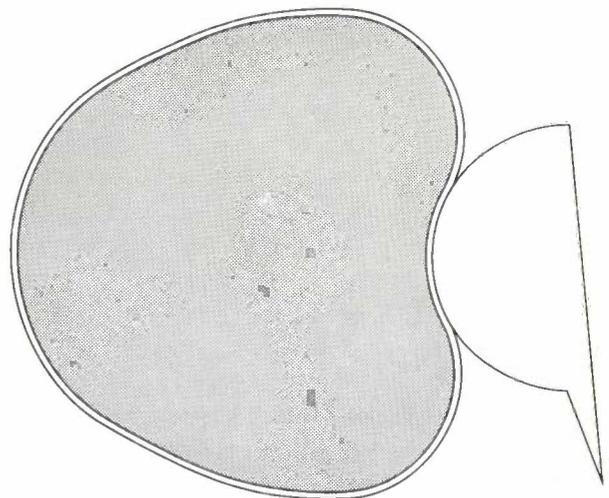


Fig. 5. Circumference of the cam and path of the cutter centre. This drawing was made by the COBRA-controlled drawing machine from the tape given by the part programme shown in fig. 3. A drawing like this may be used as a check on the control tape.

Three-dimensional workpieces

At this point we shall leave the example and assume that the whole process has now been made sufficiently clear for us to be able to describe in rather wider terms one or two methods of milling *three-dimensional* workpieces.

The simplest method is the one in which the cutter centre is made to move in a plane parallel to the base plane (the *XY*-plane), while the axis of the cutter does not change its direction. Each time that the cutter has completed a "circuit" of the workpiece, this plane is moved slightly in the *Z*-direction. This greatly simplifies the writing of the part programme. The programmer defines the cutter path as a function of the parameter *Z*. The computer programme ensures that the computer calculates the successive positions of the centre of the cutter for any desired value of *Z*. Such a part programme is often referred to as "two-and-a-half-dimensional", as it reduces a three-dimensional problem to a series of two-dimensional problems.

The APT system can also be used for real three-dimensional operation, with a three-, four- or even five-axis tool. The position and attitude of a five-axis cutter, for instance, is given by five figures: the *X*-, *Y*- and *Z*-co-ordinates of the cutter centre and two figures giving the direction of the axis. This facility is often used when the surface of the workpiece is a ruled surface. One such example is a truncated cone: through every point on the surface of the side there is a straight line which lies entirely in the surface. A five-axis cutter can be controlled in such a way that its entire length touches the side surface of the cone at any given moment.

In some milling machines the cutter position is fixed and the work-table moves with respect to the cutter. At first sight the APT language does not take this into account. The part programmer thinks in terms of a fixed work-table and a cutter moving with respect to the workpiece. The specific features of a tool of this type are only made apparent in a post-processor. This section of the APT computer programme can also be written in such a way that it converts the calculated data for the cutter path into data for control of the work-table. This method avoids unnecessary complications in the APT language.

Linear transformations in APT

The great flexibility of the APT language is also clearly shown in the ease with which changes in scale, translations, reflection and rotation can be effected. Let us illustrate this by a brief example. *Fig. 6* shows how four identical hooks should be milled from one plate. The part programmer starts by writing all the motion instructions required to mill hook *1*. Here he states, in

code, that this set of instructions must be considered as a "macro-instruction" and gives it a symbolic name, e.g. HOOK. Finally he writes the following instructions:

a) A rotation through 180° about the origin, followed by a call on the macro-instruction HOOK; this is coded in APT language as:

```
M1 = MATRIX/XYROT, 180
CALL/HOOK, MATRX = M1
```

b) A reflection in the *Y*-axis and a translation of -120 mm in the *X*-direction, followed once again by a call on the macro-instruction HOOK; these instructions, when converted into APT language, read:

```
M2 = MATRIX/MIRROR, ZXPLAN, T, -120, 0
CALL/HOOK, MATRX = M2
```

c) A reflection in the *X*-axis and a translation of -120 mm in the *X*-direction, followed by the same call as before.

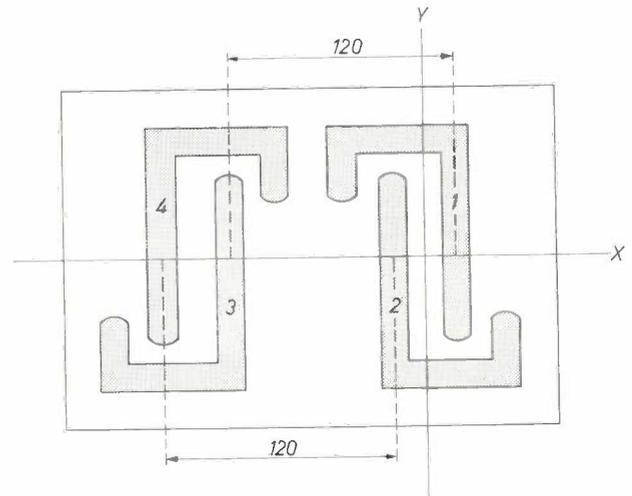


Fig. 6. Simplified working drawing for machining four identical hooks.

The APT computer programme will give the motion instructions for calculating the successive positions of the cutter centre for milling hook *1*. Next, the rotation given under (a) is performed, and this will provide all the data necessary for milling of hook *2*. The instructions given under (b) and (c) provide the same information for hooks *4* and *3* respectively. The part programmer need only concern himself with the motion instructions for hook *1*.

Future developments

Work is at present in progress on a new APT computer programme which will be radically different from existing programmes. We cannot go into this matter too deeply here, but we should nevertheless like to mention a few important advantages that the user might expect from the future APT computer programme.

First of all, attempts are being made to make it "machine-independent" as far as possible. The sections of the programme (subroutines) are, wherever possible, written in the widely-used FORTRAN programming language to make it as easy as possible for the user to adapt the programme to the computer that he has available.

Checking for syntactic errors will be intensified. In the new arrangement, the translation of APT instructions will be more rigidly separated from the calculating procedures than it is in current programmes. This will prevent the unnecessary loss of calculation time when syntactic errors are found in the part programme.

Facilities for extension will be taken into account to an even greater extent than to-day. The computer programme will be written in such a way that it will be a simple matter to add subroutines for dealing with new geometrical shapes. It will also be possible to include in the part programme subroutines which are written in FORTRAN, e.g. those for calculating parametrically defined surfaces [3]. It will then no longer be necessary for the part programmer to restrict himself to the APT language; he will also have all the facilities of FORTRAN at his disposal.

In addition, current thought is turning towards computer output devices which do not follow conventional design. APT computer programmes are at present written for the use of fast line printers, card punches and paper tape punches. There is however a pressing need for devices to make communication between man and machine easier, for example by displaying intermediate results of calculations on a cathode-ray tube in the form of legible text, graphs, diagrams, etc. Then again, the user must be enabled to pass instructions directly to the computer so that he can modify its operation on the basis of the data received. The develop-

ment of such equipment is still fairly new, but offers considerable promise.

Conclusion

We have used a few examples to show how the data for the manufacture of a workpiece can be passed to a computer by means of instructions coded in the APT language. Using the APT programme, the machine calculates from the instructions all the data required by a numerically controlled machine tool and prepares the control tape. This inclines us rather to replace the expression "numerical control" by "symbolic control", since it is not a series of numbers that is presented to a control unit, but a series of symbols to a computer.

The introduction of the APT system was an important stage on the road to the automation of machining processes. Hand punching of the control tape for some workpieces would be quite out of the question, and here the APT system has considerably widened the existing prospects. The system also makes the manufacture of simpler workpieces much easier. Once a perfect control tape has been prepared, we can repeat the production process as often as we like. Moreover, the recent standardization of the APT language allows work to be contracted out, since the method of manufacture is laid down in the part programme. The benefits which result from the development of the APT system are not therefore only experienced by the designers and part programmers; the system also helps to improve the efficiency of the production department.

Summary. This article gives a brief description of the APT programming language. It gives as an example the APT part programme for the manufacture of a cam, deals in detail with the significance of the geometrical definitions and motion instructions, and gives a general description of the way in which they are processed by an APT computer programme. Attention is also paid to checking for errors in the programme. The way in which three-dimensional workpieces are programmed is mentioned. The account of the simple way in which reflection, rotation, translation, etc., can be carried out gives an idea of the flexibility of APT. The article concludes with a brief outline of future prospects.

[3] The co-ordinates x_1 , x_2 and x_3 of the points on a parametrically defined surface are functions of two parameters u and v , i.e. $x_1 = x_1(u, v)$, $x_2 = x_2(u, v)$ and $x_3 = x_3(u, v)$.

A colour television camera with "Plumbicon" camera tubes

H. Breimer, W. Holm and S. L. Tan

The introduction of the "Plumbicon" television camera tube has brought new concepts to colour television camera design. Optical systems, mechanical arrangements and electronic circuits have all profited from these new ideas. Various aspects in the design of a Philips colour television camera and all its accessory equipment are discussed in some detail in the article below, and there is a brief account of practical experience with these cameras.

Introduction

The "Plumbicon" [*] camera tube has a number of features which allow the development of a colour television camera with these tubes to be approached in a different way from the design of a camera with image-orthicon tubes. This is true for the optical and also for the mechanical and electronic design. With the much smaller dimensions of the "Plumbicon" tube and its associated system of deflection and focusing coils, a new colour separation system can be used which is not only optically better than previous systems, but also, because of its small size, permits a completely new mechanical design to be used for the camera. The result is a colour camera which is no larger or heavier than many conventional black-and-white cameras. The main electrical features of the "Plumbicon" tube are a linear tube characteristic, which is little affected by temperature and voltage fluctuations, a signal current which is virtually free from noise and interfering components, and a black level which is nearly constant owing to the negligibly low dark current. In the following we shall see how these features have been taken into account in the design of a colour television installation [1] [2].

A studio or outside broadcast television installation almost always consists of two parts: the camera, mounted on a movable stand, and the control unit in the control room. These two parts are connected together by a cable containing a large number of single cores and several coaxial leads. For outside broadcasts, this cable may be hundreds of yards long. All electrical adjustments are kept together at the control unit where they are under the control of a technician who keeps watch on picture quality. The cameraman is only concerned with aiming and focusing the camera. Colour television installations are similarly divided. Obviously, as the electronic section of such installa-

tions is rather large every attempt is made to include as much of it as possible in the control unit.

Fig. 1 shows the camera with the "Plumbicon" tubes which we shall discuss here. The associated control unit, which will also be discussed briefly, is illustrated later on in the article (fig. 8).

Before discussing the camera, let us first briefly recapitulate the simultaneous colour television system on which all colour television studio installations are now based. The cone of light which enters the camera through the lens is distributed in three directions by a set of two colour-selective mirrors so that only a certain wavelength range of the visible spectrum passes in each direction. This gives rise to three cones of light containing the red, green and blue colour extracts from the received image. The mirror system does not affect the convergence of the rays of light to form a sharp image in the image plane, and sharp and geometrically congruent images in the primary colours red, green and blue are therefore produced on three identical camera tubes arranged in the cones of light.

The colour signals R , G and B derived from these tubes by the normal scanning process form the basis for the later reconstruction of the complete colour picture in the receiver. These signals set up three pictures in the primary colours red, green and blue in accurate geometrical register on the receiver screen, and the simultaneous presence of these three pictures gives the observer the impression of a picture in its natural colours (additive colour mixing).

[*] Registered trade mark for television camera tubes.

[1] The "Plumbicon" tube and its features are discussed in detail in E. F. de Haan, A. van der Drift and P. P. M. Schampers, Philips tech. Rev. 25, 133-151, 1963/64.

[2] A comparison between the features of the "Plumbicon" tube and those of other television camera tubes is to be found in A. G. van Doorn, Philips tech. Rev. 27, 1-14, 1966 (No. 1).

[3] The back-focus distance of a lens system is the distance between the image plane and the rearmost portion of the system, which is generally the holder of the final component lens.

The camera

The components of every colour television camera can be divided into three groups: the camera lens with the colour separation system; the pick-up section with the camera tubes and their deflection and focusing coils; and finally the electronic circuits. While the problems encountered in the first group are almost entirely optical ones, the second group requires at-

camera lens in a colour camera must be very long¹³⁾ so that the colour separation system can be inserted between the lens and the camera tubes. Such a system consists of the colour-selective mirrors mentioned above, which divert for example the red and blue cones of light, while the green is allowed to pass through directly to the "green" camera tube (*fig. 2*). The red and blue cones are reflected again by ordinary plane

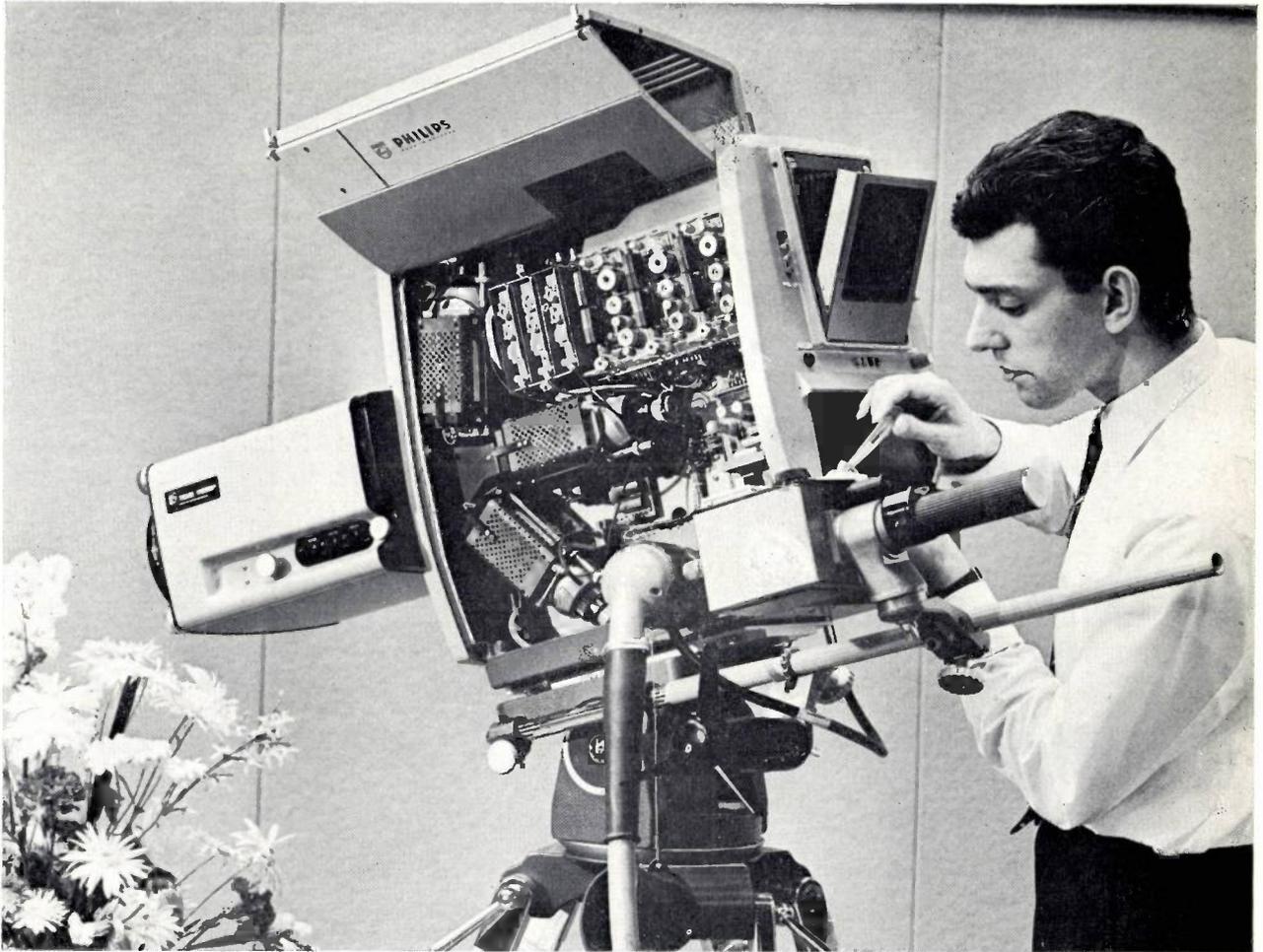


Fig. 1. The camera open but otherwise ready for use. The zoom lens in its housing, which also contains the servo drives, can be seen on the left. On its immediate right may be seen the small, encapsulated colour separation system, and the "Plumbicon" tubes, oriented in three different directions, with their deflection and focusing coil assemblies and preamplifiers. The remaining space in the camera is occupied by amplifiers and time-base circuits. The handle used for aiming the camera carries the servo controls for focusing and for choosing the focal length of the zoom lens.

tention to precision engineering, and the third group comes in the province of the electronic engineer. A few of the problems arising in these widely varying fields will now be dealt with.

The camera lens and the colour separation system

An essential difference between black-and-white and colour cameras is that the back-focus distance of the

mirrors so that they take up a direction parallel to the main optical axis before they arrive at the appropriate camera tubes.

The double reflection is important as it eliminates the lateral inversion which occurs in single specular reflection. This inversion could also be compensated by reversing the direction of scan in the camera tube. However, as the coil systems never have completely symmetrical deflection characteristics, there would be

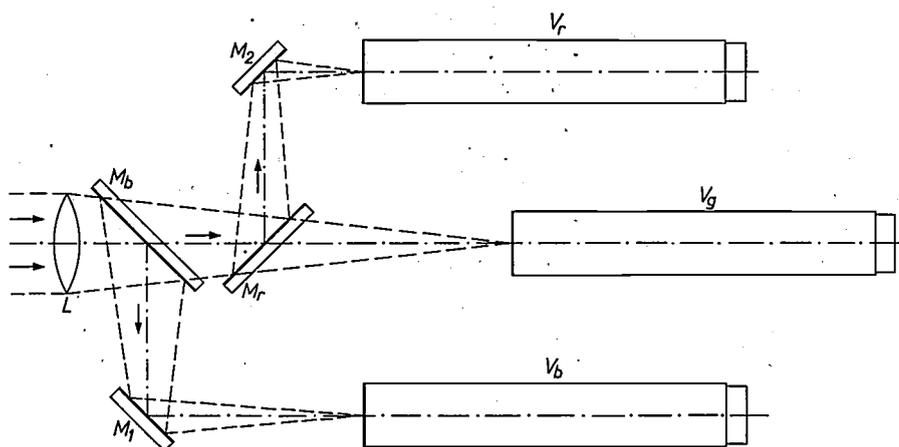


Fig. 2. Colour separation principle in television cameras. L camera lens, M_b and M_r colour-selective mirrors deflecting the blue and red components of the light successively, while the green component passes straight through to the camera tube V_g ; M_1 and M_2 ordinary plane mirrors to reflect the deflected blue and red components of the light again, so that camera tubes V_r and V_b can be arranged parallel to V_g . The second reflection also eliminates lateral inversion.

differences in geometry between the normally scanned green image and the red and blue images scanned in the opposite direction. It is virtually impossible to correct for these differences, and the three primary colour pictures on the receiver screen would therefore not be completely in register.

The parallel arrangement of the camera tubes also assists the accurate register of the three pictures. With this arrangement, the effects of the Earth's magnetism or any other interfering fields on the deflection system are much the same for all three tubes, so that distortions of the scanning geometry are virtually identical for all three pictures.

In the camera described here these interfering effects are reduced adequately in another way, described below.

A colour separation system like that shown in fig. 2 requires so much room between the camera lens and the tubes that lenses with unsuitably long focal lengths and rather small relative apertures would have

to be used. In image orthicon cameras, this difficulty is overcome by the addition of an optical system of long focal length (the relay system) which transfers the intermediate image obtained with a normal camera lens across the required large distance to the camera tubes (fig. 3a).

The small dimensions of the "Plumbicon" camera tube enabled us to develop a different method, described earlier in this journal [4], for our colour television camera. In this method the free space between the camera lens and the tubes is filled with glass at the location of the light beams. The convergence of the rays to the image plane is thus reduced and the effective back-focus distance of the lens increased. Numerically, this increase is roughly equal to the refractive index of the type of glass, $n = 1.52$ in the present case. The focal

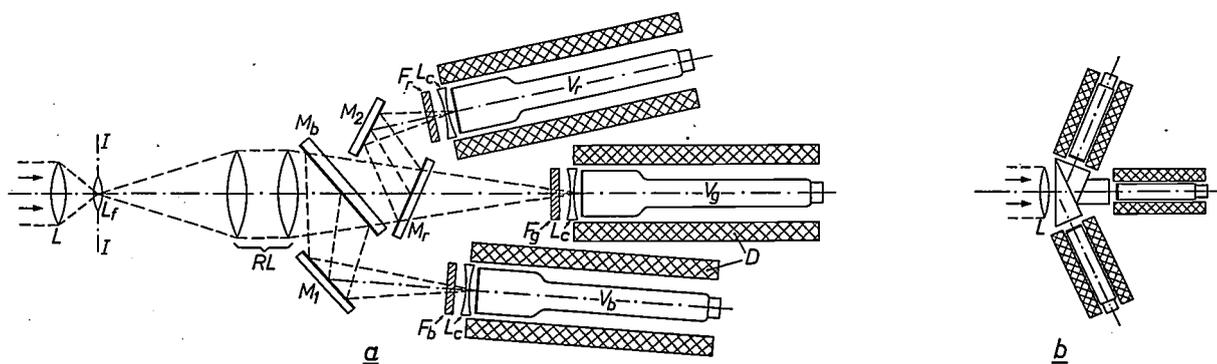


Fig. 3. Practical arrangement of the colour separation systems in a) image-orthicon cameras, b) the "Plumbicon" colour camera, the components being illustrated here on the same scale as those in (a).

In (a), L is the camera lens. To allow lenses with a large relative aperture and normal focal lengths to be used, an optical relay system RL is included. A focal length here of about 150 mm gives sufficient clearance for the colour separation system. L_t field lens for the concentration of the light beam in the optical relay system. The curvature of the image produced by L_t is compensated by the three correction lenses L_c . F_r , F_g and F_b are colour-correction filters to limit more sharply the wavelength ranges separated by the colour-selective mirrors. D deflection and focusing coil systems. The other letters are as in fig. 2.

length of the lens, the relative aperture and the magnification are unaffected by the added glass.

Colour separation must now take place inside the glass. This is achieved by sub-dividing the glass body into three adjacent prisms. The selectively reflecting layers are applied to the separating surfaces, while the re-reflection of the red and blue cones of light as mentioned above is obtained by means of total reflection (figs. 3b and 3c).

We can just briefly point out that with "Plumbicon" tubes, with their small image format (12.8×17.1 mm) and relatively small deflection-coil system, the method described gives considerably better results than a relay system. This has enabled the "Plumbicon" colour camera to be used with first-class lenses of relative aperture up to $f/2$ and a focal length down to 18 mm. These are the same as the values used with black-and-white cameras.

Because the number of optical components, and hence the number of transitions from glass to air, has been reduced to a minimum, the light losses of the entire optical system, which can amount to more than 50% in relay systems, are greatly reduced here, to about 20%. Moreover, definition and contrast of the image are considerably improved by the use of this method. Another advantage is that the light strikes the two

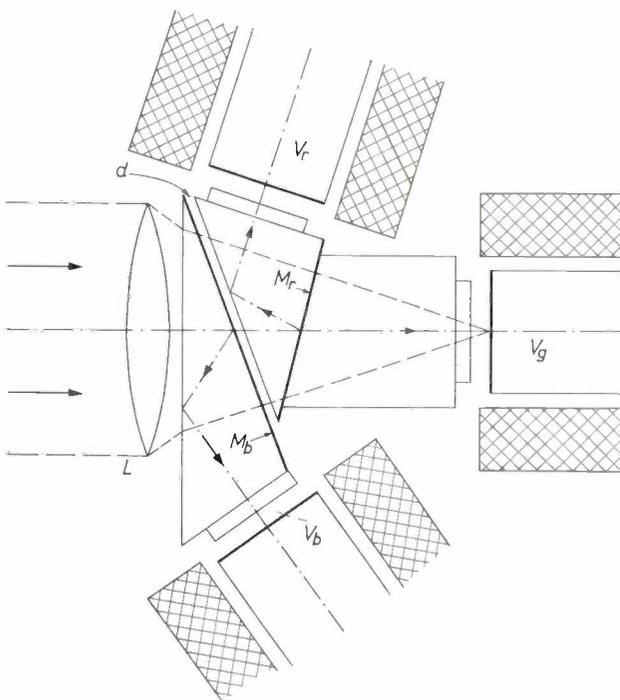


Fig. 3. c) Prismatic colour separation system of the "Plumbicon" camera as in fig. 3b, shown on a larger scale. L , again, is the camera lens. The colour-selective layers M_b and M_r are evaporated on to the rear of the two triangular prisms. The second reflection takes place by total reflection at the front of both prisms; a small air-gap d is provided for this purpose. The other letters are as in fig. 2.

colour-selective reflecting layers at relatively small angles. This means that polarization effects, which can lead to colour errors, particularly if there are specular reflections^[4], occur only to a very slight extent. Finally, as we noted in the Introduction, the reduction of the size of the camera to that of a black-and-white model is primarily due to the compactness of the prismatic colour separation system. Fig. 4 shows the prisms cemented together to form a block, while fig. 1 shows how the block with its protective housing is arranged between the camera tubes.

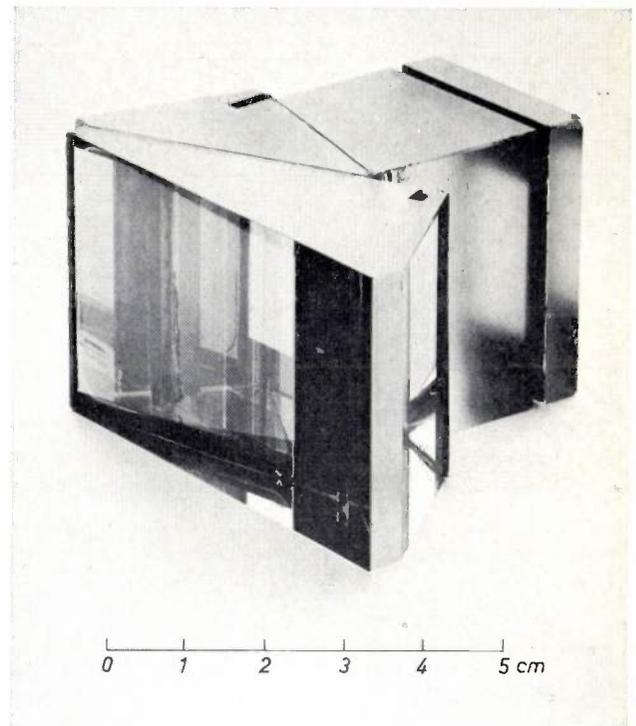


Fig. 4. The glass prisms in the colour separation system are cemented together to form one unit.

If this colour separation system is to be used, two conditions must be satisfied. First of all, the camera lens requires a slight additional correction because the light paths on one side of the lens are almost entirely in glass, not air. This is a matter for the manufacturers of the lenses, and presents them with no difficulties. Secondly, the camera tubes, which are not parallel as in fig. 2, must be given very effective magnetic screening to prevent interfering magnetic fields from causing different distortions of the scanning pattern in the three tubes. The screening is obtained by enclosing the tubes almost completely in mu-metal cases and by applying screening foil to the inside of the camera.

[4] H. de Lang and G. Bouwhuis, Colour separation in colour television cameras, Philips tech. Rev. 24, 263-271, 1962/63.

The pick-up section

Television studio work requires the camera tubes to be quickly and easily interchangeable. With colour cameras, we have the special problem that newly installed tubes must be very precisely positioned in relation to the cones of light from the colour separation system, so as to obtain congruent images on all the tubes. This means that the following conditions have to be satisfied:

- 1) The optical axis of each of the three cones must coincide with the longitudinal axis of the appropriate tube.
- 2) The three image planes of the lens must coincide with the surface of the signal electrode for each of the three tubes.
- 3) The horizontal direction in each of the three images must coincide with the direction of scan.

The mechanical design must be drawn up with this in mind, and the deflection coils and the camera tubes must be engineered to such a degree of precision that these three conditions can be met and continue to be met after a tube has been changed. In fact, the relatively simple construction of the "Plumbicon" tube and its set of deflection coils enabled a system to be devised for changing a tube without the need for mechanical readjustment. The initial setting of the three deflection-coil assemblies is carried out with the aid of an optical alignment device which permits six different adjustments to be made for each coil assembly. These adjustments are given by displacement along and rotation about three axes which are perpendicular to one another and pass through the centre of the signal electrode of the tube. (The rotation about the longitudinal axis of the tube is not in fact made during optical adjustment but by comparison with the horizontal direction of scan.)

Although we have been successful in eliminating the need for mechanical readjustment when a new tube is put in, a certain number of electrical adjustments are still required, as the electrode systems in different tubes are never completely identical.

The circuits

The circuits in a colour television camera may be subdivided into groups comprising of the three signal amplifiers, the deflection, focusing and blanking circuits, the supply section, the electronic view-finder and the signalling and telephone circuits.

The circuits in the first two groups mainly determine the performance of the camera and the quality of the pictures obtained. The design of these circuits is fundamentally affected by the particular features of the "Plumbicon" tube and its deflection system.

The "Plumbicon" tube supplies a signal current that,

in practice, with appropriate lighting of the scene, can vary between about 0.3 μA at the points with the highest luminance and only a few nanoamperes in the darkest shadows. (The tube can be regarded as a source of high internal impedance, which means that the current supplied is almost independent of the load impedance.) These currents are extremely low, 50 to 100 times lower than those for image orthicons. This does not mean, however, that the image orthicon is 50 to 100 times more sensitive than the "Plumbicon". In fact, the sensitivity of a television camera is determined by the scene illumination at which a signal is obtained with a just tolerable signal-to-noise ratio. While the signal current of an image orthicon reaches this signal-to-noise ratio only at rather high values, the noise contribution in the signal current of the "Plumbicon" is negligibly low even at the smallest currents normally occurring in practice (see reference [2]). The signal-to-noise ratio of a "Plumbicon" camera is therefore entirely determined by the noise performance of the *signal amplifier*, which in turn is largely determined by its input stage. On this account the input stage is the one to which most attention has to be paid when designing the amplifier circuit.

To obtain a good signal-to-noise ratio, the amplifier must have as high an input impedance Z_1 as possible. This can be understood in the following way. In our circuit, the input impedance consists of an input resistance R_1 and a capacitance C_p connected in parallel. This capacitance C_p is formed by the input capacitance of the amplifier, in parallel with the signal-electrode capacitance of the "Plumbicon" tube, and the stray capacitance of the connecting leads. $|Z_1|$ (at frequency f) is then given by:

$$|Z_1| = \frac{R_1}{\sqrt{1 + (2\pi f R_1 C_p)^2}} \quad \dots \quad (1)$$

Since we can regard the "Plumbicon" tube as a constant-current source, the signal voltage v_s is proportional to $|Z_1|$:

$$v_s = i_s |Z_1|,$$

where i_s is the signal current. The noise, however, consists mainly of two contributions, one of which, the noise from the first valve or transistor in the amplifier, is independent of Z_1 , while the other, originating in R_1 , is given by

$$v_n = \sqrt{4kT\Delta f} \frac{\sqrt{R_1}}{\sqrt{1 + (2\pi f R_1 C_p)^2}},$$

where v_n = effective noise voltage, k = Boltzmann's constant, T = temperature and Δf = bandwidth. If

$|Z_i|$ is increased, which, as equation (1) shows, may be done by making R_1 greater or C_p smaller, the signal voltage increases proportionally to $|Z_i|$, while the first-stage noise contribution remains constant and the one due to R_1 increases much less rapidly: the signal-to-noise ratio will therefore improve.

It can be seen from equation (1) that the term $(2\pi f R_1 C_p)^2$ in the denominator indicates a decrease in $|Z_i|$ at higher frequencies, and thus a falling response. If we assume, for example, that $R_1 = 1 \text{ M}\Omega$ and $C_p = 25 \text{ pF}$, and if the signal current i_s is $0.3 \text{ }\mu\text{A}$, the signal voltage decreases from 300 mV at zero signal frequency to about 0.5 mV at 5 MHz . This amplitude loss must be compensated in a later stage of the amplifier, by extra amplification of the higher-frequency components of the signal, but this compensation has the adverse effect of accentuating the amplifier noise from the first stage, which is not itself frequency-dependent. On this account, it is desirable to prevent too rapid a decrease of Z_i at high frequencies and therefore not to make the product $2\pi f R_1 C_p$ too great. This is another good reason for keeping C_p as low as possible. Keeping R_1 low, however, has an adverse effect on the signal-to-noise ratio, which as we have shown in the previous paragraph, requires a high value of R_1 .

The two conflicting requirements for the value of R_1 can be largely reconciled by the use of *negative feedback*, as will be seen in the following more detailed discussion of our signal amplifier.

To reduce the amplifier noise from the input stage, nuvistor input circuits (the nuvistor is a small thermionic triode system) have usually been used both for “Plumbicon” and vidicon cameras, although the vidicon cameras were otherwise fully transistorized. Nuvistors give considerably less noise than *P-N-P* or *N-P-N* transistors. Their noise performance has however been surpassed by that of the field-effect transistors which are now available. Moreover, these give a lower input capacitance than nuvistors. For the input stage of the signal amplifier we have therefore chosen a combination of a field-effect transistor and an *N-P-N* transistor. This gives a signal-to-noise ratio 2 to 3 dB better than that obtained with a nuvistor; in an experimental version the improvement was as much as 7 dB.

The basic circuit of the input stage, a cascode circuit with feedback, is given in *fig. 5*. Part of the signal voltage amplified in *Tr1* and *Tr2* and taken from emitter follower *Tr3* is fed back to the input via R_3 and R_2 . The actual control voltage of *Tr1* then becomes:

$$v_1 = i_1 Z_1 + (i_1 + i_3) Z_2.$$

Z_1 and Z_2 are the impedances of the circuits formed by R_1 with its parallel capacitance C_1 (which is

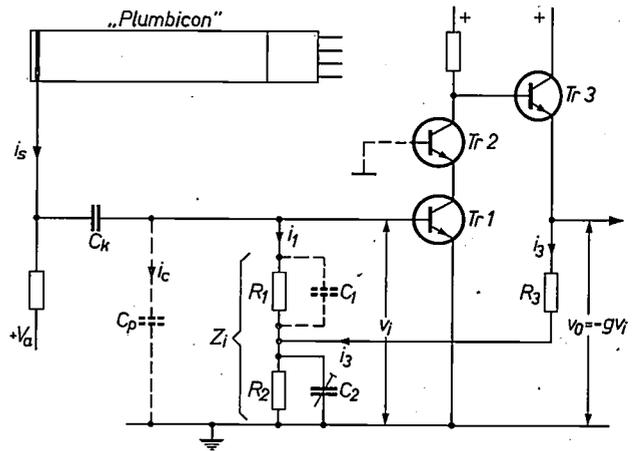


Fig. 5. Simplified diagram of the input stage of a signal amplifier for the “Plumbicon” tube with negative feedback cascode. The tube, which may be regarded as a source of constant current, supplies a signal current i_s , part of which (i_c) flows through the stray capacitance C_p , while another part (i_1) passes through the input impedance Z_1 of the amplifier. The amplified feedback current i_3 generates an opposite voltage across R_2/C_2 so that the control voltage v_1 is lower than that calculated from the product $i_1(R_1/C_1 + R_2/C_2)$. The input impedance is therefore reduced by the feedback.

significant here), and by R_2 and the trimming capacitance C_2 . After substituting

$$i_3 = \frac{-g v_1 - i_1 Z_1}{R_3 + Z_2},$$

where g is the total amplification between *Tr1* and the emitter of *Tr3*, then:

$$v_1 = i_1 \frac{Z_1 + \frac{Z_2 R_3}{Z_2 + R_3}}{1 + g \frac{Z_2}{Z_2 + R_3}}$$

and the “dynamic” input impedance is:

$$Z_i = \frac{v_1}{i_1} = \frac{Z_1 + \frac{Z_2 R_3}{Z_2 + R_3}}{1 + g \frac{Z_2}{Z_2 + R_3}}$$

This expression can be reduced to the form for a simple *RC* circuit if $R_2 C_2$ is made equal to $R_1 C_1$ with the aid of the trimming capacitance C_2 and if $R_2 \ll R_1$. Then:

$$Z_i = \frac{R_1}{1 + g \frac{R_2}{R_2 + R_3}} \frac{1}{1 + j\omega C_1 \frac{R_1}{1 + g \frac{R_2}{R_2 + R_3}}}$$

The total impedance presented to the signal current i_s

supplied by the "Plumbicon" tube is that due to Z_i in parallel with the stray capacitance C_p , and is thus:

$$Z_{tot} = \frac{kR_1}{1 + j\omega kR_1(C_1 + C_p)},$$

where

$$k = \frac{1}{1 + g \frac{R_2}{R_2 + R_3}}.$$

This means that, for a given amplitude-frequency response, R_1 may be $1/k$ times greater than in a non-feedback circuit. To put it another way, the later stages will now only have to compensate for a k -times smaller reduction in amplitude at higher frequencies.

at frequencies up to 6 MHz are fed to the camera cable at equal amplitude. The level of these components lies far above that of any interference that might be picked up in the long cable between camera and control unit.

The excellent noise performance of these signal amplifiers ensures a low interference level in the picture, and their linearity and stability ensure very faithful colour rendering. Another important factor in the quality of a colour television picture is the accuracy of register of the three primary colour pictures: this is determined by the characteristics of the deflection circuits. The designs of the vertical and horizontal deflection circuits are quite different, because of the difference in scanning speed.

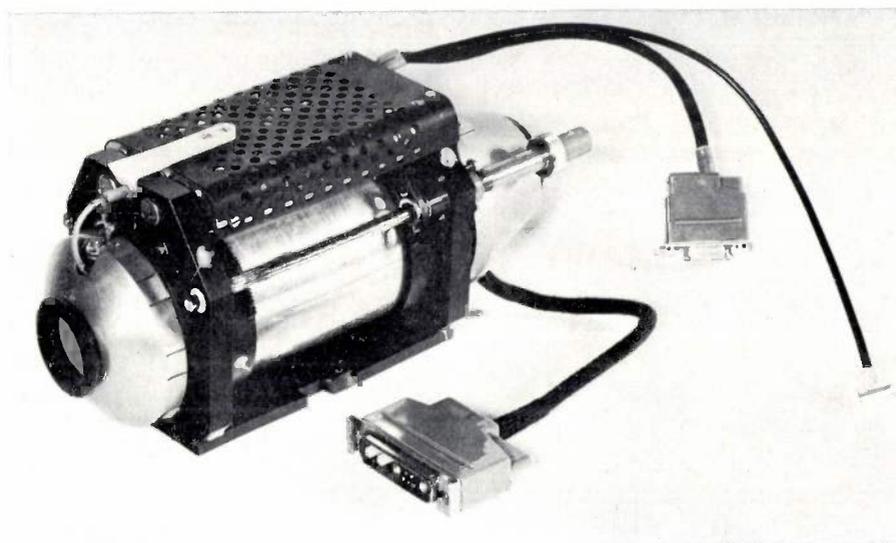


Fig. 6. Deflection and focusing coil assembly for a "Plumbicon" tube with the preamplifier fitted to it and protected by a perforated screening plate. (In the model illustrated here, the preamplifier is an earlier version, with nuvistors.)

With the circuit in question, we have obtained a signal-to-noise ratio of 45 dB at the signal current of 300 nA quoted above and a bandwidth of 5 MHz [5].

To keep the stray input capacitance C_p as low as possible, the three preamplifiers, which are printed wiring boards carrying the input stages, are each fitted directly on top of the appropriate deflection systems, giving a very short distance between input stage and the signal electrode of the "Plumbicon" tube (see fig. 6). With this arrangement, interference due to stray electric fields at the input lead is also kept to a minimum.

Each preamplifier contains only the input stage and emitter follower. From the low output impedance of the emitter follower the signals are fed via screened leads to further amplifier stages in the camera. The reduction in amplitude which we mentioned earlier is compensated in these stages, so that signal components

The *vertical* deflection circuit is located in the control unit and the three finally-shaped saw-tooth currents for vertical deflection in the three tubes are fed directly to the deflection coils via the camera cable. All the controls and adjustments can thus be made at the control unit. As the impedance of the deflection coils is very nearly purely resistive at the low frequency of the saw-tooth currents (50 Hz), this method presents no difficulty. Sufficient stabilization of the desired shape and amplitude of the saw-tooth currents against fluctuations in the temperature of the camera cable and the deflection coils is obtained by current feedback in the output stages of the deflection circuits.

Matters are not so simple for the *horizontal* deflection circuits. A pair of horizontal deflection coils requires a current with a saw-tooth wave form between -100 and $+100$ mA and a repetition rate of 15 625 kHz. The

resistance R of the pair of coils is 3Ω and its inductance L is 1 mH, so that the impedance at a frequency of 15 kHz is $Z = R + j2\pi fL = 3 + 100j (\Omega)$. In order to obtain a saw-tooth current in this almost pure inductance, a pulsed voltage is required whose amplitude is given by the expression:

$$v_L = L \frac{di_{st}}{dt}$$

In practice, a peak voltage of about 60 V is used. Because of the ohmic resistance, a saw-tooth component

$$v_{st} = Ri_{st}$$

has to be added to this voltage (see fig. 7).

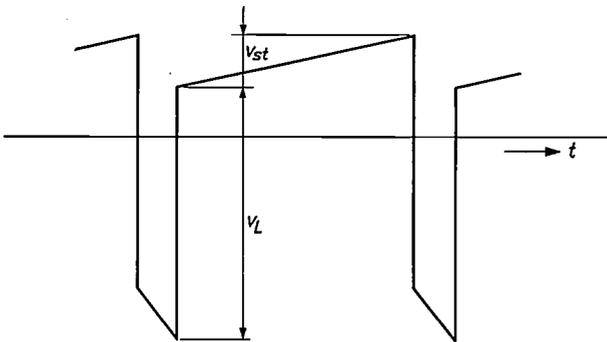


Fig. 7. The voltage waveform required for the horizontal deflection coils of a "Plumbicon" tube. v_L pulse voltage which sets up a saw-tooth current in the coils; v_{st} saw-tooth voltage for linearizing the saw-tooth current.

While the value v_L determines the amplitude of the saw-tooth current, and hence the scanning width, the linearity of the saw-tooth current and the scanning sweep is governed by v_{st} . Apart from these two parameters, the position of the pulsed voltage in relation to the zero line is significant, as this determines the symmetry of the scanning sweep in relation to the vertical centre line of the signal electrode of the camera tube. This requires the addition of an adjustable d.c. component.

For practical reasons, these three important adjustments are made in the central control booth, i.e. at the control unit, and not at the camera. Up to now it has been usual for the basic horizontal deflection waveforms for colour cameras to be generated at the control unit, and distributed from there to the three channels. The three adjustments mentioned above for each channel were also made at the control unit and the individually adjusted waveforms were fed to the deflection coils in the camera via three coaxial leads in the camera cable. This method not only involved appreciable wastage of power on account of

the 75Ω characteristic impedance of the cable (cf. the coil resistance of 3Ω), there were also considerable complications in maintaining the correct voltage waveforms when cables of different lengths and types were used and when there were temperature variations.

The development of very stable transistor circuits has made it possible to shift the whole of the horizontal deflection circuit for the colour camera with "Plumbicon" tubes to the camera itself and so to eliminate the difficulties just mentioned. Now only a trigger pulse for the pulse generator circuit in the camera — this pulse is not critical in shape or amplitude — and a number of d.c. voltages for the control of amplitude, linearity and the d.c. component of the scan are fed through the camera cable. The controls for these voltages are arranged on the control unit and remain thus under the charge of the technician responsible for all the adjustments in the camera installation.

The accurate alignment and focusing of the electron beam in each tube are also important for the precise register of the primary colour images from the three camera tubes. Since the electrode system is never perfectly symmetrical, the beam from the electron gun is generally at a slight angle to the axially directed magnetic focusing field in which the entire tube is located. An oblique entry gives a helical motion to the beam, which upsets not only the exact focusing but also the deflection geometry. This fault can be remedied by aligning the beam with the aid of a transverse magnetic field, adjustable in strength and direction, arranged immediately in front of the electron gun. It is possible to align the beam correctly only if the magnetic focusing field is homogeneous and parallel to the axis. Disturbances of the homogeneity of the field within the tube and especially in front of it must therefore be prevented. The deflection and focusing coils must on this account be manufactured and fitted with the utmost precision. The currents in the focusing coils and in the coils for beam alignment are of course stabilized.

The other circuits in the camera are little different from those of other cameras. For completeness we shall just give a very brief description of the function of these circuits.

The *blanking circuit* is designed to suppress the scanning beam during the horizontal and vertical fly-back of the scan to prevent partial erasure of the charge pattern on the tube target in these periods. This is done

[5] In television, the signal-to-noise ratio is regarded as the ratio between the maximum signal amplitude at a black-and-white transition and the r.m.s. value of the noise. For the "weighting" of the noise and the associated filter, see reference [2], page 5. Measurement with a noise-weighting filter, as used for black-and-white television, has under the given conditions indicated a signal-to-noise ratio of as high as 60 dB.

with blocking pulses obtained from the horizontal and vertical deflection circuits in the camera and combined to form a common series of pulses. These pulses are added to three negative d.c. voltages adjusted individually at the control unit and fed to the camera via the cable. The sum of these d.c. voltages and the blanking pulses is applied to the control grids of the three "Plumbicon" tubes. The d.c. voltages determine the beam currents of the tubes.

In television cameras, the *electronic view-finder* replaces the optical type found on cine cameras and allows the cameraman to aim his camera accurately. The view-finder is simply a small monitor which displays as a black-and-white picture the video signal received from the control unit after processing for transmission. In the vast majority of cases the green picture is most suitable for black-and-white reproduction, but, by pressing the appropriate button, the cameraman can select the red or blue picture or any combination of the three for display in black-and-white on his view-finder screen. Furthermore, the signal from another camera can be superimposed on his own view-finder image; this is very useful in special work where pictures from two cameras are combined.

The *signalling* and *telephone* circuits which include various indicator lamps arranged at suitable points on the camera and two headphone connections, give a communication link between cameraman and assistant and the technician in the control room and the director. These circuits are also used to give an indication to those in the studio that the camera is on the air. The indicator lamps for this are switched on automatically as soon as the camera signal is applied to the preview monitor in the control room, giving the final check before the actual transmission.

The control unit

The control unit of every colour television camera has four main functions: the "refinement" in various aspects of the three colour signals supplied by the camera, which takes place in the *processing amplifiers*; adjustment of the signals for optimum colour rendering under different conditions of illumination; adjustment for optimum register and definition of the three primary colour images; checking these adjustments with the aid of a signal monitor and a picture monitor; and finally the provision of power for the entire installation.

Instead of fitting all the components for these functions into a single constructional unit, as in the camera, it is preferable to arrange them according to their function in sub-units on a vertical rack. This facilitates check measurements and servicing. All important units are of the slide-in type and the circuits are arranged on quickly interchangeable printed-wiring boards.

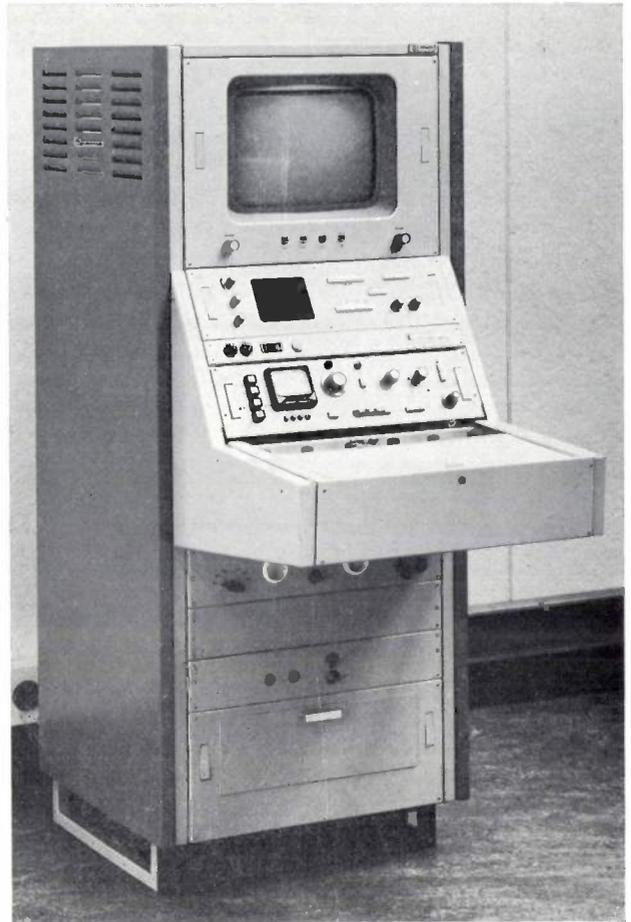


Fig. 8. Control unit for the "Plumbicon" colour television camera. From top to bottom: the picture monitor; the signal monitor; operational panel with controls for diaphragm adjustment, maximum and zero level setting (right-hand half), meter for indicating the focal length of the zoom lens, and main switch; this panel also contains push-buttons for selecting the signal to be displayed on the picture monitor and for the signalling system. The desk section contains all the control knobs for adjusting the colour balance and the register of the three primary colour pictures. The lower part of the cabinet is occupied mainly by the supply section and a few auxiliary circuits. It also contains a circuit with a selector switch for adjustment of the correction for different lengths of camera cable.

Fig. 8 illustrates the complete control unit. Those sections which have not yet been dealt with will now be discussed.

The processing amplifiers

As stated, the processing amplifiers "refine" the camera signals in various aspects. The maximum and zero levels are firmly established, extreme peaks in the signals are rendered innocuous, special measures are applied to improve the definition of the picture, and "gamma correction" is applied to compensate for the non-linear response of the receiver picture tube. Finally the blanking signal prescribed in the television standards is added to each signal to obtain the three primary colour signals, which can be combined to form a com-

posite colour signal in an encoder appropriate for the transmission system to be used.

We shall discuss these functions in slightly more detail with reference to the block circuit diagram (fig. 9) of one of the three almost identical processing amplifiers.

First of all, firm zero levels are established in the signals arriving from the camera. Due to the blank-

ment of the definition by the electronic accentuation of contours, known as *aperture* or *spot correction*. We shall go into this in slightly greater detail.

Contours in the image give rise to sudden changes in amplitude in the electrical signal from the camera tube. For a number of reasons, in particular the finite diameter of the scanning spot, the changes in amplitude are not abrupt but take the form of more or less gradual

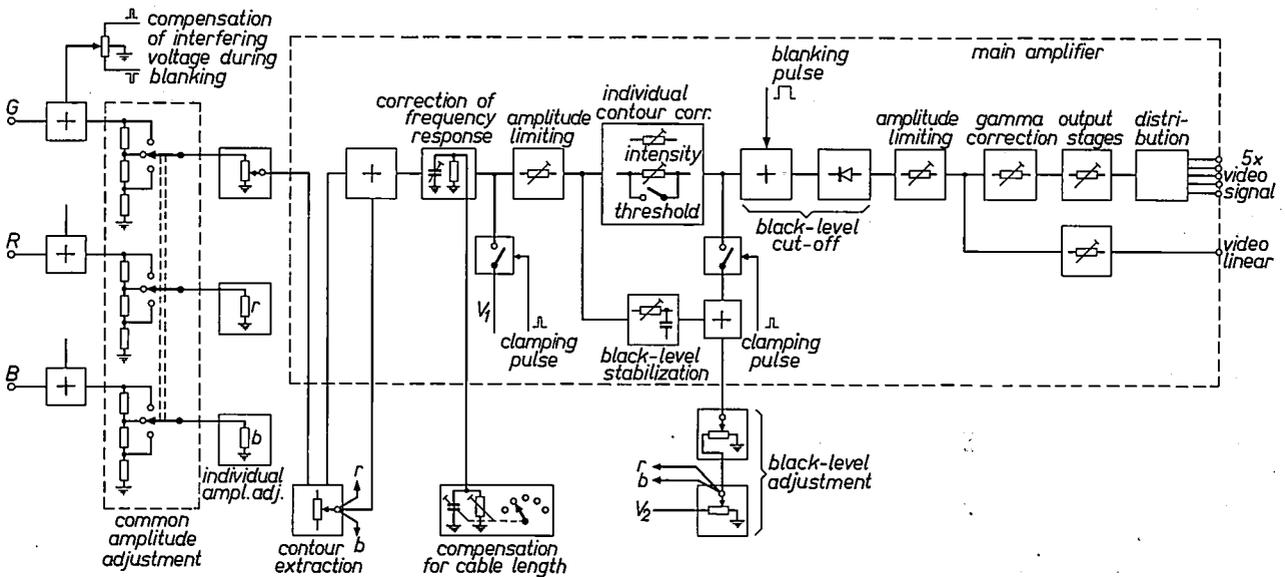


Fig. 9. Block circuit diagram of the processing amplifier for the “green” signal with the external adjustments. Apart from the contour extraction for the common aperture correction (on the “green” amplifier only), all three amplifiers are identical.

Each amplifier also contains an individual aperture corrector; both the amplitude and the threshold of the signal at which this correction should start can be adjusted.

After passing through the gamma corrector and after a final adjustment of the frequency response, the signal is distributed over five independent outputs. The signal as it was before gamma correction is available at the sixth output for measurement purposes.

ing during each flyback period, the picture signal would contain a zero level, but this level generally shows a certain sinusoidal variation as a result of an interference voltage induced in the electrodes of the tube during the rapid reversal of the line-deflection field. In practice it is sufficient to establish a firm zero level each time just for the short period during which the clamping pulses are effective in the following amplifier stages. For this reason, a small pulse adjustable in amplitude and polarity is derived from the clamping pulses, and added to the signal in the first stage of the amplifier. This pulse is adjusted for each tube in such a way that the interfering voltage is exactly compensated during the pulse (see fig. 10).

After common coarse amplitude adjustment of all three signals and an individual fine adjustment (highly important for faithful colour rendering), a process is applied which has recently acquired great significance, particularly for colour cameras: this is the improve-

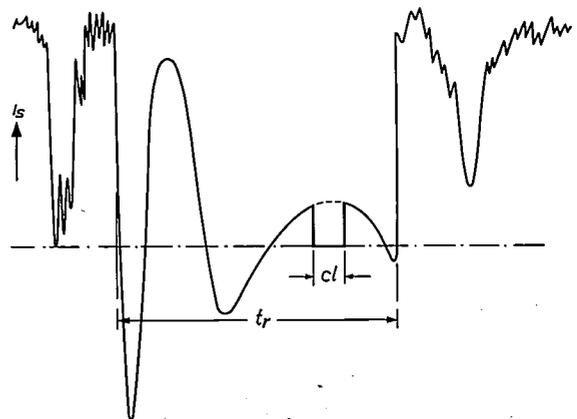


Fig. 10. The beam is suppressed during every horizontal scanning flyback period t_r and the picture signal supplied by the “Plumbicon” tube drops back to zero (chain-dotted line). Since, however, during this period there is a substantially sinusoidal voltage due to interference, the zero level is set exactly during the clamping pulse cl (falling within t_r) by compensating the momentary interference voltage with an adjustable voltage pulse.

transitions (see *fig. 11a*). This decreases the sharpness of the contours in the received picture.

There are several methods of aperture correction to compensate for this loss of definition, and, because the image is scanned in successive lines, a distinction must be drawn between "horizontal" and "vertical" aperture correction. A method of horizontal aperture correction which we have found very useful with the "Plumbicon" camera is to make the signal transition even worse by restricting the bandwidth to 2 MHz and then to subtract this degraded signal from the original signal (see *fig. 11b*). The result is a pure contour signal, i.e. a signal whose average level is zero and which consists only of small "pulses" of opposite polarity at those points in the picture where the scanning beam cuts a contour (*fig. 11c*). This contour signal, added at the correct amplitude to the original signal, provides a noticeable accentuation of the changes in amplitude of the original signal and thus an accentuation of the contours of the picture (*fig. 11d*).

It is easy to see that this accentuation is most effective for *vertical* contours, and decreases as the contours depart from the vertical. The effect is zero where the

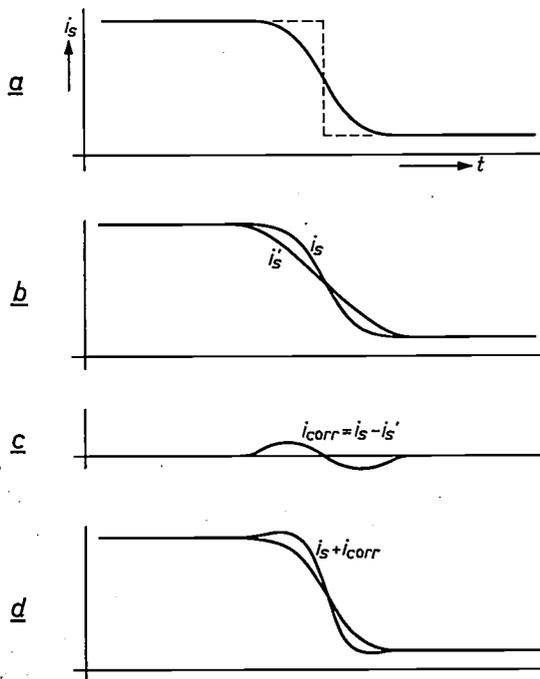


Fig. 11. Electronic accentuation of *vertical* image contours.
a) With a sharp image contour, the abrupt change in amplitude shown in the dashed line should occur in theory. In reality, however, the change in signal amplitude is continuous, as shown in the solid line, for a number of reasons.
b) An even more "blurred" signal i'_s is obtained by limiting the bandwidth, and this signal is subtracted from i_s .
c) The difference between i_s and i'_s is a pure contour signal i_{corr} with an average value of zero.
d) Adding the contour signal i_{corr} to the original signal i_s accentuates the abrupt transition and thus improves the definition of the vertical contours.

contours are exactly in the direction of the scan. Contours like this are much more difficult to accentuate as this requires signals of successive lines to be subtracted from one another. After a thorough investigation of existing methods, we developed the following method for the colour camera with "Plumbicon" tubes (see *fig. 12*).

Part of the video signal is delayed by exactly twice the line period ($2 \times 64 \mu s$) by converting it into ultrasonic vibrations and sending these vibrations along two special glass rods of appropriate length, connected in cascade. The conversion between electronic and ultrasonic signals is made by means of piezo-electric transducers. The amplitude of the double-delayed signal i_{s2} is halved and its polarity is reversed. The same occurs with the original undelayed signal i_s , and both signals are added in a mixing stage to an unattenuated signal of the original polarity, which has been delayed by only one line period (i_{s1}). *Fig. 13* shows the whole process. The addition of $\frac{1}{2}i_{s1}$ and $\frac{1}{2}i_{s2}$ has the same effect as the bandwidth limitation in *fig. 11* and the result of the complete operation is a pure contour signal i_{corr} which is at its greatest for *horizontal* contours, decreases as the contours depart from the horizontal and disappears for *vertical* contours.

If both the undelayed signal i_s and the double-delayed signal i_{s2} are limited in bandwidth to about 2 MHz (by *LP* in *fig. 12*) before they are fed to the mixing stage, then we have also in a rather simple way included the accentuation of the vertical contours as described above. We then have effective contour accentuation for *all* directions of contour.

The reader may wonder why a separate contour signal $i_{corr} = i_{s1} - \frac{1}{2}(i_s + i_{s2})$ is derived in the circuit we describe, instead of a directly corrected picture signal $i_{s1} + i_{corr} = 2i_{s1} - \frac{1}{2}(i_s + i_{s2})$. Such a signal could readily be obtained if the signal delayed by one line period was doubled before being fed to the mixing stage. There would indeed be some point to this for black-and-white signals [6], but not for colour signals. In fact, in colour television, the idea is not just to accentuate each separate signal — which could be done by such a "direct correction" — but, in the final instance, to improve the definition of the total picture, which is obtained by the superimposition of the three pictures in the primary colours. Since it is virtually impossible to obtain the ideal case where the three pictures are exactly in register over the entire surface of the screen, accentuation of the contours of each *separate* signal provides not only no essential improvement, but may even adversely affect those points of the

[6] Cf. also the method described in this journal some time ago: C. F. Brockelsby and J. S. Palfreeman, Philips tech. Rev. 25, 234-252, 1963/64.

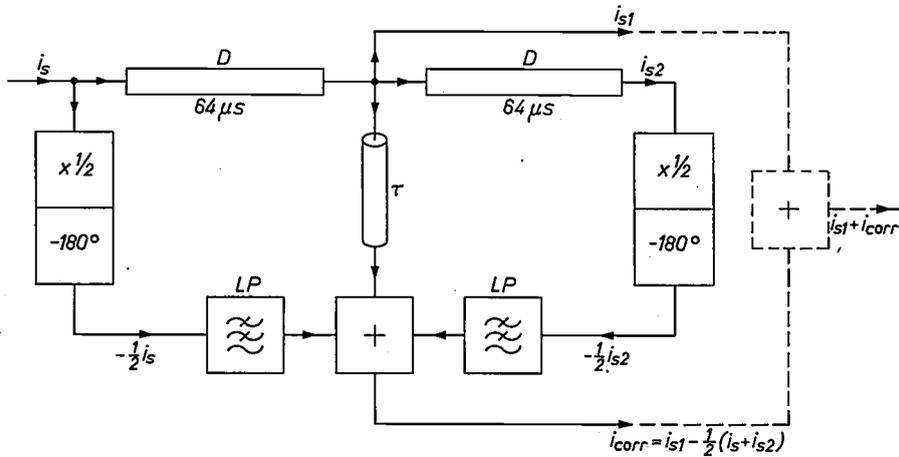
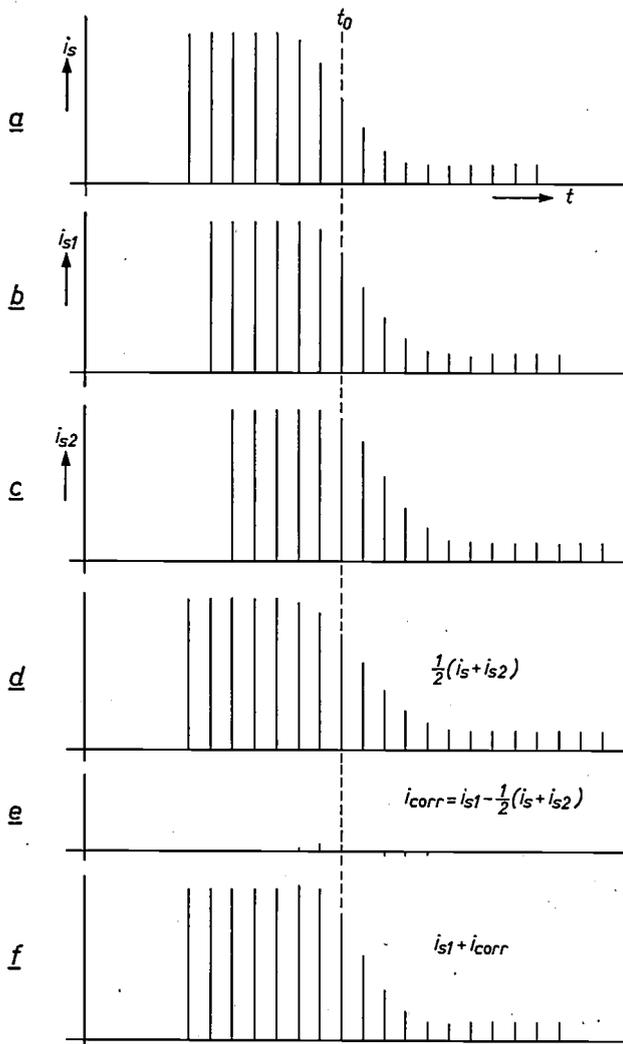


Fig. 12. Block circuit diagram of the arrangements for the electronic accentuation of *horizontal* contours. The signal i_s is delayed by exactly one line period ($64 \mu s$) in each of the two delay lines D . The undelayed signal i_s and the twice-delayed signal i_{s2} are each reduced to half amplitude and their polarities are reversed. They are both added to the signal i_{s1} , which has been delayed by one line period (cf. fig. 13). The result is a pure contour signal which, on being added to the signal i_{s1} (dashed lines), accentuates the horizontal contours.

The bandwidth of the signals i_s and i_{s2} that are used for correction is limited to 2 MHz, by filters LP . This provides at the same time in a simple manner the accentuation of the *vertical* contours as shown in fig. 11. The result is therefore a contour signal i_{corr} that accentuates *all* the contours. The bandwidth limiting gives the signals i_s and i_{s2} a small extra delay τ ; this delay must also be given to signal i_{s1} before the addition is made.



picture where there is a discrepancy between the primary colour pictures. These faults are then made even more apparent. It is much better to use a *common* correction signal and, since it must be usable for three different colour extracts, this can, of course, only be a pure contour signal.

In theory there are several ways in which such a common contour signal could be derived from the three colour signals and added to the composite colour signal. In our opinion, the best method is to derive the contour signal from the *green* signal only and then to add it at equal amplitude to all three. This means that, whether or not the three pictures are in themselves completely congruent, they are all given congruent contour lines. Discrepancies between the pictures are thus not only not accentuated but may even disappear if they are small. The addition of the contour signal at equal amplitude to all three picture signals accentuates the

Fig. 13. Operation of the circuit in fig. 12.

- a) Signal amplitudes i_s at points exactly one beneath the other in successive lines in a picture at an abrupt transition from light to dark in the scene. The transition is gradual because of blurring. The line at time t_0 is half-way between light and dark.
- b) The same signal amplitudes as in (a) but delayed by one line period with respect to t_0 .
- c) Signal amplitudes i_{s2} delayed by two line periods with respect to t_0 .
- d) Half the sum of i_s and i_{s2} .
- e) The difference $i_{s1} - \frac{1}{2}(i_s + i_{s2})$ is a pure correction signal i_{corr} with an average value of zero.
- f) The addition of the contour signal i_{corr} to the signal i_{s1} , which is delayed by one period, gives an accentuated transition in amplitude and thus an improvement in the impression of sharpness for horizontal contours.

contours, so to speak, in black-and-white (a method which has also been found to be successful in colour printing). The level of the added signal i_{corr} can be set visually to the optimum value (the facility of such a simple adjustment is another advantage over "direct correction"). The green signal is the preferred choice for the common correction signal because, unlike blue and red, it occurs in most natural colours, and often predominates; in white, for instance, the proportion of green is 59%. It has, in fact, been found that extremely good results are obtained with "contours from green" on black-and-white receivers — and for a number of years these will still be the receivers most widely used for receiving colour television broadcasts.

For the sake of completeness we should mention that the "contours from green" method which has just been described is just as powerless as any other to correct lack of register between pictures which may be caused by a fault in the receiver.

Apart from the provisions for the common contours-from-green improvement of definition, the block diagram in fig. 9 also shows an *individual* aperture correction for the red, green and blue signals. Level and threshold of this correction are adjustable. This allows the user to make a further correction, if desired, to the definition at the camera; this can be useful for the highest signal frequencies.

We now come to *gamma correction*. The characteristic of the "Plumbicon" camera tube is linear, i.e. the signal current is proportional to the locally varying luminance of the received image. In contrast to this, however, picture tubes have a non-linear characteristic such that the luminance increases with (at least) the square of the control voltage, so that low signal amplitudes are reproduced by a luminance which is relatively far too low. This would lead to very "hard" pictures in black-and-white television, but in colour television there would also be errors in colour, since the intensity ratio of the three sets of colour information is no longer correct at different signal amplitudes. The remedy here is to reshape the colour signals before transmission in an amplifier stage whose amplitude characteristic has the opposite curvature to the characteristic of the picture tube (see fig. 14). If the picture tube has a characteristic $L \propto v_g^\gamma$, where L is the luminance, v_g the control voltage at the Wehnelt cylinder and γ the exponent corresponding to the curvature, the characteristic of the correction stage must be $v_o \propto v_i^{1/\gamma}$, where v_o and v_i are the input and output voltages. Such a characteristic can be obtained with a semiconductor diode whose current-voltage characteristic has the desired shape at the beginning of the conducting range. This correction circuit, which can be basically very simple, should allow the degree of

curvature to be adjusted between about $\gamma = 0.4$ and $\gamma = 1.0$; the circuit is made slightly more complicated by the additional requirement that when making adjustments the maximum amplitude of the signal, once set, should not vary.

We should point out here that gamma correction is always effected in the studio equipment and not in the receiver. With compensation at the camera the transmitted signal becomes less sensitive to any interference that can arise along the transmission path.

The prerequisite for the correct operation of a gamma correction circuit is a well-established zero level that remains stable under all lighting conditions. As the application of gamma correction has the result that changes in the lowest signal amplitudes receive the greatest amplification there is the danger that annoying colour errors could occur in the shadows if the zero levels of the three signals are not perfectly stable. An unacceptable change in the zero level was found to occur with large fluctuations of the average luminance of the scene. The explanation for this is that a considerable amount of the light striking the signal electrode of the "Plumbicon" tube is scattered from it, and in spite of anti-reflection measures in the colour separation system, is reflected back to the tube as faint, diffuse light. Variations in the average luminance of the scene then cause proportional variations in the scattered light and thus an incorrect zero level.

This difficulty was eliminated by using a control circuit which automatically alters the zero-level setting for each of the three camera tubes in proportion to and

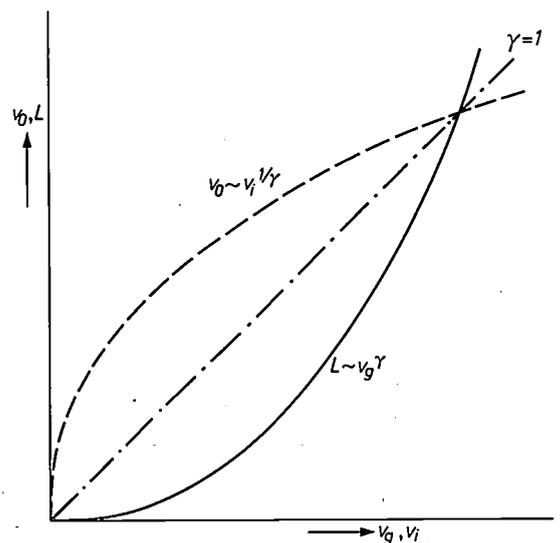


Fig. 14. Principle of gamma correction. The luminance L in the picture tube of the receiver varies with the control voltage v_g as $L \propto v_g^\gamma$, where $\gamma \approx 2.2$. The marked curvature of this line can be pre-compensated in the processing amplifier of the camera installation with the aid of the oppositely curved characteristic of a crystal diode. This requires the diode characteristic to be $v_o \propto v_i^{1/\gamma}$.

in the same direction as the change in average signal strength (the proportionality factor can be accurately set).

Here we should mention another annoying phenomenon that is connected with scattered light: this is the halo effect. Total reflection in the glass faceplate of the "Plumbicon" tube produces an annular brightening of the picture around the highlights. The surface of the signal electrode looks slightly reddish, and the halo effect is therefore most marked in the tube in the red channel. An adequate cure for this has now been found^[7]. The thickness of the window of the tube is increased to more than 7 mm by cementing a plane-parallel ground glass plate to it. The totally reflected

points in each processing amplifier. In one of the first stages, all signals above 130% of the rated maximum are clipped, thus protecting the following stages from being overloaded. The second limiting occurs after the circuit for setting the zero level and accurately determines the maximum signal amplitude, for which a peak value of 1 V is usually chosen.

Monitoring and control circuits

After this description of the most important functions of the amplifier section, we shall just mention a few of the special features of the other circuits.

The signal monitor serves to check whether the signal levels and the gamma characteristics of the three pro-

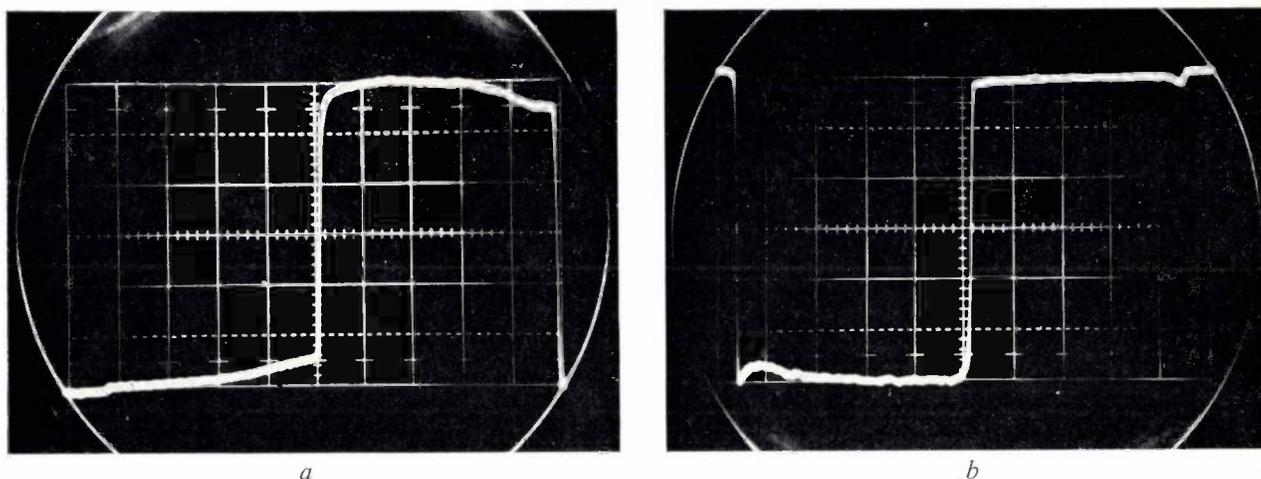


Fig. 15. Elimination of the halo effect by a plane-parallel glass plate cemented to the window of the "Plumbicon" tube. The curve of the signal at a black-white transition is recorded on an oscillograph *a*) for a tube without and *b*) for one with a glass plate. The error in amplitude caused by the halo effect is 9.5% in *a*) and only 2% in *b*). (Taken from^[7].)

light now falls upon the blackened cylindrical rim of the plate and is absorbed there. Fig. 15 shows two oscillograms demonstrating the effect of the glass plate.

In colour television, the accurate setting of the maximum signal level (equivalent to the white level in black-and-white television) is just as important as the stabilization of the three zero levels. Specular reflections in the scene to be taken are almost inevitable. While with expert adjustment the "Plumbicon" tube can handle bright spots of luminance up to 6 to 8 times the normally occurring maximum value without unfortunate side effects, a result of the linear characteristic is that these bright spots give peaks in the tube current which are 6 to 8 times the rated maximum. This rated maximum is the most important parameter for the entire transmission channel and all the circuits are therefore based on it. The occurrence of unexpected bright spots would, therefore, lead to overloading, with all its adverse consequences on picture quality.

To prevent this, there is an amplitude limiter at two

processing amplifiers have been properly adjusted. This monitor is a cathode-ray oscilloscope with three identical vertical amplifiers and a time-base which can be synchronized with either the horizontal or the vertical deflection of the camera tubes. Electronic switching allows the three colour signals to be displayed beside one another or superimposed, thus facilitating very precise relative adjustment of these signals. Once this adjustment has been made, the signal monitor is used during normal operation solely for checking the maximum level of the signals. Deviations from the rated values are then simply compensated by the readjustment of the camera lens diaphragm or, if an abrupt change is permissible, by the common coarse amplitude adjustment.

The signal monitor is the only instrument that is actually necessary for checking during normal operation.

^[7] F. W. de Vrijer, S. L. Tan and A. G. van Doorn, Advanced techniques for "Plumbicon" cameras, *J. SMPTE* 75, 1080-1082, 1966 (No. 11).

The picture monitor is required in the first place for the preliminary mechanical and electronic setting of the camera tubes and for checking and adjusting the scanning geometry and hence the register of the primary colour pictures. Because of space restrictions a colour monitor is not used; instead there is a black-and-white monitor, which can display any of the primary colour pictures or all three of them superimposed. During normal operation, the picture monitor is only used to check on camera operations and to see if everything is proceeding according to plan.

From the above it can be seen that the adjustments can be divided into settings that must be made before and those that have to be made during operation. The only ones made during operation are the setting of the diaphragm, the common regulation of the maximum level and the common adjustment of the zero level. The control knobs for these adjustments are on the operational panel, but their connections can also be switched over to a second, identical, panel which may be arranged in a common control desk for several cameras. The technician at this desk then has to compare the brightness and contrast of the pictures from the different cameras with the aid of monitors and bring them into line with one another.

It is usually necessary to make the numerous adjustments to the camera tubes, signal amplifiers and deflection circuits only when camera tubes have been changed or after the installation has been in use for a considerable time. The appropriate controls are to be found on the desk of the control unit (*fig. 16*). They are protected against inadvertent or unauthorized operation. The prerequisite for the reliable operation of all controls and circuits is, of course, careful stabilization of all supply voltages and protection of the circuits from any kind of interference or interaction effects.

Practical results

The practical experience obtained with the type of camera described has completely fulfilled our expectations of the use of "Plumbicon" tubes in colour television cameras. The whole installation is simple to operate and the operations are easy to grasp. No complications have been found in the adjustments before operation, and the set values remain stable and need no further readjustment for considerable periods of time. The results obtained are hardly affected at all by fluctuating ambient temperatures, and extreme changes in the lighting, such as those which occur when changing from indoor to outdoor shots, are handled without difficulty and without any loss of quality.

These cameras have now been in use in many colour television studios for more than two years, either for

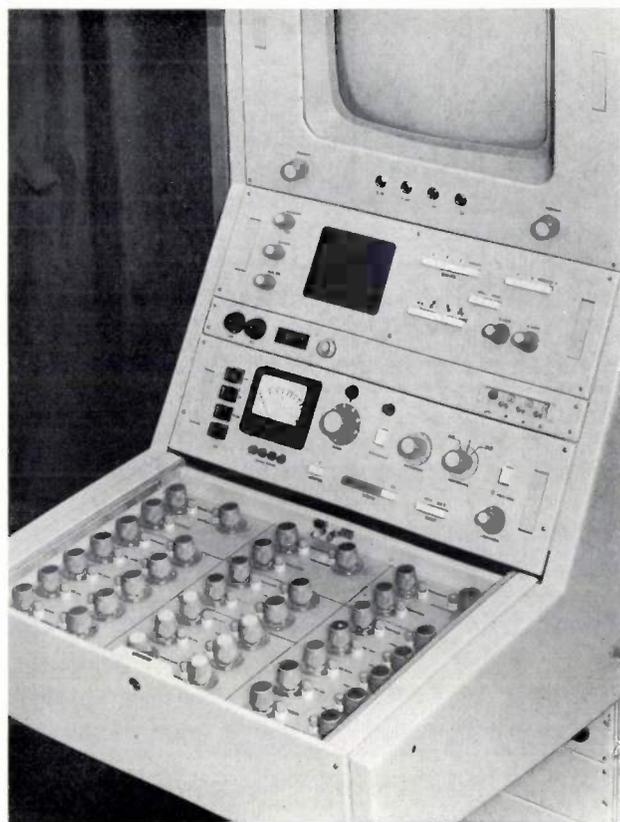


Fig. 16. Control desk with controls for setting the signal amplifiers to the correct colour balance and the deflection circuits for the best possible register of the primary colour pictures. These controls do not need to be operated during shooting, and are then protected by a cover from inadvertent alteration. In the figure, the cover has been pushed back into the cabinet.

normal programme production, as in the United States, Canada and Japan, or for experimental colour television work, as in most of the countries of Western and Central Europe. To obtain an assessment of the performance of the camera and the quality of the pictures it gives, many comparisons have been made, under fully controlled conditions, with other colour cameras, of both the three- and the four-tube types. The three-tube type includes the conventional image-orthicon camera, and the four-tube type includes those with one image orthicon for the luminance signal plus three vidicons for the colour signals and those with four "Plumbicon" tubes, where the fourth is used for the luminance signal.

All comparisons have shown that the type of camera described here requires markedly less light for high-quality studio shots than the others, including four-tube types. The colour rendering of the camera in outdoor scenes, where the lighting is generally considerably less uniform than in the studio, is clearly better than that of the other types. The contrast range of the scene which can be dealt with is better under all conditions

than the range which can be handled with image-orthicon cameras. The definition and the register of the three primary colour pictures is considerably better than in three-tube image-orthicon cameras and as good as those in such cameras with four tubes. We should like to emphasize once more that the picture definition is important not only for reproduction in colour receivers but also for reproduction of the colour transmissions in black-and-white receivers (this is the compatibility requirement). The good results which we have obtained in this respect with the camera with three "Plumbicon" tubes — due to the use of "contours from green" — may perhaps settle the question, discussed over the past few years, as to whether three or four tubes are best. As compared with the theoretically and practically

more complicated four-tube cameras, which were primarily developed to provide sharper reproduction of colour programmes in black-and-white receivers, we now have one three-tube camera which is their equal in this respect, but provides better colour rendering and much simpler operation.

Even though we already have results which are excellent, the scope for the further development of this type of camera is considerable. Such development may be directed towards further circuit refinement made possible by the recent rapid progress in solid-state components, towards the increasing use of automation in television practice, and towards further perfection of the "Plumbicon" camera tube, which is still a relatively recent development.

Summary. The small size of the "Plumbicon" camera tube and its relatively simple operation in comparison with image orthicons permit the design of a colour television camera in which the mechanical and optical arrangements and the circuits are different from those in conventional cameras. This article discusses the most important features of the new prismatic colour separation system, the simple, practical methods of adjusting the camera tubes, and the specially developed amplifiers and deflection cir-

uits. The performance of the circuits is greatly assisted by the linear characteristic of the "Plumbicon" tube and the negligibly low noise level of its signal current, provided that these particular features are taken into account in the design of the circuits. Finally, an account is given of the practical experience gained so far in ordinary studio work. The conclusion drawn is that the type of three-tube camera described is to be preferred to cameras with four tubes.

Recent scientific publications

These publications are contributed by staff of laboratories and plants which form part of or co-operate with enterprises of the Philips group of companies, particularly by staff of the following research laboratories:

Philips Research Laboratories, Eindhoven, Netherlands	<i>E</i>
Mullard Research Laboratories, Redhill (Surrey), England	<i>M</i>
Laboratoires d'Electronique et de Physique Appliquée, Limeil-Brévannes (S.O.), France	<i>L</i>
Philips Zentrallaboratorium GmbH, Aachen laboratory, Weisshausstrasse, 51 Aachen, Germany	<i>A</i>
Philips Zentrallaboratorium GmbH, Hamburg laboratory, Vogt-Kölln-Strasse 30, 2 Hamburg-Stellingen, Germany	<i>H</i>
MBLE Laboratoire de Recherche, 2 avenue Van Becelaere, Brussels 17 (Boitsfort), Belgium.	<i>B</i>

Reprints of most of these publications will be available in the near future. Requests for reprints should be addressed to the respective laboratories (see the code letter) or to Philips Research Laboratories, Eindhoven, Netherlands.

- J. Adams:** X-ray detection by channel electron multipliers.
Adv. in Electronics and Electron Phys. **22A**, 139-153, 1966. *M*
- C. S. Aitchison, R. Davies & P. J. Gibson:** A simple diode parametric amplifier design for use at *S*, *C*, and *X* band.
IEEE Trans. on microwave theory and techniques **MTT-15**, 22-31, 1967 (No. 1). *M*
- D. Andrew:** The performance assessment of sputter ion pumps.
Vacuum **16**, 653-657, 1966 (No. 12). *M*
- E. Andrich:** Properties and applications of PTC thermistors.
Electronic Appl., Comp. and Mat. **26**, 123-144, 1965/66 (No. 3). *A*
- P. Beekenkamp & G. E. G. Hardeman:** Etude par résonance magnétique nucléaire de la structure des verres du système $\text{Na}_2\text{O}-\text{B}_2\text{O}_3-\text{P}_2\text{O}_5$.
Verres et Réfr. **20**, 419-426, 1966 (No. 6). *E*
- C. W. Berghout:** Precipitatie in magnetische legeringen. Metalen Constr.mat., spec. No. Precipitatie in legeringen, 1966, p. 76-80. *E*
- G. Blasse & A. Brill:** Fluorescence of Eu^{3+} -activated oxides of the type AB_2O_6 .
Philips Res. Repts. **22**, 46-54, 1967 (No. 1). *E*
- H. W. Bodmann:** Der Helligkeitsunterschied als Funktion von Objekt- und Umfeldleuchtdichte.
Internat. Farbtagung, Luzern 1965, p. 203-207; Musterschmidt, Göttingen 1966. *A*
- H. W. Bodmann:** Grundlagen der Großraumbürobeleuchtung.
Dtsch. Bauzeitschr. **14**, 1109-1110, 1113-1114, 1966 (No. 6). *A*
- H. W. Bodmann, G. Söllner & E. Senger:** A simple glare evaluation system.
Illum. Engng. **61**, 347-352, 1966 (No. 5). *A*
- A. J. van Bommel & F. Meyer:** LEED measurement of H_2S and H_2Se adsorption on germanium (111).
Surface Sci. **6**, 391-394, 1967 (No. 3). *E*
- H. Bosma:** Performance of lossy *H*-plane *Y* circulators.
IEEE Trans. on magnetics **MAG-2**, 273-277, 1966 (No. 3). *E*
- G. Brédart & Ph. van Bastelaer:** Les paramètres matriciels du transistor et leur emploi dans l'étude du comportement transitoire des circuits transistorisés (cas des petitssignaux). I repartie: Les paramètres matriciels.
Rev. MBLE **9**, 152-163, 1966 (No. 3). *B*
- J.-J. Brissot:** Préparation et propriétés des monocristaux destinés aux lasers.
Acta electronica **10**, 7-35, 1966 (No. 1). *L*
- J. van den Broek:** Optical absorption of red lead monoxide.
Philips Res. Repts. **22**, 36-45, 1967 (No. 1). *E*
- P. H. Broerse:** Electron bombardment induced conductivity in lead monoxide.
Adv. in Electronics and Electron Phys. **22A**, 305-314, 1966. *E*
- J. Burmeister:** Beobachtungen zur Polymorphie des Rickardits Cu_4Te_3 .
Z. Metallk. **57**, 325-326, 1966 (No. 4). *A*
- K. H. J. Buschow, A. M. van Diepen & H. W. de Wijn:** Anomalous behaviour of the Knight shift in SmAl_2 .
Physics Letters **24A**, 536-537, 1967 (No. 10). *E*
- H. J. Butterweck:** Scattering bounds for the general loss-less reciprocal three-port.
IEEE Trans. on circuit theory **CT-13**, 290-293, 1966 (No. 3). *E*

- A. Charles-Georges & A. Salmon:** Taillage des monocristaux pour lasers.
Acta electronica **10**, 37-54, 1966 (No. 1). *L*
- P. Chevalier & J. Nussli:** Photomultiplicateur à haute résolution utilisant un multiplicateur semi-conducteur.
C. R. Acad. Sci. Paris **264B**, 462-465, 1967 (No. 6). *L*
- A. Cohen:** Eigenschappen van de spreektaal.
Natuurk. Voordr. 1965-1966, No. 44, p. 99-111; Diligentia, Den Haag 1966. *E*
- A. Cohen, I. H. Slis & J. 't Hart:** On tolerance and intolerance in vowel perception.
Phonetica **16**, 65-70, 1967. *E*
- B. J. Curtis:** The precipitation of graphite from the hexaborides of some rare earth elements and yttrium.
Carbon **4**, 483-488, 1966 (No. 4). *A*
- J. Domain:** Contribution à la réalisation d'un laser au néodyme. Etude spectroscopique de l'ion néodyme trivalent dans le tungstate de calcium et dans le verre.
Acta electronica **10**, 71-109, 1966 (No. 1). *L*
- W. Elenbaas:** The influence of cluster formation on the evaporation rate of hot metals.
Philips Res. Repts. **22**, 1-4, 1967 (No. 1).
- W. Elenbaas:** Rate of evaporation and heat dissipation of a heated filament in a gaseous atmosphere. Part II: The tungsten transport through the Langmuir layer.
Philips Res. Repts. **22**, 5-9, 1967 (No. 1).
- J. R. van Geuns:** A study of a new magnetic refrigerating cycle.
Thesis, Leiden 1966. *E*
- H. C. de Graaff:** High-frequency measurements of thin-film transistors.
Solid-State Electronics **10**, 51-56, 1967 (No. 1). *E*
- H. G. Grimmeiss & H. Scholz:** Einige mögliche Bauelemente aus GaP.
Proc. IFAC/IFIP Symp. on microminiaturization, Munich 1965, publ. No. 2.2, 11 p.; Oldenbourg, Munich 1966. *A*
- R. Groth:** Untersuchungen an halbleitenden Indiumoxydschichten.
Phys. Stat. sol. **14**, 69-75, 1966 (No. 1). *A*
- G. J. van Gorp:** The effect of structure on the superconducting properties of vanadium and niobium foils.
Philips Res. Repts. **22**, 10-35, 1967 (No. 1). *E*
- G. J. van Gorp:** On the dynamic intermediate state in superconductors.
Physics Letters **24A**, 528-530, 1967 (No. 10). *E*
- G. J. van Gorp & D. J. van Ooijen:** The influence of dislocations on superconductivity.
J. Physique **27**, Colloque C3, 51-67, 1966. *E*
- C. Haas:** Phase transitions in transition-metal chalcogenides.
Solid State Comm. **4**, 419-421, 1966 (No. 9). *E*
- R. F. Hall & H. C. Wright:** Hall effect measurements on Au-Cs-O.
Brit. J. appl. Phys. **18**, 33-35, 1967 (No. 1). *M*
- E. E. Havinga:** Shell model of interionic interactions for BaTiO₃.
J. Phys. Chem. Solids **28**, 55-64, 1967 (No. 1). *E*
- L. Heijne:** Einige physikalische und chemische Aspekte der Photoleitung.
Festkörperprobleme **6**, 127-173, 1967. *E*
- E. L. Hentley:** Superconducting magnet for an 8 mm travelling wave maser.
Cryogenics **7**, 33-35, 1967 (No. 1). *M*
- F. N. Hooge:** Spectra of praseodymium in yttrium gallium garnet and in yttrium aluminum garnet.
J. chem. Phys. **45**, 4504-4509, 1966 (No. 12). *M*
- D. van Houwelingen, A. L. Luiten & J. Volger:** A superconducting dynamo and solenoid combination operating at high current level.
Bull. Inst. Int. Froid, Annexe 1966-5, p. 545-553. *E*
- G. H. Jonker:** Halogen treatment of barium titanate semiconductors.
Mat. Res. Bull. **2**, 401-407, 1967 (No. 4). *E*
- W. Kebschull:** Leuchtdichteverhältnisse auf feuchten Straßen.
Lichttechnik **18**, 109A-114A, 1966 (No. 9). *A*
- G. Klein & H. Hagenbeuk:** An accurate triangular-wave generator with large frequency sweep.
Electronic Engng. **39**, 388-390, 1967 (No. 472). *E*
- M. Klerk & E. Roeder:** Grain boundary migration in hot-pressed tantalum carbide.
The Electron Microprobe, p. 642-652; Wiley, New York 1966. *A*
- J. T. Klomp:** Solderen in de hoogvacuümtechniek.
Meded.blad Ned. Vacuümver. **4**, No. 4/5, 48-65, 1966. *E*
- H. Klotz:** Verfahren zur autogenen Verschweißung gasgefüllter Keramikampullen.
Vakuum-Technik **15**, 63-66, 1966 (No. 3). *A*
- K. G. Knauff:** Darstellung und Eigenschaften dünner Schichten der Boride von Übergangsmetallen.
Basic problems in thin film physics, Proc. int. Symp., Clausthal-Göttingen 1965, p. 207-211; Vandenhoeck & Ruprecht, Göttingen 1966. *A*
- K. Lagemann:** Das DV-Flipflop, ein neuartiges Schaltglied und seine Vorzüge gegenüber dem JK-Flipflop.
Elektron. Rechenanl. **9**, 9-16, 1967 (No. 1). *H*
- J. Liebertz:** Züchtung von CuCl-Einkristallen aus salzsaurem Lösung.
Phys. Stat. sol. **15**, K 123-125, 1966 (No. 2). *A*
- J. Loeckx & M. Sintzoff:** Les langages de programmation et leur traitement automatique.
Rev. MBLE **9**, 126-146, 1966 (No. 3). *B*

- Mme Lottin & J. Bonnefous:** Les éléments lasers: contrôle des cristaux, finition optique et mesures. *Acta electronica* **10**, 55-70, 1966 (No. 1). *L*
- G. Meijer:** The influence of light on plant development. *Congrès et Colloques Univ. Liège* **38**, 141-149, 1966. *E*
- H. J. G. Meyer, G. Ahsmann & J. W. van der Laarse:** A new class of low-pressure arc columns with positive *V-I* characteristics. *Appl. Phys. Letters* **10**, 124-126, 1967 (No. 4). *E*
- E. J. Millett, J. C. Brice, P. A. C. Whiffin & P. W. Whipps:** Coupled substitution of chromium and lithium in zinc tungstate. *Crystal Growth, Proc. int. Conf. Boston, 1966 (suppl. to J. Phys. Chem. Solids)*, p. 673-677; Pergamon Press, Oxford 1967. *M*
- J. Neiryneck:** The aperiodic step response. *Rev. MBLE* **9**, 147-150, 1966 (No. 3). *B*
- J. A. van Nielen & O. W. Memelink:** The influence of the substrate upon the DC characteristics of silicon MOS transistors. *Philips Res. Repts.* **22**, 55-71, 1967 (No. 1). *E*
- P. Penning:** Theory of X-ray diffraction in unstrained and lightly strained perfect crystals. Thesis, Delft 1966. *E*
- U. Pick:** An improved electron bombardment evaporation source for powdered materials. *J. sci. Instr.* **44**, 70-71, 1967 (No. 1). *M*
- A. Rabenau:** Chemical problems in semiconductor research. *Endeavour* **25**, 158-165, 1966 (No. 96). *A*
- A. Rabenau:** Chemistry of the incandescent lamp. *Angew. Chemie: Int. Edit. in English* **6**, 68-73, 1967 (No. 1); *German Edit.* **79**, 43-49, 1967 (No. 1). *A*
- A. Rabenau, E. Roeder & S. Scholz:** Méthode d'étude du frittage des réfractaires jusqu'à 3000 °C et à 1000 kg/cm². *Rev. int. hautes Temp. Réfract.* **3**, 85-90, 1966 (No. 1). *A*
- J. A. Rietdijk:** The expansion ejector, a new device for liquefaction and refrigeration at 4 °K and lower. *Bull. Inst. Int. Froid, Annexe* 1966-5, p. 241-248. *E*
- C. J. M. Rooymans:** Hochdruck-Hochtemperatur-Umwandlungen der Monotelluride der seltenen Erden. *Berichte Bunsenges. phys. Chemie* **70**, 1036-1041, 1966 (No. 9/10). *E*
- H. J. Schmitt:** Hochfrequenz-Permeabilität von Ferriten bei dynamischer Vormagnetisierung. *Z. angew. Physik* **22**, 95-103, 1967 (No. 2). *H*
- H. Scholz & R. Kluckow:** Temperature-gradient reversal methods for crystal growth. *Crystal Growth, Proc. int. Conf. Boston, 1966 (suppl. to J. Phys. Chem. Solids)*, p. 475-482; Pergamon Press, Oxford 1967. *A*
- G. Schulten:** Microwave optical ring resonators. *IEEE Trans. on microwave theory and techniques MTT-15*, 54-55, 1967 (No. 1). *H*
- G. Schulten & J. P. Stoll:** A note on "Submillimeter wave harmonic mixing". *IEEE Trans. on microwave theory and techniques MTT-15*, 60, 1967 (No. 1). *H*
- G. Söllner:** Blendungsbegrenzung in der Innenraumbeleuchtung. *Techn. Rdsch.* **58**, No. 6, 41-45, 1966. *A*
- B. J. Stocker & G. F. Weston:** The effect of acceleration on a filamentary discharge in inert gases. *J. sci. Instr.* **43**, 913-916, 1966 (No. 12). *M*
- H. D. Stone:** Preparation of high-resolution phosphor screens. *Adv. in Electronics and Electron Phys.* **22A**, 565-570, 1966. *M*
- F. L. Stumpers:** Ir. H. Rinia neemt afscheid. *T. Ned. Elektronica- en Radiogen.* **31**, 245-249, 1966 (No. 12). *E*
- D. G. Taylor:** The measurement of the modulation transfer functions of fluorescent screens. *Adv. in Electronics and Electron Phys.* **22A**, 395-405, 1966. *M*
- B. D. H. Tellegen:** On nullators and norators. *IEEE Trans. on circuit theory CT-13*, 466-469, 1966 (No. 4). *E*
- M. G. Townsend & J. W. Orton:** Spin-lattice relaxation of Rh(II) in single crystals of ZnWO₄. *J. chem. Phys.* **45**, 4135-4140, 1966 (No. 11). *M*
- J. Volger:** Recente toepassingen van supergeleiders. *Natuurk. Voordr.* 1965-1966, No. 44, p. 43-53; *Diligentia*, Den Haag 1966. *E*
- W. L. Wanmaker, A. H. Hoekstra & J. G. Verriet:** The preparation of Ca, Sr, Cd and Mn antimonites. *Rec. Trav. chim. Pays-Bas* **86**, 537-544, 1967 (No. 5).
- H. W. Werner & H. A. M. de Grefte:** Measurement of small ion currents in a mass spectrometer with a scintillation detector. *1965 Trans. 3rd Int. Vacuum Congress, Stuttgart, Vol. 2, Part 2*, p. 493-496; Pergamon Press, Oxford 1967. *E*
- W. J. Witteman:** Rate determining processes for the production of radiation in high power molecular lasers. *IEEE J. of quantum electronics QE-2*, 375-378, 1966 (No. 9). *E*

A fast cryopump system for ultra-high vacuum

A. Venema

Now that pumps which have a high pumping speed at low pressure have been developed it is easier to pump down to ultra-high vacuum, i.e. pressures of the order of 10^{-9} torr. However, the times in which such pressures can be attained — which do not only depend upon pumping speed, but on other factors as well — still leave something to be desired. In the article below a system is described in which, by the appropriate use of low temperature, the pumping time is considerably shortened.

Introduction

In the last ten years the art of achieving ultra-high vacuum, that is to say, pressures of less than 10^{-9} torr, has evolved from a complicated and time-consuming procedure to an established technique. The statement still holds even if the vacuum vessel is extremely large. A typical example of such a case is to be found in experiments designed to simulate the conditions prevailing in outer space. Apart from the classical method of pumping, using a diffusion pump, still widely employed, other methods have gained ground in which use is made of a getter ion pump, or of a cryopump [1]. The chief advantage of these pumps is that they cannot contaminate the space being evacuated, as they do not use a pump fluid. This advance has not however been associated with an appreciable shortening of the time in which the required vacuum is reached. The traditional method of reducing the pumping time is to bake out the system, but this results in a pumping time no shorter than a few hours. For experiments or processes of short duration which have to take place in a vacuum, a much greater reduction of the pumping time would therefore be very welcome. We have found that the pumping time can in fact be substantially reduced by the appropriate use of low temperature.

Generally speaking, lowering the temperature of part of the wall of a vacuum vessel has three effects, each of which causes a drop in pressure. The first is trivial: as the wall becomes colder the temperature of the gas drops and so too does the pressure (Charles's Law). The second effect is connected with the desorption, during evacuation, of gas which is adsorbed on

the wall of the vacuum vessel and absorbed in the wall itself. Reduction of the temperature causes a considerable decrease in the rate of desorption of this gas. The pressure in the last phase of the pumping process, called the end pressure, is determined to a great extent by the competition between this desorption of gas and the performance of the pump. Lowering of the desorption rate and of the diffusion rate therefore leads to a lowering of the end pressure. The third and last effect is found only in gases whose partial pressure is higher than the saturation pressure corresponding to the temperature of the cooled part of the wall. A gas of this kind partly *condenses* on the cold wall, and its partial pressure then approaches the saturation pressure. The latter process, removal of the gas by condensation, is known as *cryopumping*.

The operation of the system described in this article [2], which can be used to achieve an ultra-high vacuum in a short time, is based on a combination of cryopumping and reduction of the rate of desorption. The Charles's Law temperature effect is also present, of course, but, as in other systems, it plays a relatively minor part. The required refrigeration is produced by a Philips two-stage gas-refrigerating machine [3], both

[1] For a survey of modern vacuum techniques see S. Dushman, *Scientific foundations of vacuum technique*, Wiley, New York 1962, or M. Wutz, *Theorie und Praxis der Vakuumentchnik*, Vieweg, Braunschweig 1965. An ultra-high vacuum system using a diffusion pump is described by A. Venema and M. Bandringa, *Philips tech. Rev.* 20, 145, 1958/59.

[2] A. Venema, paper read at 11th A.V.S. Symposium, Washington 1964. See also: R. David and A. Venema, 1965 *Trans. 3rd Int. Vacuum Congress, Stuttgart*, Vol. 2, Part 3, p. 577; Pergamon Press, Oxford 1967.

[3] G. Prast, *Philips tech. Rev.* 26, 1, 1965.

stages of which are used — the first for reducing the desorption rate and the second for cryopumping. The walls to be cooled are in direct mechanical contact with the machine, so that no gas or fluid is necessary for the cold transport. Before describing the system we shall briefly discuss the two temperature effects underlying its operation.

Since low-temperature pumping calls for relatively expensive machines, it will not be the most economic method in all cases. It is undoubtedly an economic proposition for large systems, but for small systems it is only economic in certain cases. For example, a system of the type described in this article is attractive for short term experiments and processes, i.e. cases in which the pumping time figures largely in the costs.

Finally, it should be noted that circumstances may arise where an ultra-high vacuum can be attained quickly only by pumping simultaneously with pumps of different types. We shall not discuss situations of this kind here. Our purpose in the following is simply to show the contribution which the carefully considered use of low temperature can make to shortening the pumping time.

The two fundamental processes

Adsorption and desorption of residual gas

After pumping for some time, a point is reached where the pressure in a vacuum system begins to drop very slowly. The evacuated space then contains a gas mixture, the "residual gas", whose composition generally differs very considerably from that at the beginning of the pumping process, and whose pressure depends on a variety of processes. Chief among these processes is the release of gas adsorbed on the wall, but, as we indicated above, gas in solution inside the walls is also desorbed. Since different gases are adsorbed on the wall, whose desorption depends in different ways on temperature and pressure, the composition of the gas gradually changes during evacuation. If the system is baked out in a furnace, or a filament is heated to incandescence in the evacuated space, the composition may in addition be changed by interactions between the gas and the hot walls or the hot filament, resulting in chemical reactions.

In recent years investigations using mass spectrometers have yielded a great deal of information on the composition of the residual gas, but it is still not possible to give a quantitative prediction of how the total pressure in a particular pumping process will vary with time, particularly when dealing with high or ultra-high vacuum. This is easily understood if one considers that in a sphere of 1 dm³ in which the gas pressure is 10⁻² torr, there are just as many molecules on the wall

as inside the sphere when the wall is covered with only a monomolecular layer. If the pressure in the sphere is reduced to 10⁻⁸ torr the ratio between the number of adsorbed and the number of free molecules is no less than 10⁶ to 1. Obviously, therefore, the gas desorption from the wall has a decisive bearing on the course of the pumping process.

The rate of desorption, i.e. the number of molecules leaving the surface per second and per cm², is given by the expression

$$dN/dt = N/\tau, \quad \dots \dots \dots (1)$$

where N is the number of molecules per cm² of surface and τ is the average time a molecule remains on the surface. This time is given by:

$$\tau = \tau_0 \exp(E_d/RT). \quad \dots \dots \dots (2)$$

In equation (2) τ_0 is a constant, E_d is the desorption energy, R the gas constant and T the absolute temperature.

It is clear from (1) and (2) that a change in the temperature T of the wall from which the gas is desorbed has a marked effect on the rate of desorption. Cooling reduces the desorption rate considerably and thus results in a much lower end pressure.

The speed at which the gases dissolved in the wall reach the surface also decreases considerably as the temperature falls; the reciprocal of the diffusion constant varies with T in much the same way as τ .

By cooling the walls of a vacuum vessel and everything contained inside it to the temperature of liquid nitrogen, it is therefore possible in principle to produce an ultra-high vacuum without the need for any baking out [4].

Of course, in many vacuum systems it is not possible to make all the surfaces cold. The difficulties due to this differ from one case to another, and the right answer has to be found for each individual case. If, for example, a layer has to be deposited on a substrate by vacuum evaporation, the source has to be heated and a great deal of gas is released. The obvious answer here is a drastic increase in the pumping speed, i.e. the ratio of the quantity of gas pumped out per second and the pressure. Here too, as we shall now see, the use of low temperature can be very useful.

Condensation of gas; cryopumping

The saturation pressure of any substance depends to a great extent upon temperature. *Fig. 1* shows the variation with temperature of the saturation pressure of various gases frequently found in vacuum systems. It can be seen that cooling to the temperature of liquid nitrogen (77 °K) is more than sufficient to reduce the vapour pressure of water to a negligible value, but that

at this temperature pressures in the ultra-high vacuum range cannot be reached if gases such as carbon monoxide, carbon dioxide or methane are present. Broadly speaking, such pressures can be reached when cooling to the temperature of liquid hydrogen (20 °K); at this temperature only hydrogen, helium and neon still have too high a pressure. For hydrogen the temperature of liquid helium (4 °K) is still too high; at this temperature hydrogen has a vapour pressure of about 10^{-7} torr. (If these gases are present in a quantity sufficient to upset the experiments, cryopumping must be combined with some other method of evacuation.)

If the pressure of a gas at a certain temperature is equal to the saturation pressure at that temperature, then the number of evaporating molecules is equal to the number of condensing molecules. Since we are concerned here with very low pressures, and the mean free path of the molecules is much larger than the dimen-

second with the wall — and hence with the part whose temperature has been reduced — is given from the kinetic theory of gases, by:

$$v_1 = 3.513 \times 10^{22} p_1 (MT_1)^{-1/2} \dots (3)$$

Here p_1 is the gas pressure in torr, M the molecular weight and T_1 the temperature of the gas, i.e. the temperature of the walls apart from the area at reduced temperature. Not every molecule colliding with this part of the surface condenses, however, although there is a great probability of this happening if the molecule is not travelling exceptionally fast. Let $\alpha_{12}v_1$ be the number of condensing molecules per cm^2 and per second; the factor α_{12} , the fraction that condenses, we shall call the condensation coefficient of a molecule of temperature T_1 which comes into contact with a surface of temperature T_2 . The number of molecules evaporating per cm^2 and per second from the part reduced in temperature is

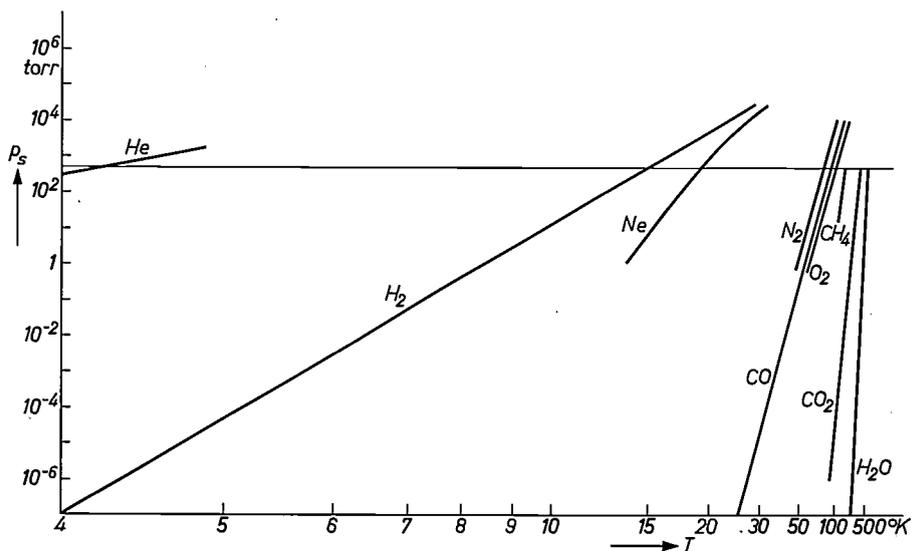


Fig. 1. The saturation pressure p_s as a function of absolute temperature T for various gases frequently found in vacuum systems.

sions of the vacuum system, the temperature of the gas is entirely determined by the temperature of the walls and of the objects contained in the vacuum system. If now the temperature of part of the surface is lowered, so that the number of molecules evaporating from it is reduced, gas will condense there continuously; the part of the wall whose temperature has been reduced then operates as a cryopump.

The pumping speed of such a pump and the lowest attainable pressure can easily be calculated as follows. The number of molecules v_1 colliding per cm^2 and per

equal to $\alpha_{22}v_2$. Here α_{22} is the condensation coefficient for molecules of temperature T_2 that collide with a surface having the same temperature. The number v_2 is found from (3) by substituting for p_1 and T_1 in that expression the values of p_2 and T_2 ; p_2 is the saturation pressure at the temperature T_2 . For every cm^2 of the surface of the cold wall, therefore, there disappear per second from the evacuated space a number of molecules

[4] See also A. Venema, Trans. 2nd Int. Vacuum Congress, Washington 1961, Pergamon Press, Oxford 1962, Vol. 1, p. 1.

equal to $\alpha_{12}v_1 - \alpha_{22}v_2$. If this number is converted into practical units, the pumping speed S per cm^2 of cold surface is found to be:

$$S = 3.64 \alpha_{12} \left(\frac{T_1}{M}\right)^{\frac{1}{2}} \left\{ 1 - \frac{\alpha_{22}}{\alpha_{12}} \frac{p_2}{p_1} \left(\frac{T_2}{T_1}\right)^{-\frac{1}{2}} \right\}. \quad (4a)$$

The lowest pressure attainable is the pressure at which S has decreased to zero. This value of p_1 , usually denoted by p_{eq} , is given by:

$$p_{\text{eq}} = \frac{\alpha_{22}}{\alpha_{12}} p_2 \left(\frac{T_2}{T_1}\right)^{-\frac{1}{2}}. \quad (5)$$

As can be seen, p_{eq} is *not* equal to the saturation pressure p_2 that corresponds to the temperature T_2 of the pumping plate. This is because only part of the wall has the temperature T_2 .

With the aid of equation (5) the equation for the pumping speed can be reduced to:

$$S = S_0 (1 - p_{\text{eq}}/p_1), \quad (4b)$$

where $S_0 = 3.64 \alpha_{12} (T_1/M)^{-1/2}$. This expression (see *fig. 2*) shows rather more clearly than (4a) the way in

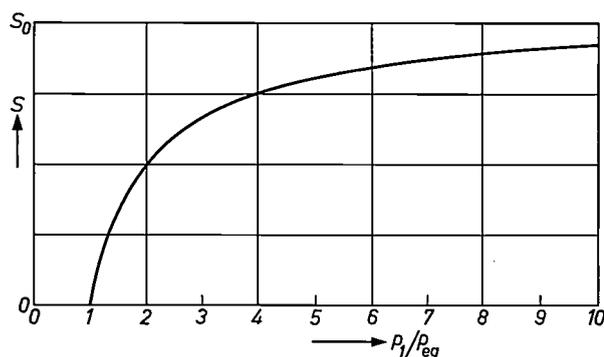


Fig. 2. Variation of pumping speed S with pressure p_1 in a vacuum chamber (equation 4b). The lowest pressure (p_{eq}) is reached when S is zero. At pressures which are high compared with p_{eq} , S is nearly independent of p_1 and approximately equal to S_0 .

which the pumping speed drops to zero during evacuation in a particular case (i.e. at a given S_0). The advantage offered by cryopumping in terms of gas kinetics also becomes clear: if the "temperature" of the molecules is not too far above that of the pumping surface, then α_{12} is roughly unity, which is 4 to 5 times greater than the constant, corresponding to α , occurring in the expression for the pumping speed of a well-designed diffusion pump.

A vacuum system with a two-stage gas-refrigerating machine

The heart of our vacuum system for rapid production of ultra-high vacuum is the two-stage gas-refriger-

ating machine mentioned earlier. In the design of the system good use was made of the ability of this machine to supply cold at two temperatures, the cold production at the lowest temperature (about 12 °K) being only slightly influenced by the cold production at the other temperature (50 °K to 80 °K.)

A diagram of the whole system is shown in *fig. 3*. The space to be evacuated is contained inside the copper housing W_1 , mounted on the freezer F_2 of the first stage. The head of the machine, with freezer F_1 , is inside this housing. Fitted to the top of F_1 is a copper plate C with a surface area of about 80 cm^2 , which functions as the cryopump. To prevent the condensation of atmospheric vapours and gases on the outside wall of the vacuum chamber — this would give a loss of cold and could cause short-circuiting between electric leads through the wall — the whole system is contained inside a much larger vacuum vessel (wall W_2) in which a pressure of 10^{-3} to 10^{-5} torr prevails. This pressure can easily be obtained by means of a rotary oil pump

Since there are advantages in putting the cryopump into operation only when the pressure is already fairly low, the vacuum chamber can be connected by a pipe P to another type of pump, for example a rotary pump or a combination of a sorption pump using zeolite [5] and a diffusion pump, or a getter ion pump. In addition, the wall W_1 is fitted with a number of flanges (not shown in the diagram) which can be used to connect an ionization gauge for measuring the total pressure, to connect instruments for measuring the partial pressures, for feed-through of electrical leads, etc.

Since, after completion of an experiment, a vacuum vessel of this kind can be opened only when the temperature has again become roughly equal to room temperature, the outside of the wall W_1 is fitted with an electric heating element to raise the temperature quickly. This element can, if necessary, also be used for baking out the wall to some extent before or during the first phase of the pumping process. Another method of heating, which is undoubtedly attractive, is to let the gas-refrigerating machine itself produce the heat. In principle, this only requires the direction of rotation to be reversed. With the machines at present on the market, however, this would involve some slight modifications.

Experience gained with the system

The experience gained with the system has completely fulfilled our expectations. Starting from atmospheric pressure, a vacuum of less than 10^{-9} torr can be reached within about an hour. This result will not, of course, be achieved easily in all circumstances. It is necessary to take account of the particular characteristics of the instruments arranged inside the vacuum space and of

the pressure gauges connected to it, and suitable measures must be taken to offset any adverse effects they may have. To some extent, these are precautions which have to be taken in any ultra-high vacuum system, but they are more important here owing to the shortening of the time scale. These precautions are also applied because the gas cannot be completely evacuated from a

gauge. After out-gassing the pressure again dropped quickly, and in a good hour a pressure as low as 10^{-10} torr was reached.

Another feature of the system came to light in experiments designed to simulate the gas yield due to desorption from the wall; this was done by admitting a gentle stream of nitrogen into the vacuum space while

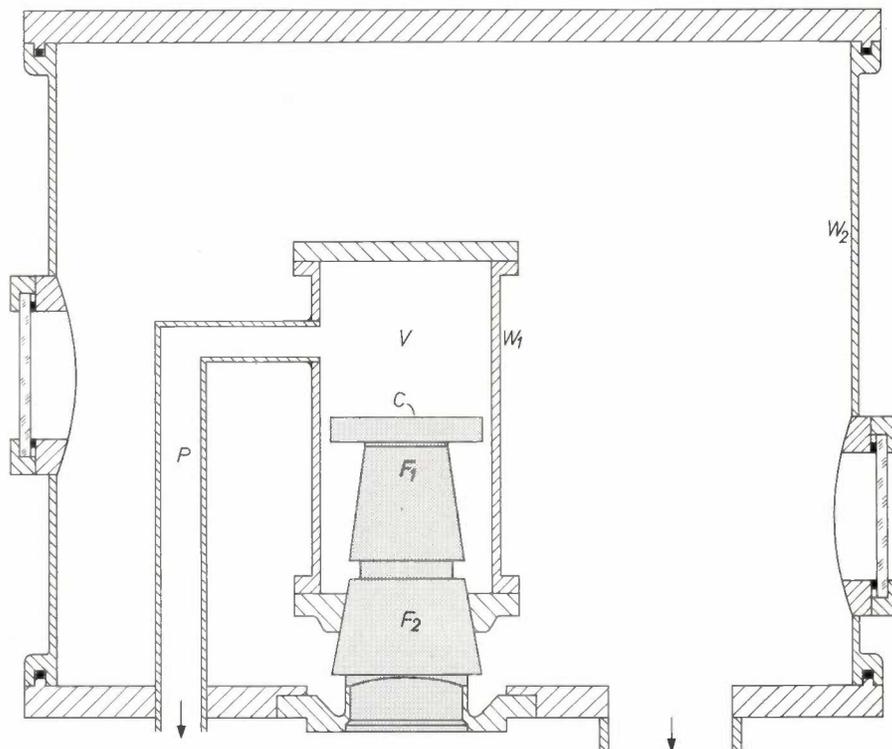


Fig. 3. Diagram of the cross-section of the fast cryopump system for ultra-high vacuum. *V* vacuum chamber with copper wall *W*₁. *F*₂ freezer of the first stage of a two-stage gas-refrigerating machine. *F*₁ freezer on the head (second stage) of the refrigerating machine surmounted by copper plate *C*, acting as the cryopump. At the beginning of the pumping process a diffusion pump is used, which is connected to *V* by pipe *P*. The complete assembly is enclosed in a second vacuum chamber (outer wall *W*₂) in which a low vacuum is maintained.

system like the one illustrated in fig. 3, some gas always remaining on the wall.

One of the most troublesome components of the system itself is the ionization gauge. Its hot filament gives off gas and moreover heats the wall, giving rise to some unwanted desorption. This makes it necessary to out-gas the gauge and everything near it. The importance of this can be clearly understood from fig. 4. In the experiment which gave these results the refrigerating machine was connected in after a pressure of 7×10^{-5} torr had been reached with an auxiliary pump. The pressure then dropped within half an hour to about 10^{-8} torr, but from then on the pressure dropped very slowly because of the release of gas from the ion

pumping. The end pressure established in these experiments was 10^{-6} to 10^{-7} torr. When the supply of nitrogen was stopped, the pressure dropped much more slowly than one would have expected. The reason proved to be that the admitted nitrogen was quickly adsorbed on the wall and was gradually desorbed when the supply was stopped; the temperature of the wall in our equipment (about 55 °K) is apparently not low enough

[5] Zeolites are Al-Mg silicates which, upon heating, lose their water of crystallization without any change taking place in their crystal structure, giving a molecular-porous material. This material has an extremely large internal surface area and can bind a great deal of gas. Since this takes place both by adsorption and absorption, the term sorption pumps is used, which also distinguishes these pumps from getters which trap gases *chemically*.

to bind nitrogen drastically and not high enough for rapid desorption of nitrogen. This is the very same effect which is found with water vapour in systems at room temperature. This difficulty can easily be overcome by increasing the temperature of the wall by 5 to 10 °K. The nitrogen is then released more quickly. This sends the pressure up for a little while, but after this it falls rapidly.

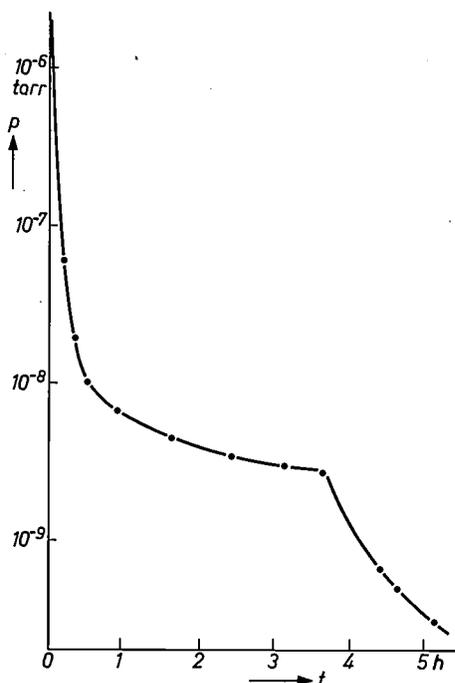


Fig. 4. The effect of out-gassing the ion gauge on the pressure p that can be reached in a time t . Before out-gassing, the pressure after 3½ hours had dropped to a value no lower than about 4×10^{-9} torr. After out-gassing, the pressure was ten times lower within 1½ hours.

When an evaporation process is carried out in the vacuum chamber a considerable amount of heat is generated which reaches the wall as radiation. How does this affect the temperature of the surface acting as the cryopump? The situation is relatively favourable, since the refrigerating capacity of the gas-refrigerating machine is already fairly high at temperatures only just above the minimum temperature (as much as 85 W at 20 °K). Even so, it is more advantageous to dissipate the entire radiant energy through the first stage of the machine. For if the pumping surface becomes a few degrees warmer, this can have a fairly considerable effect on the pressure if the residual gas has a component whose partial pressure is close to the saturation pressure. A slight increase of temperature in the first stage of the machine, on the other hand, has hardly any effect on the temperature of the pumping surface

and cannot therefore seriously affect the pumping speed. The pumping surface can be very effectively screened by means of a grid connected to the wall of the vacuum chamber and directly connected to the first stage of the machine. A grid of this kind, with a cross-section like that shown in *fig. 5*, lets gas through readily but not the radiation. It does, of course, reduce the pumping speed slightly.

We may note in passing that the refrigerating capacity of the machine at, say, 20 °K, is generally much greater than the heat of condensation from the pump plate plus the energy of the incident radiation. We might, therefore, have made the pump plate and hence the pumping speed much larger. For our purpose however, this was not necessary. In fact, in systems for carrying out certain heat-evolving processes under vacuum a large pump plate can be undesirable. It is better then to keep a fair amount of refrigerating capacity "in reserve".

If required, it is possible, using a two-stage gas-refrigerating machine, to build an ultra-high vacuum system of an entirely different type, leaving the walls of the vacuum chamber at room temperature. A diagram of the cross-section of such a system is shown in *fig. 6*. The cryopumping surface is surrounded on all sides by a screen of the type illustrated in *fig. 5*, which is fitted to the first stage of the machine. In a system of this kind a substantially higher pumping speed can be reached than is possible with the pumping methods used hitherto. Very large vacuum systems using a cryopump are nearly always designed on the lines illustrated in *fig. 6*; the screen is usually cooled with liquid nitrogen, and the pumping surface with liquid hydrogen or helium.

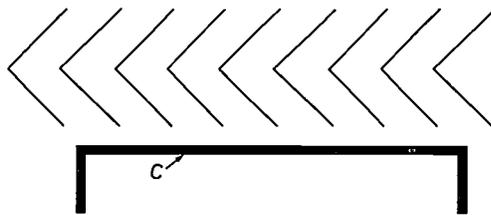


Fig. 5. Diagram of the cross-section of a grid of copper angle strips for screening the pump plate C (*fig. 3*) from heat radiation. The screen presents no obstruction to the gas.

Arrangement for measuring condensation coefficients

To conclude, as an example of a special application of our vacuum system, we shall briefly describe an arrangement for measuring condensation coefficients (*fig. 7*). The arrangement consists basically of two vacuum chambers I and II , interconnected by means of a circular hole of known diameter (conductivity F). Immediately above the plate C , which operates as the cryopump, there is a diaphragm, and this means that the area of the pumping surface is also exactly known. In the equilibrium state the quantity of gas flowing per

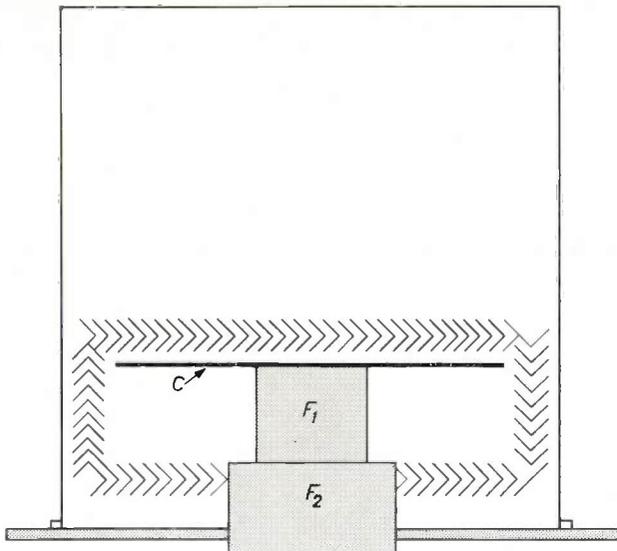


Fig. 6. Schematic diagram of a vacuum system in which the walls are at room temperature. The pump plate *C* is surrounded on all sides by screens to protect it from heat radiation. *F*₁ and *F*₂ have the same significance as in fig. 3.

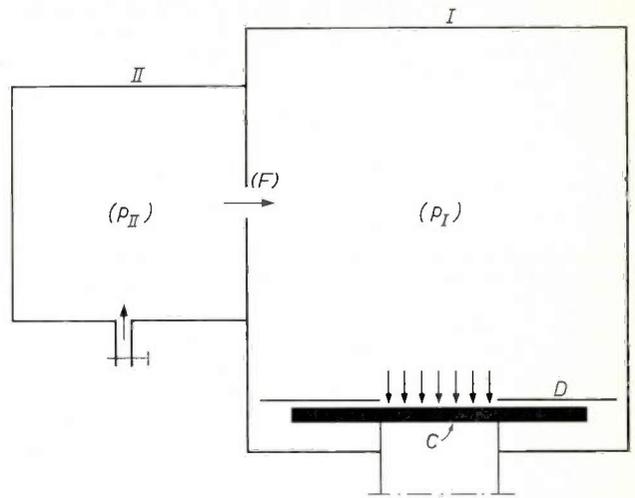


Fig. 7. Arrangement for the measurement of condensation coefficients. In the equilibrium state the gas flow from chamber *II* to chamber *I* through an orifice of conductivity *F* is equal to the quantity condensing on the pump plate *C*, the effective surface area of which is determined by the hole in diaphragm *D*.

second from *II* to *I* is equal to the quantity condensed per unit time on the pumping plate *C*. Expressed as an equation:

$$(p_{II} - p_I)F = p_I S; \quad \dots \quad (6a)$$

hence:

$$S = F \left\{ \left(\frac{p_{II}}{p_I} \right) - 1 \right\}. \quad \dots \quad (6b)$$

Using equation (6b) the pumping speed can be directly derived from the ratio of the pressures *p*_{II} and *p*_I. From the value of *S* then found the condensation coefficient can be calculated directly. In this way we found a coefficient of 0.98 for the condensation of nitrogen at 60 °K on a surface at 18 °K.

Summary. Now that pumps are available which have a high pumping speed at low pressure, it is no longer particularly difficult to produce an ultra-high vacuum (< 10⁻⁹ torr). The pumping time can be drastically shortened by cooling the wall of the system to about 50 °K. This minimizes the rate of gas desorption. The Philips two-stage gas-refrigerating machine is particularly useful in such a system. The wall is then mounted on the first stage. The head of the machine (at about 12 °K) is inside the vac-

uum and on top of it there is a copper plate which acts as a cryopump, i.e. it lowers the partial pressure of the gas components to a value slightly above their saturation pressure at 12 °K. With an arrangement of this kind 10⁻⁹ torr can be reached in one hour. If the system contains an internal heating source, as in a vacuum-evaporation equipment, a grid is arranged in front of the plate to shield it from radiation: the heat received by the grid is taken off through the first stage of the machine.

Measurement of the solderability of components

In the modern manufacture of electronic equipment large numbers of connections are often simultaneously soldered by dipping printed wiring panels into a bath of molten solder for a few seconds. Since the reliable operation of electronic equipment depends on the quality of the soldered joints, it is important to test the solderability of the components by taking samples at random. Once the process conditions such as temperature, flux and solder have been established, the solderability is finally governed by the nature of the surface to be soldered.

A criterion for the solderability of a surface is the way in which it is wetted by the solder. The methods used up to now for measuring the degree of wetting [1] all have one or more disadvantages that make them rather unsuitable for quality control in the works. Although the split-globule method earlier described in this journal [2] can be used for this purpose, its use is limited to tests on wire of round section.

The device described here makes it possible to record the wetting process quickly and simply, thus providing a figure of merit which is closely connected with practical soldering. The test also provides an idea of the wetting rate which is also of great importance. For economic reasons and on account of technological limitations such as the limited ability of components and insulating material to withstand high temperature, the dipping time should be very short.

The method is based on the following effect, previously used by Earle [3].

When a sample is partly immersed in molten solder (*fig. 1*) a quantity of solder is displaced. As well as the solder displaced by the part of the sample below the surface level, an additional quantity of solder is displaced by the action of surface tensions (the surface tension of solid metal, of liquid solder and of the metal-solder interface). If the sample is solderable, it becomes

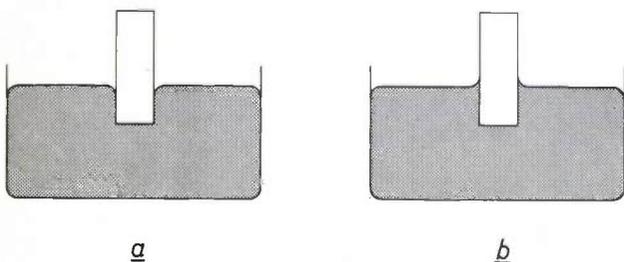


Fig. 1. If a sample is not wetted when it is immersed in a bath of solder, an additional quantity of solder is displaced, thereby increasing the upward force (*a*). When the sample is wetted, a "collar" of solder forms whose weight exerts a downward force (*b*).

wetted shortly after immersion. When equilibrium is established between the surface tensions, the solder creeps upwards, forming a "collar". During immersion the change in the total force acting upon the sample is measured, and the original additional upward force is seen to give way to a downward force — the weight of the collar of solder. The time which this takes, together with the final weight of the collar of solder, gives the required measure for the solderability.

The test is performed with the device illustrated in *fig. 2*. The sample, coated with flux, is attached to the arm *A* of a balance. The bath of solder *B* can be raised until the sample is immersed to an adjustable depth in the solder. (The sample could for example be a tag with a rectangular cross-section of 0.4×1.0 mm, with immersion depth between 3 and 5 mm.) The force acting on the sample as a result of the wetting is about 100 mg.

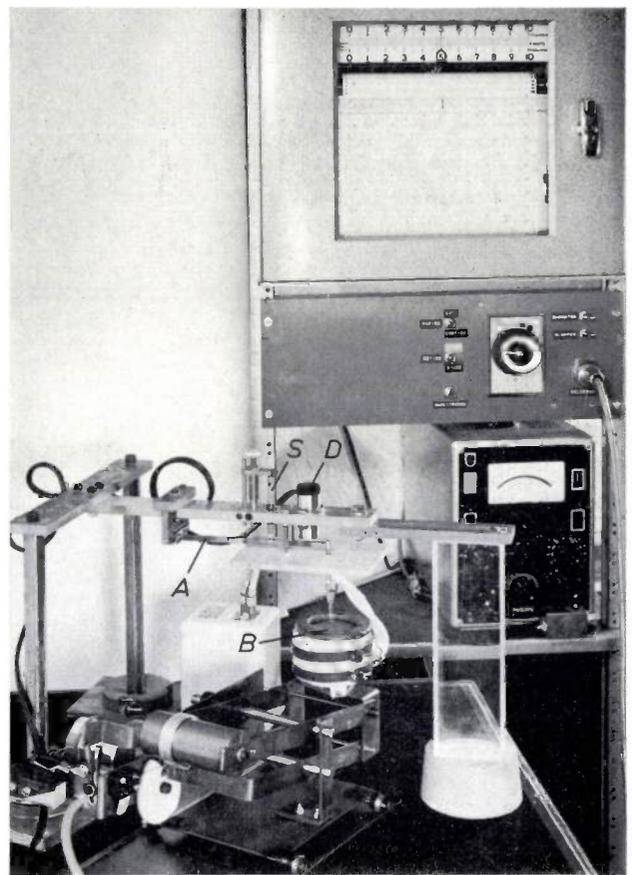


Fig. 2. The dipping apparatus. *A* balance arm. *B* solder bath with sample above it. *D* displacement gauge (type PR 9310/03). *S* is a spring for compensating the weight of the arm and sample, and which also provides initial pressure for the measuring pin of the gauge. Mounted in the rack are the bridge for the displacement measurement (type PR 9303) and a pen recorder type PR 2210 (paper speed 1 cm/s).

The change in the force during the dipping process is measured by the displacement of the balance arm. The displacement must only be very slight as it affects the upward force. The very small displacement can be measured with sufficient accuracy by a displacement gauge D . The initial force upon the measuring pin of

the opposite extreme a case in which no wetting takes place. In *fig. 4* these two cases are combined with a number of intermediate ones for samples of identical shape.

Information about the solderability is obtained not only from the behaviour as the sample is immersed but

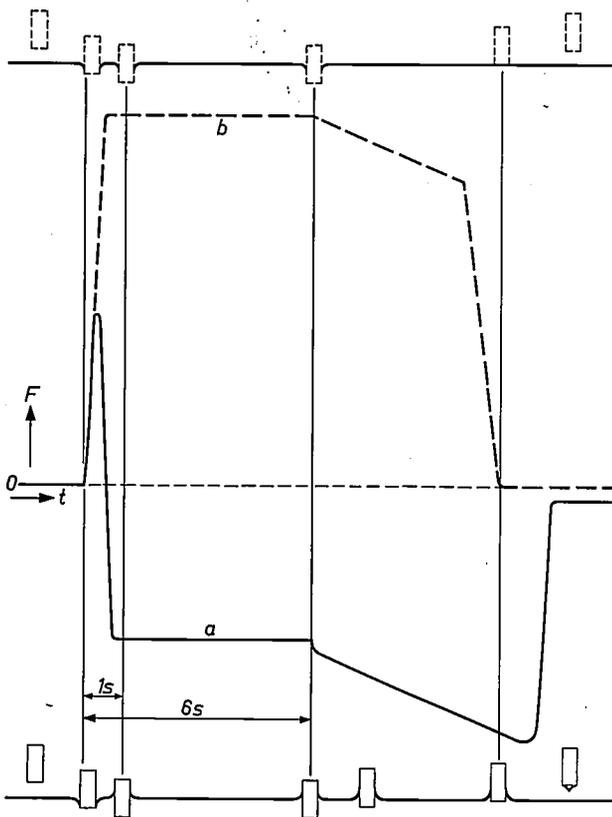


Fig. 3

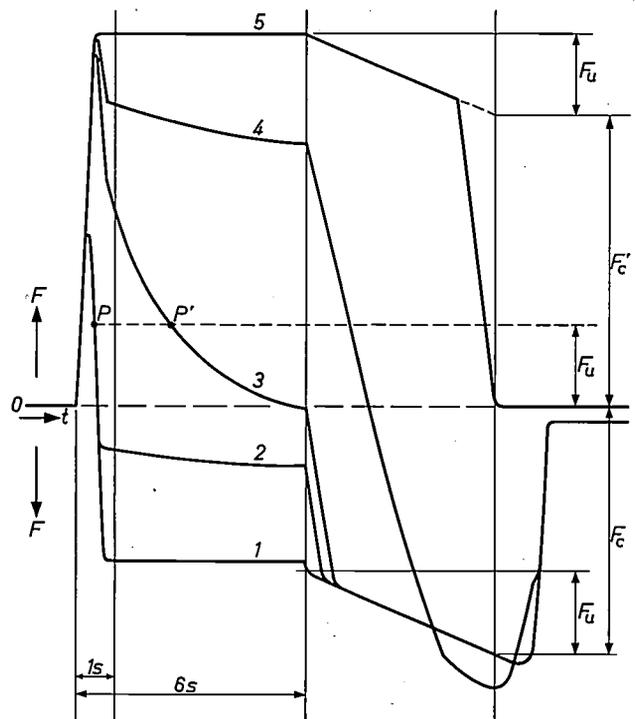


Fig. 4

Fig. 3. Wetting diagrams (schematic) with corresponding dipping situations (in reality the soldering bath is moved and the sample remains almost stationary). *a*) Very good wetting. *b*) No wetting.

Fig. 4. Wetting diagrams (schematic) for a number of differently wetted samples. Diagrams 1 and 5 are the extreme cases which have already been shown in fig. 3. Diagrams 2, 3 and 4 are intermediate cases. F_u is the upward force due to the hydrostatic pressure. F_c is the downward force (the weight of the collar of solder) upon wetting; F_c' is the upward force of the extra quantity of solder displaced when no wetting takes place. It can be seen that for the small samples, to which these diagrams relate, the weight of the collar of solder may exceed the upward force.

the gauge (20 g) and the force compensating for the weight of arm A and the sample are provided by a weak spring S . In this arrangement a difference in force of 100 mg in our example causes a displacement of the sample by 0.008 mm, which is negligible compared with the immersion depth. The amplified signal from the displacement gauge is fed to a fast pen recorder.

Fig. 3, curve *a*, shows a schematic diagram of the complete dipping process for a sample whose solderability is very good. First there is an upward force equal to the weight of the total quantity of solder displaced; this is followed by a rapidly increasing downward force due to wetting. Curve *b* in fig. 3 shows as

also as the bath descends. As a result of the inertia of the fluid as it flows back, the collar of solder is somewhat higher during the descent of the bath than it is in the stationary state, and the effect of this is an initial sharp increase in the downward force. Assuming that the collar of solder then remains unaltered, the time-linear increase of the downward force is a simple function of the decreasing depth of immersion. This

- [1] B. Keysseltz, *Metall* 18, 816-820, 1964.
 R. Barrie, *Mullard tech. Comm.* 9, 2-9, 1966.
 C. J. Thwaites, *Publ. 358 Tin Research Inst.*, 1964.
 W. B. Harding, *Plating* 52, 971-981, 1965.
 [2] J. A. ten Duis, *Philips tech. Rev.* 20, 158-161, 1958/59.
 [3] E. G. Earle, *J. Inst. Metals* 71, 45-72, 1945.

linear region can therefore be used to derive the constant upward force F_u due to hydrostatic pressure, which is operative during the stationary part of the dip cycle. If we now mark off a distance corresponding to F_u above the zero (dashed line) we find the point P (P') at which the angle between the solder surface and the

tic differences. In such cases the wetting diagram should be compared with a standard diagram of a readily-solderable sample of the same shape.

Practical examples of wetting diagrams can be seen in *fig. 5*. Both diagrams are for brass tags which have been electroplated with a tin-lead alloy. The wetting

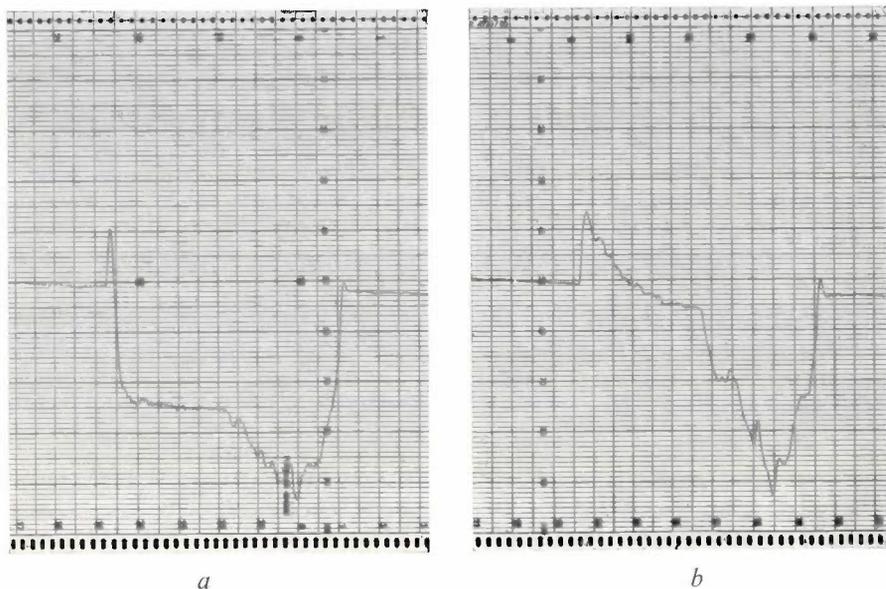


Fig. 5. Two wetting diagrams obtained in practice for brass soldering tags electroplated with a tin-lead alloy.

a) Good wetting. The curve obtained as the bath is removed is not very smooth because of roughness at the surface due to diffusion of zinc from the interior.

b) The wetting is found to be insufficient owing to prolonged storage at room temperature.

surface of the sample is exactly 90° . Point P is particularly useful for judging the scatter in the behaviour data of a number of samples of the same kind.

In order to be able to complete the cycle uniformly, the movement of the soldering bath is programmed by means of a cam disc. The bath rises quickly (at about 50 mm/s), remains stationary for about six seconds and then descends slowly (at 1 mm per second).

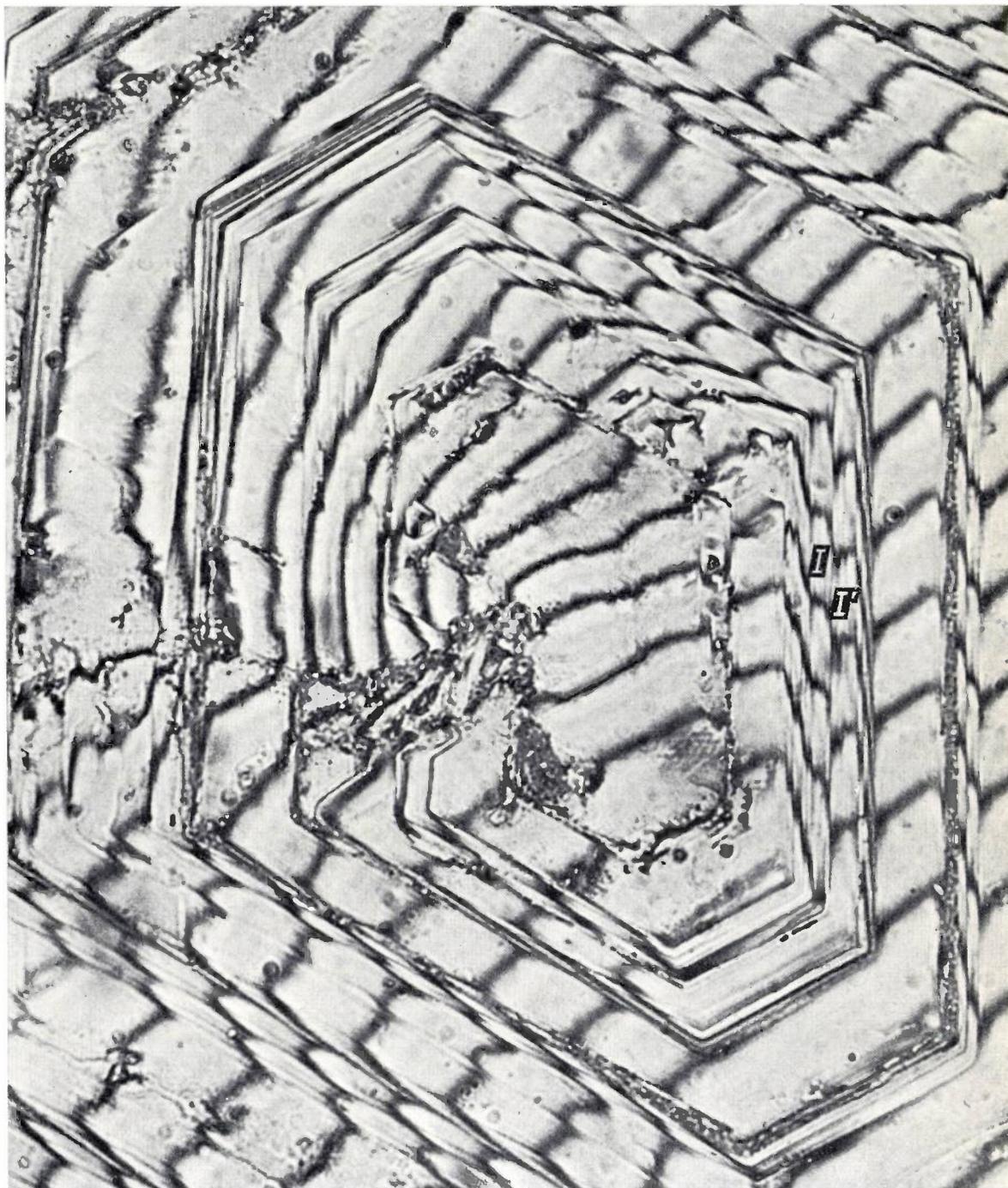
Diagrams like those of *fig. 4* can be expected only for the immersion of samples of simple shape. Protrusions, sharp edges and holes give rise to characteris-

tic behaviour, however, is quite different. The tag of *fig. 5b* came from a batch in which the solderability of the tin-lead plating had deteriorated through prolonged storage at room temperature.

J. A. ten Duis
E. van der Meulen

J. A. ten Duis and E. van der Meulen are with the Philips Electronic Components and Materials Division (Elcoma), Eindhoven.

Crystal growth of ferroxplana



Growth spiral on the surface of a crystal of ferroxplana ZnY ^[1], made by a method employed at the Philips laboratories in Eindhoven and Hamburg (magnification $500\times$). The hexagonal symmetry of the crystal is beautifully demonstrated by the spiral.

It is known that growth spirals frequently occur at screw dislocations. Contrary to what one might expect in such a case, the centre of the growth spiral shown here lies deepest, i.e. at the bottom of a pit.

The photomicrograph, obtained from the work of C. J. M. Rooymans, J. A. Schulkes and A. J. G. Op het Veld, was made with monochromatic illumination (wavelength $\lambda = 537\text{ nm}$) in

a microscope, by the multiple interference method ^[2]. The height of each "growth step" can be deduced from the displacement of the fringes at the step: for a displacement of one fringe spacing this height is $\frac{1}{2}\lambda$. The growth step between *I* and *I'* is therefore about 135 nm high, corresponding in this case to about 30 unit cells of the crystal structure.

^[1] G. H. Jonker, H. P. J. Wijn and P. B. Braun, Philips tech. Rev. **18**, 145-154, 1956/57.

^[2] See W. Dekeyser and S. Amelinckx, Les dislocations et la croissance des cristaux, Masson, Paris 1955.

The skin effect

III. The skin effect in superconductors

H. B. G. Casimir and J. Ubbink

In parts I and II of this article we studied the skin effect in ordinary metals possessing finite conductivity [1]. In this third part we shall consider the skin effect in metals that are in the superconducting state [2]. The superconducting state is limited to temperatures and magnetic fields that do not exceed certain critical values. Between the critical field H_c and the temperature T there exists the empirical relation:

$$H_c/H_{c0} = 1 - (T/T_c)^2, \quad \dots \quad (1)$$

where H_{c0} is the critical field at $T = 0$ and T_c the critical temperature at $H = 0$.

The most striking property of a superconductor is its zero resistance to direct current; not merely extremely low but really zero, as is evident from the existence of persistent currents. From the equation for the classical skin depth (I,10), $\delta_k^2 = 2/\omega\mu\sigma$, one might at first be inclined to think that the supercurrent would be a purely surface current: with $\sigma \rightarrow \infty$ one finds $\delta_k = 0$. This was the main line of thought before 1933. In that year Becker, Heller and Sauter [3] pointed out that a current layer of *finite* thickness is arrived at if the inertia of the electrons is taken into account. In fact their reasoning corresponds in broad lines to that given in II for deriving the skin depth in the relaxation limit (in regions B and E in part II), and which therefore leads to the same result: if all electrons are superconducting, the penetration depth is $\lambda_p = \sqrt{(m/\mu n e^2)}$, cf. equation (23) in part II. Both cases involve electrons which are entirely free but possess inertia. Both constitute the limiting case for $\omega\tau \rightarrow \infty$; in the extreme relaxation case, however, the emphasis is placed on $\omega \rightarrow \infty$ for moderate τ , whereas in the reasoning of Becker, Heller and Sauter it is σ , and hence τ , that tends to infinity while ω remains small.

All that one can really conclude from the reasoning of Becker, Heller and Sauter is that a field *variation* cannot penetrate beyond a depth λ_p . The fact that also a *constant* magnetic field is expelled from the interior of a metal when it changes to the superconducting state (i.e. the Meissner effect [4], 1933), induced the brothers London [5] (1935) to postulate *ad hoc*, in addition to the equation giving the acceleration of an entirely free elec-

tron in an electric field, a new equation for the current. In terms of the skin effect their hypothesis might be formulated by saying that a superconductor, *even at zero frequency*, behaves like a normal metal in the relaxation limit. Thus, even at $\omega = 0$, they find a penetration depth equal to the relaxation skin depth λ_p if all electrons are superconducting; if, more generally, n_s is the concentration of the superconducting electrons, the "London penetration depth" is $\lambda_L = \sqrt{(m/\mu n_s e^2)}$. The fact that this is a new hypothesis, and cannot be deduced by taking the limit for $\sigma \rightarrow \infty$ of a normal conductor, is immediately apparent from *fig. 1*, an ω - τ diagram like *fig. 11* in part II. For $\tau \rightarrow \infty$ at low frequencies we do not arrive in the relaxation region B or E but in the region of the anomalous skin effect D.

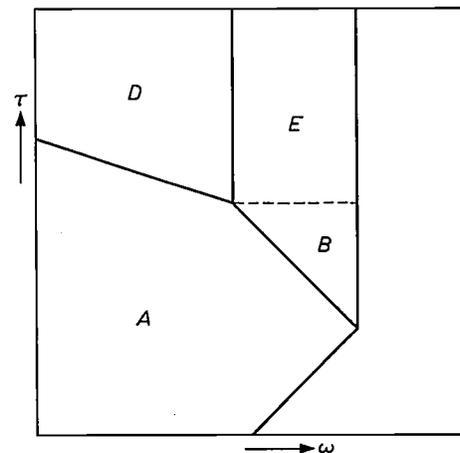


Fig. 1. The ω - τ diagram for a normal conductor, with the region of the normal skin effect (A), relaxation (B and E) and the anomalous skin effect (D). See also II, fig. 11.

Supplementing London's theory with the two-fluid model of Gorter and Casimir [6] (1934), we obtain the picture of superconductivity already outlined in the introduction to part I: there are "superconducting electrons" in addition to "normal electrons" in a ratio which is governed by the temperature. The superconducting electrons have the effect of screening off electromagnetic fields, even at zero frequency. At zero frequency they short-circuit any electric field, but at non-zero frequencies there exists, owing to the inertia

Prof. Dr. H. B. G. Casimir is a member of the Board of Management of N.V. Philips' Gloeilampenfabrieken; Dr. J. Ubbink is with Philips Research Laboratories, Eindhoven.

of the electrons, an electric field in the penetration layer which gives rise to energy dissipation via the normal electrons. This model served its purpose for a long time. Although it has since been superseded by more comprehensive theories and now possesses little more than historical interest, it is still a useful guide in classifying quite a number of phenomena associated with superconductivity.

In discussing the skin effect we shall use this model (the "GCL model") as a guide. In this discussion there are two dominant aspects: the *thickness of the penetration layer* and the *magnitude of the absorption*. These are both comprised in the *complex skin depth* $\delta = \delta' + j\delta''$, introduced in part II. Under the conditions $\delta' \ll \delta'' \ll c/\omega$ — implying that the absorption is low and the penetration depth small compared with the wavelength in free space — the penetration depth is equal to δ'' and the absorption is given by δ' (see II, 14). The results of absorption measurements are usually given in terms of the "surface-resistance ratio" $q = R/R_n = \delta'/\delta_n'$; here the subscript n refers to the metal just above the critical temperature. In the following we shall briefly describe a few methods of measuring both the penetration depth and the absorption, and we shall use the results to test the GCL model.

In broad lines the GCL model gives a reasonable description of the experimental results. Pippard [7], however, found a marked discrepancy: he discovered that the skin depth is dependent on the mean free path (the degree of impurity), whereas according to the GCL model the skin depth ought to be determined solely by the concentration of the superconducting electrons. This, together with other considerations, led Pippard to propose a modification of London's theory, along the lines suggested by the theory of the anomalous skin effect: he argued that it was necessary to take into account the possibility that the relation between current and field is not a purely local one.

To give a rough idea of this line of thought, we shall first reconsider for a moment the possible situations in a normal conductor (fig. 2, see also fig. 2 in part II). We can distinguish between three limiting cases, depending on whether l , v/ω , or δ is much smaller than the other two quantities. In the normal skin effect (A), l is the smallest of the three; with extreme relaxation (B, E), the smallest is v/ω , and with the anomalous skin effect (D) it is δ . Provided $\delta' \ll \delta''$, we may take the quantity $-1/\delta^2$ as a measure of the "screening". In B and E $-1/\delta^2$ has the value $1/\lambda_p^2$. Disregarding details, the results of II for A and D can now be obtained by saying that, coming from B, E, the screening $-1/\delta^2$ (or, rather, its modulus) is reduced in A by a factor $l/(v/\omega)$ and in D by a factor $\delta/(v/\omega)$.

As already noted, the London superconductor is

closely related (even at zero frequency) to the normal conductor in the relaxation limit B, E. The screening ($-1/\delta^2$) is $1/\lambda_L^2$. If, however, l or the penetration depth is too small, the latter is no longer equal to λ_L . We introduce λ_s to denote, quite generally, the penetration depth [8]. With what are we now to compare l or λ_s in this case, at zero frequency? The quantity v/ω is

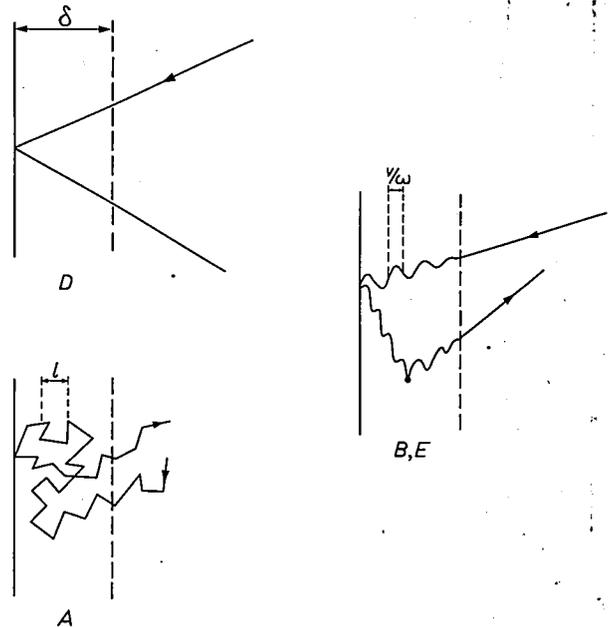


Fig. 2. In a normal metal the various limiting cases (see fig. 1) can be characterized by indicating which of the three lengths l , δ and v/ω is the shortest. In A it is l , in B and E it is v/ω and in D it is δ .

obviously ruled out. As a parallel for v/ω we now have a new length ξ_0 , introduced by Pippard, called the coherence length of the clean superconductor. If $\xi_0 \ll l$ and $\xi_0 \ll \lambda_s$, the superconducting electrons have their full screening effect (fig. 3, B') as predicted by London. In a very "dirty" superconductor ($l \ll \xi_0$ and $l \ll \lambda_s$) the coherence is "curtailed" to l and the screening effect

- [1] H. B. G. Casimir and J. Ubbink, The skin effect, I. Introduction; the current distribution for various configurations, II. The skin effect at high frequencies, Philips tech. Rev. 28, 271-283 and 300-315, 1967 (Nos 9 and 10). These parts are referred to as I and II, and the equations as (I,1), etc.
- [2] A general survey of superconductivity will be given in an article by J. Volger, shortly to appear in this journal.
- [3] R. Becker, G. Heller and F. Sauter, Z. Phys. 85, 772, 1933.
- [4] W. Meissner and R. Ochsenfeld, Naturwiss. 21, 787, 1933.
- [5] F. London and H. London, Proc. Roy. Soc. A 149, 71, 1935; Physica 2, 341, 1935.
- [6] C. J. Gorter and H. B. G. Casimir, Physica 1, 306, 1934; Phys. Z. 35, 963, 1934; Z. techn. Phys. 15, 539, 1934.
- [7] A. B. Pippard, Proc. Roy. Soc. A 216, 547, 1953.
- [8] To link up with Part II it would be consistent to represent the penetration depth by δ'' . We prefer here, however, to follow the practice in the literature of superconductivity, where the penetration depth is denoted by λ with a subscript.

is reduced by a factor l/ξ_0 (fig. 3, A'). The third case (fig. 3, D') is the one in which λ_s is much smaller than ξ_0 and l . Compared with B' the screening is now reduced by a factor λ_s/ξ_0 . A classification of superconductors made on this basis, i.e. according to the values of ξ_0 and l , will be found in figs. 12 and 13.

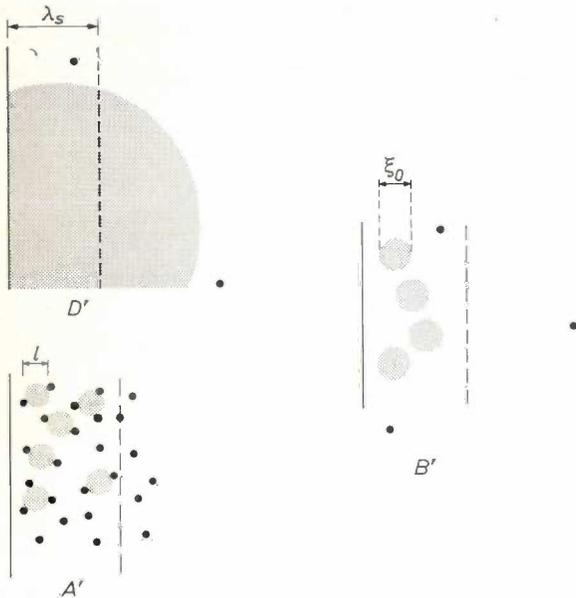


Fig. 3. By analogy with fig. 2 for the normal metal, three limiting cases A', B' and D' can be distinguished in a superconductor, depending on which of the three lengths l , ξ_0 or λ_s respectively is the shortest. ξ_0 is Pippard's coherence length, which is the formal analogue of v/ω in the normal metal. The shaded regions indicate regions of coherence.

The coherence length ξ_0 , which also appears in the theory of superconductivity in other ways with various interpretations, plays in Pippard's theory the role of the distance over which the field at a given point influences the current in the environment. The situation D' has a typical non-local character: the current density in the middle of the skin layer is partly determined by the (stronger) field at the surface and the (weaker) field in the interior of the metal. This situation is characteristic of nearly all elementary metals that become superconductive.

Finally, it may be asked how Pippard's modification affects the theory in its prediction of absorption. The surface-resistance ratio depends on the temperature and on the frequency. Confining ourselves to the situation D', then by means of a simple estimate based on the two-fluid model and Pippard's theory, we arrive at the interesting conclusion that the frequency-dependent part of q is a "universal" function^[9] of the "reduced frequency" $h\omega/kT_c$, that is to say, independent of l , ξ_0 , λ_p and v . This conclusion seems to be supported by the experimental results (see fig. 8).

Our subject is the skin effect. We are not, therefore, concerned with those situations in which the field penetrates far into the metal. Examples of these in superconductivity are to be found in a superconductor of type I in the intermediate state, where the metal divides into superconducting and normal regions and the type II superconductor in the mixed state, threaded by "fluxoids". The theoretical line of thought traced above, Gorter & Casimir — London — Pippard, which we shall deal with at greater length in the following sections, is a suitable basis for a discussion of the skin effect, in particular because of the relation to the anomalous skin effect in normal metals. We shall completely disregard another theoretical development, with which in particular the names of Ginzburg, Landau and Abrikosov are associated, and which predicts, among other things, the mixed state. The theories mentioned so far are phenomenological. Since 1957 there has also been a fundamental microscopic theory of superconductivity, the theory of Bardeen, Cooper and Schrieffer (BSC)^[10]. This, too, we shall disregard, although occasionally we shall use its results.

The finite penetration depth according to Becker, Heller and Sauter

We shall now briefly examine the argument put forward by Becker, Heller and Sauter^[3] showing that the absence of friction ($\sigma \rightarrow \infty$) does not lead to zero penetration depth, as would be predicted from the equation for the classical skin depth, but to a finite penetration depth because of the inertia of the electrons. They consider a superconducting sphere which is set into rotation about an axis through its centre. At first sight we might think that, owing to the complete absence of friction and to the inertia of the electrons, it is not possible to set the electrons in motion in this way. However, since initially the electrons do *not* move, whereas the lattice ions *do*, a time-varying current exists and as a result of the magnetic and electric fields associated with it the electrons in the sphere are finally swept into motion too, all except those in a thin shell under the surface. Becker, Heller and Sauter also showed that a close analogy exists between the setting in motion of a superconducting body and the switching phenomena in superconductors. We shall briefly recapitulate their reasoning here, adapted to the case of a superconducting

[9] See T. E. Faber and A. B. Pippard, Proc. Roy. Soc. A **231**, 336, 1955.

[10] J. Bardeen, L. N. Cooper and J. R. Schrieffer, Phys. Rev. **108**, 1175, 1957.

[11] This "perfect diamagnetism" is sometimes expressed as zero permeability. In our description, however, zero induction inside the body is due to zero *net* magnetic field (due to external coils and induced currents) and μ is just the permeability of the "medium" (see I, Introduction), being usually equal to $\mu_0 = 1.26 \times 10^{-6}$ H/m.

cylinder situated in a varying magnetic field parallel to the axis of the cylinder. The crux of their argument is that, although in a stationary state there can be no electric fields present in a perfect conductor, this does not apply to changing fields, owing to the inertia of the electrons. For electrons without friction but possessing inertia the equation of motion in an electric field \mathbf{E} is

$$m\dot{\mathbf{v}}_e = e\mathbf{E}, \quad \dots \dots \dots (2)$$

which, with $\mathbf{J} = ne\mathbf{v}_e$, yields

$$\mathbf{E} = \mu\lambda_p^2\dot{\mathbf{J}}, \quad \dots \dots \dots (3)$$

where

$$\lambda_p^2 = m/\mu ne^2 \quad \dots \dots \dots (4)$$

(see II, 23). With this equation, together with the Maxwell equations, the finite penetration depth of the varying part of the magnetic field is easily found. Substitution of (3) in Maxwell's second relation: $\text{curl } \mathbf{E} = -\dot{\mathbf{B}}$, with $\mathbf{B} = \mu\mathbf{H}$, yields:

$$\text{curl } \dot{\mathbf{J}} = -\dot{\mathbf{H}}/\lambda_p^2. \quad \dots \dots \dots (5)$$

Integration with respect to time — assuming that at a given instant the magnetic field is \mathbf{H}_a and the current density \mathbf{J}_a (both being then functions of place but not of time) — we find:

$$\text{curl } (\mathbf{J} - \mathbf{J}_a) = -(\mathbf{H} - \mathbf{H}_a)/\lambda_p^2. \quad \dots \dots (6)$$

With $\text{curl } (\mathbf{H} - \mathbf{H}_a) = \mathbf{J} - \mathbf{J}_a$ (since both $\text{curl } \mathbf{H} = \mathbf{J}$ and $\text{curl } \mathbf{H}_a = \mathbf{J}_a$) this results in:

$$\text{curl curl } (\mathbf{H} - \mathbf{H}_a) = -(\mathbf{H} - \mathbf{H}_a)/\lambda_p^2,$$

or (as $\text{curl curl} = \text{grad div} - \Delta$ and $\text{div } (\mathbf{H} - \mathbf{H}_a) = 0$):

$$\Delta(\mathbf{H} - \mathbf{H}_a) = (\mathbf{H} - \mathbf{H}_a)/\lambda_p^2. \quad \dots \dots (7)$$

If the radius of the cylinder is large compared with λ_p , then to a good approximation we can treat the surface of the cylinder as a plane surface. For the simple configuration introduced earlier (\mathbf{H} and \mathbf{J} parallel to the plane surface of the conductor and mutually perpendicular) eq. (7) leads to:

$$H - H_a = (H - H_a)_0 \exp(-z/\lambda_p), \quad \dots \dots (8a)$$

$$J - J_a = (J - J_a)_0 \exp(-z/\lambda_p). \quad \dots \dots (8b)$$

The suffix 0 relates to the values at the surface. The current and field changes therefore effectively penetrate to a depth λ_p .

Becker, Heller and Sauter show that this result can also be obtained if, starting from a normal conductor, one goes to the limit $\sigma \rightarrow \infty$, taking the inertia of the electrons into account right from the beginning. In shortened form the reasoning is as follows. In Drude's theory allowance is made for the inertia of the elec-

trons. From the equations (II, 18, 19) a complex frequency-dependent conductivity can be introduced:

$$\sigma^* = \sigma/(1 + j\omega\tau) = ne^2\tau/m(1 + j\omega\tau).$$

The limit of this expression for $\sigma \rightarrow \infty$ (i.e. for $\tau \rightarrow \infty$) is:

$$\sigma^* = ne^2/j\omega m.$$

In the classical region, the complex skin depth (see II, 27) is given by $\delta^2 = j/\omega\mu\sigma$. Substituting for σ in this equation the above value for σ^* one finds:

$$\delta^2 = -m/\mu ne^2 = -\lambda_p^2, \quad (\text{cf. eq. 4}),$$

or

$$\delta'' = \lambda_p, \quad \delta' = 0. \quad \dots \dots \dots (9)$$

The skin effect is "purely reactive": the alternating fields are indeed shielded but there is no energy dissipation.

As already remarked, this reasoning is basically identical with the theory of extreme relaxation presented in part II. In view of the complications of the anomalous skin effect, however, the step proposed by Becker, Heller and Sauter is not permissible at low frequencies. For this reason alone one may not consider the superconductor as the limiting case of a normal conductor for $\tau \rightarrow \infty$.

The Meissner effect, the London equations and the two-fluid model

Although before 1933 it was taken as established that it is not possible to change the magnetic induction in the interior of a superconductor, there was at the time no reason to suppose that it would not be possible to "freeze in" the flux by first placing the body in a magnetic field and then making it superconducting by cooling it. In fact, however, the flux is excluded in such an experiment; this is the Meissner effect already referred to, which Meissner and Ochsenfeld^[4] discovered by accurately measuring the magnetic field around a superconducting body. Since 1933, therefore, it has been considered that, apart from perfect conductivity, zero magnetic induction inside the body ("perfect diamagnetism") is an essential property of a superconductor^[11].

The London theory^[5] gives a description of these facts. F. and H. London postulated that, in addition to the Maxwell equations, the following two equations are applicable in a superconductor:

$$\dot{\mathbf{J}} = \mathbf{E}/\mu\lambda_L^2 \quad \dots \dots \dots (10)$$

and

$$\text{curl } \mathbf{J} = -\mathbf{H}/\lambda_L^2 \quad \dots \dots \dots (11a)$$

or

$$\mathbf{J} = -\mathbf{A}/\mu\lambda_L^2, \quad \dots \dots \dots (11b)$$

where \mathbf{A} is the *vector potential* defined by $\text{curl } \mathbf{A} = \mathbf{B}$, $\text{div } \mathbf{A} = 0$, and

$$\lambda_L^2 = m/\mu n_s e^2, \quad \dots \dots \dots (12)$$

n_s being the concentration of the superconducting electrons. Eq. (10) is exactly analogous to the acceleration equation (3) for frictionless electrons. Eq. (11a) would follow from (6) if the time-independent part of field and current were zero. The London hypothesis — which, incidentally, was also supported by quantum-mechanical speculations — therefore boils down to taking the integration constant equal to zero when integrating eq. (5). With these London equations, in the same way as with (7), it now follows that

$$\Delta \mathbf{H} = \mathbf{H}/\lambda_L^2, \quad \dots \dots \dots (13)$$

from which it again follows, just as (7) leads to (8), that the field in a superconductor, apart from a skin of thickness λ_L , is virtually zero. The London brothers thus made no attempt to explain superconductivity as a limiting case of normal conductivity, but postulated eqs. (10) and (11) as electromagnetic relations for a superconductor, in the place of Ohm's law for a normal conductor.

The Meissner effect implies that the superconducting-to-normal transition is reversible. Fig. 4 shows an H - T diagram. If we change a sphere from $A(T > T_c, H = 0)$ via $B(T = 0, H = 0)$ to $C(T = 0, 0 < H < H_{c0})$, then the flux is excluded. If we now raise the temperature again ($\rightarrow D$) the flux penetrates the sphere. Before the Meissner effect was discovered it was thought that cooling from state D would bring about a superconductive state C' with frozen-in flux. The transition from C to D would therefore not be reversible. According to the Meissner effect it is not C' that is obtained upon cooling from D , but C ; in other words, the transition C - D is reversible. This is of great importance because it gives confidence that reversible thermodynamics may also be applied to the superconductor.

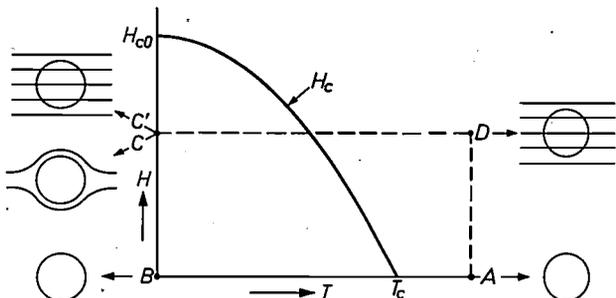


Fig. 4. The phase diagram of a superconductor in the H - T plane. Inside the parabolic curve (eq. 1) the metal is superconducting. If the metal is cooled from D (with flux through the sphere) to 0°K , the result is not C' (with "frozen-in flux") but C (with excluded flux). This is the Meissner effect.

Superconductivity is a new phase of the metal, and a well-defined state of the superconductor is associated with each point of the H - T diagram.

By means of the two-fluid model of Gorter and Casimir [6], it is possible to describe the properties of the superconducting phase in simple terms. In this model the electrons can be in either of two kinds of state; a fraction x are in the "excited" or "normal" states and a fraction $1 - x$ in the "condensed" or "superconducting" states. The thermodynamic functions — as regards their dependence on x — can be chosen in a simple manner such that a number of thermodynamic properties of the superconductivity — in particular the form of the critical field curve (1) — can be satisfactorily described. It is then found that

$$x = \vartheta^4, \quad \dots \dots \dots (14)$$

where

$$\vartheta = T/T_c. \quad \dots \dots \dots (15)$$

In this two-fluid model the concentration of the superconducting electrons is $n_s = n_e (1 - x)$. In combination with the London theory this now gives for the penetration depth λ_L :

$$\lambda_L^2 = m/\mu n_s e^2 = m/\mu(1 - x)n_e e^2 = \lambda_p^2/(1 - x). \quad (16)$$

From this model we expect the penetration depth λ_L to be dependent upon temperature in accordance with:

$$\lambda_L = \lambda_p(1 - \vartheta^4)^{-1/2}. \quad \dots \dots \dots (17)$$

The complex skin depth in the GCL model

With the model described above it is a simple matter to calculate the complex skin depth, giving both the penetration and the absorption.

We again take the simple configuration given in fig. 1 of part II (\mathbf{E} and \mathbf{J} // x -axis, \mathbf{H} // y -axis, z -axis surface) and we again assume the fields and currents to be proportional to $\exp j(\omega t - z/\delta)$. We disregard the displacement current, but we do take into account the relaxation of the normal electrons.

Under these conditions the Maxwell equations (I, 1 and 2) give $jH/\delta = J$ and $-jE/\delta = -j\omega\mu H$. The current J is now composed of J_s , the current of the superconducting electrons, and J_n , the current of the normal electrons: $J = J_n + J_s$. The London equations for the superconducting electrons give $j\omega J_s = E/\mu\lambda_L^2$ and $-jJ_s/\delta = -H/\lambda_L^2$. The concentration of the normal electrons is $n_n = xn_e$. The normal electrons respond to a given electric field in the same way as in the normal conductor; only their *number* is reduced. Introducing λ_n as a measure of the concentration of normal electrons,

$$1/\lambda_n^2 = \mu n_n e^2/m = x/\lambda_p^2, \quad \dots \dots (18)$$

we then have (cf. II,19):

$$J_n = \sigma_n E / (1 + j\omega\tau), \dots (19)$$

where $\sigma_n = \tau / \mu \lambda_n^2 = x\tau / \mu \lambda_p^2$ (cf. II,25). Elimination of the fields and currents, and making use of (16), yields the result:

$$-\frac{1}{\delta^2} = \frac{j\omega\tau x}{1 + j\omega\tau} \frac{1}{\lambda_p^2} + \frac{1-x}{\lambda_p^2} \dots (20)$$

The "screening" $-1/\delta^2$ is thus composed of two additive terms: the first is the $-1/\delta^2$ that would be obtained only for normal electrons in a concentration xn_e (see II,30; the first term there originates from the displacement current, here neglected), and the second term is the $-1/\delta^2$ that would be obtained only for superconducting electrons in a concentration $(1-x)n_e$ (see eq. 9).

Eq. (20) can be written in a more abbreviated form:

$$\lambda_p^2 / \delta^2 = -1 + \frac{x}{1 + j\omega\tau} \dots (21)$$

Fig. 5 shows a few curves described by δ in the complex plane in accordance with eq. (21) when $\omega\tau$ goes

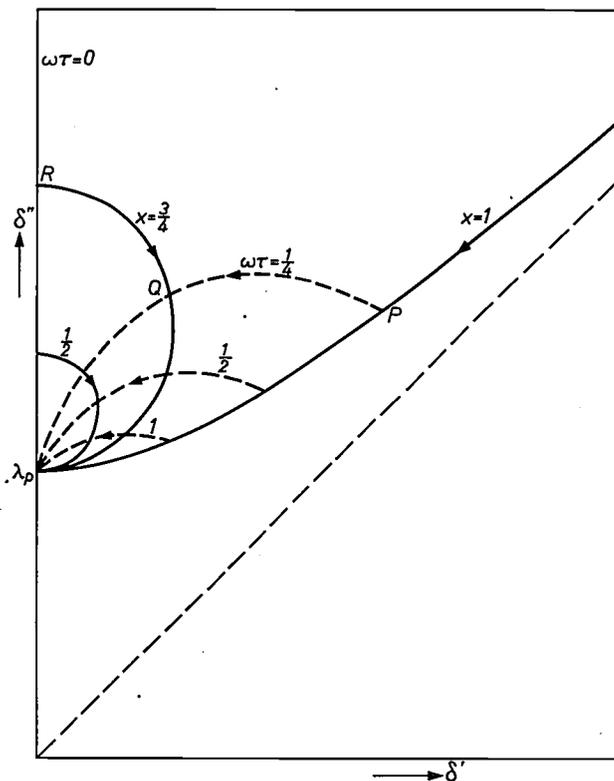


Fig. 5. The complex skin depth δ in the two-fluid model according to (21). This equation is derived from the Maxwell equations in conjunction with the London equations (10, 11) for the supercurrent and Drude's equation (19) for the normal current. The solid curves show how δ changes as ω goes from 0 to ∞ for given fixed values of x (i.e. at specific temperatures). The dashed curves represent δ as the temperature goes from T_c to zero (i.e. x from 1 to 0) for given fixed values of $\omega\tau$, i.e. at given frequencies. (For $\omega\tau = 0$, the curve is coincident with the δ' -axis.) The displacement current is neglected.

from zero to infinity at a given x (solid lines) and when x goes from 1 to 0 at a given $\omega\tau$ (broken lines). The line " $x = 1$ " gives δ for a conductor in the normal state. For ω increasing, δ goes in the direction of the arrow. If we now keep $\omega\tau$ constant when P is reached, and we cool the superconductor to below the critical temperature, then δ follows the curve " $\omega\tau = \frac{1}{4}$ ". If, having arrived at Q we now keep the temperature constant but let the frequency drop to $\omega = 0$, then δ follows the curve " $x = \frac{3}{4}$ " (opposite to the direction of the arrow) and arrives at R : this is "pure London penetration" without absorption.

The figure illustrates the relation between superconductivity and extreme relaxation: the point $j\lambda_p$ can be reached both by $x \rightarrow 0$ (100% superconducting electrons) and by $\omega\tau \rightarrow \infty$ (extreme relaxation). In both cases the electrons behave as completely free particles.

For $\omega\tau \ll 1 - x$ (low frequencies and temperatures well under to the critical temperature) we find from (21):

$$\delta'' = \lambda_p (1 - x)^{-1/2} = \lambda_L, \dots (22)$$

$$\delta' = \frac{1}{2} \omega\tau \lambda_p x (1 - x)^{-3/2} = \frac{1}{2} \omega\tau \lambda_L^3 / \lambda_n^2. (23)$$

δ'' and δ' depend on the temperature via x only. We see that, under the stated condition, δ'' is dependent only on the choice of the metal (λ_p) and the temperature (x), while δ' depends in addition upon the purity of the metal (τ) and the frequency ω .

As we shall see, some characteristic properties of superconductivity in an alternating electromagnetic field are well described by this theory; on a number of important points, however, there is a considerable difference.

Nevertheless, it is possible to make two simple amendments to this theory that bring us closer to a correct description of superconductivity in an alternating electromagnetic field.

With a view to these amendments, let us consider once again the physical background to equations (22) and (23). The penetration depth in the superconductor, which can have different values in different theories, is denoted generally in this article by λ_s , as above. Eq. (22) scarcely requires any further comment: under the condition stated ($\omega\tau \ll 1 - x$), δ' is much less than δ'' , and in that case δ'' is effectively the penetration depth: $\lambda_s = \delta''$. Eq. (22) expresses the fact that λ_s does not depend on frequency and is equal to λ_L , the penetration depth previously found for $\omega = 0$. In eq. (23), the following will show that the expression $\frac{1}{2} \omega\tau \lambda_L^3 / \lambda_n^2$ has a more general significance. δ' is a measure of the energy absorbed by the normal electrons in the penetration layer. The restriction of the fields and the currents to a layer (of thickness λ_s) is entirely due to the superconducting electrons. Without causing

dissipation they carry the current which according to Maxwell's first relation (I,1), must exist in the presence of our (rotational) magnetic field. In the layer the varying magnetic field induces an electric field, which according to Maxwell's second relation (I,2) is given by $E/\lambda_s = j\omega H$; E is 90° out of phase with H . This electric field causes a current of normal electrons: $J_n = \sigma_n E$ (where $\sigma_n = \tau/\mu\lambda_n^2$) and thereby performs work. The energy delivered to the electrons per second and per m^3 , averaged over the time, is $\overline{J_n E} = \frac{1}{2}\sigma_n E^2$. By integration over the layer λ_s we find the power per m^2 of surface: $\frac{1}{4}\sigma_n \lambda_s E_0^2 = \frac{1}{4}\sigma_n \lambda_s^3 \mu^2 \omega^2 H_0^2 = \tau \lambda_s^3 \mu \omega^2 H_0^2 / 4\lambda_n^2$, where E_0 and H_0 are the amplitudes of the fields at the surface. This power is therefore proportional to the concentration of the normal electrons ($1/\lambda_n^2$) and to τ , the average time of uninterrupted interaction between electric field and normal electron. We now find a more general form of (23), the expression for δ' , by equating absorbed power = absorptivity \times incident power:

$$\tau \lambda_s^3 \mu \omega^2 H_0^2 / 4\lambda_n^2 = (4\omega\mu\delta'/Z_0) \times \frac{1}{8}Z_0 H_0^2.$$

(The incident power is $\frac{1}{2}Z_0 H_1^2$ where H_1 is the amplitude of the magnetic field of the incident wave. $H_0 = 2H_1$ since the magnetic fields of the incident and reflected wave at the surface are in phase. $Z_0 = (\mu_0/\epsilon_0)^{1/2}$ is the impedance of free space. For the absorptivity, cf. II,14 and 12.) We obtain:

$$\delta' = \frac{1}{2}\omega\tau\lambda_s^3/\lambda_n^2. \quad \dots \quad (24)$$

From the above we can give eq. (24) a wider interpretation: τ is now the average time during which the field acts uninterruptedly on a normal electron, λ_s is the thickness of the penetration layer and λ_n is a measure of the number of normal electrons (cf. eq. 18).

The first of the amendments referred to above comes from Pippard's non-local theory of superconductivity, which runs parallel with the anomalous skin effect in normal conductors. This theory results in a value for the thickness λ_s of the penetration layer which differs from the London value λ_L (just as the theory of the anomalous skin effect results in a skin depth differing from the classical one).

The second amendment takes account of the fact that, in general, the mean free path of the normal electrons is greater than the penetration depth. We make allowance for this by filling in for τ , instead of the time between two collisions, the time needed to pass through the penetration layer. In eq. (24), therefore, we must substitute a new value both for τ and for λ_s .

After briefly discussing the methods of measuring the penetration depth and the surface resistance, we shall compare the results of the measurements with the predictions of the theory before and after amendment.

Measurements of the penetration depth

In principle, the penetration depth λ_s can be determined by establishing for a given configuration and a given external field the difference between the magnetic flux actually measured and the flux one would expect to find if there were no penetration. The latter would have to be calculated from the geometry. For small penetration depths this method is very difficult, and it has not in fact proved possible to determine λ_s in this direct way. It is, however, much more feasible to measure small *changes* of flux, and if λ_s can be influenced in one way or another one can then obtain in this way some information on λ_s . Here eq. (17) comes to our assistance: the penetration depth can be influenced by the temperature. The methods of determining λ_s therefore all amount to an experimental arrangement in which some quantity (generally the flux) is a linear function of λ_s : the procedure is then to establish that this quantity is a linear function of $(1 - \vartheta^4)^{-1/2}$. After conversion to λ_s , the coefficient of $(1 - \vartheta^4)^{-1/2}$ is identified with λ_0 , the penetration depth at $T = 0$.

1) Since the penetration is a surface effect it is necessary to make the surface relatively large. Shoenberg [12], for example, measured the magnetic susceptibility of mercury suspensions in which the mercury was present in the form of a large number of small spheres with diameters ranging from 10 to 100 nm. Variations on this theme are susceptibility measurements on capillary mercury threads [13] and on thin films [14]. Although these measurements give a clear qualitative demonstration of the penetration, only the latter has yielded reliable quantitative results.

2) In the Casimir method [15] the magnetic flux *outside* the body proper is made as small as possible. The method consists in measuring, at low frequencies, the mutual induction between two coils; the primary is wound *closely* on a cylindrical superconducting body and the secondary is wound around the primary. The coefficient of mutual induction M is proportional to the cross-sectional area available to the flux through the primary coil; in other words, when the cylinder is in the superconducting state, it is proportional to $D + \lambda_s$ where D is the width of the gap between cylinder and coil. The accuracy with which variations in λ_s can be measured therefore depends on λ_s/D and on the accuracy with which M is measured. An advantage of the method is that well-defined macroscopic bodies can be used. The method, with some refinements, has been successfully employed by Laurmann and Shoenberg [16]. A variant is the measurement of changes in the resonant frequency of an LC circuit in which L is the inductance of a coil wound on a superconducting cylinder [17].

3) The measurement of changes in resonant fre-

quency was applied by Pippard [18] to superconducting resonators at microwave frequencies. The resonant frequency of a microwave cavity is always, owing to field penetration in the wall, somewhat lower than would follow from the dimensions if there were no penetration. Variations of the complex skin depth δ cause variations of the resonant frequency proportional to δ'' . It is possible to vary δ not only via the temperature but also — at a given $T < T_c$ — by means of a magnetic field that disturbs the superconducting state; δ'' then jumps from the penetration depth in the superconductor λ_s to the value δ_n'' of the normal metal. The accompanying jump in the resonant frequency gives a measure of the jump $\Delta\delta''$ at different temperatures. The variations found in $\Delta\delta''$ are entirely attributable to λ_s because δ_n'' in this temperature range can be regarded as independent of temperature. One can then find λ_0 from the slope of $\Delta\delta''$ as a function of $(1 - \vartheta^4)^{-1/2}$. An incidental result of this measurement is the determination of δ_n'' , this being the sum of λ_0 and the jump found at 0 °K by extrapolation (see fig. 6). In this way Pippard was able to verify experimentally eq. (II,41), $\delta_n'' = \delta_n'/\sqrt{3}$, in the extreme anomalous region, δ_n' having been found from absorption measurements (see II).

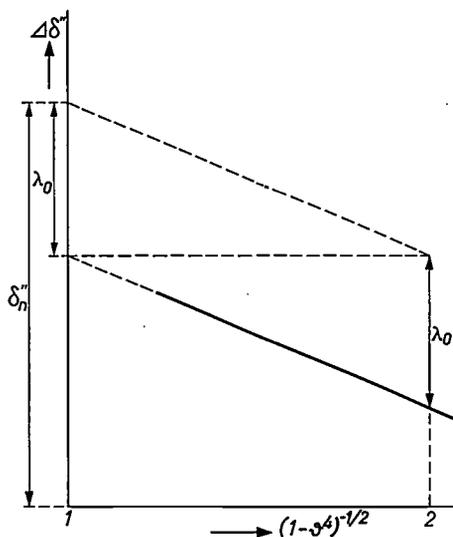


Fig. 6. The solid straight line gives $\Delta\delta''(T) = \delta_n'' - \lambda_s(T)$, the jump in δ'' when the superconducting state is perturbed by a magnetic field, as a function of $(1 - \vartheta^4)^{-1/2}$. This jump in δ'' is determined experimentally from the jump in the resonant frequency of a cavity resonator (see text). λ_0 is deduced from the slope of the line. In addition, δ_n'' is found from λ_0 and $\Delta\delta''$ at $\vartheta = 0$ using $\delta_n'' = \Delta\delta''(T = 0) + \lambda_0$.

The results of the measurements carried out so far may be summarized as follows:

- 1) Eq. (17) gives a good description of the temperature dependence.
- 2) The values found for the penetration depth at 0 °K, λ_0 , are indeed close to the theoretical value λ_p , but

exceed this value systematically by a factor of roughly 2 to 5. Table I [19] gives various values for λ_0 found from measurements; they may be compared with the value of λ_p for the standard metal introduced in part II (see II, Table I): $\lambda_p = 22$ nm.

Al	50
Cd	130
Hg	38-45
In	64
Pb	39
Sn	{ 51
	{ 47-60
Tl	92

Table I. Experimental values (in nm) for λ_0 , the penetration depth at $T = 0$ in a number of superconductors. The variation in the values for Hg and Sn is due to anisotropy. The table is taken from Lynton [19].

3) Finally, there is Pippard's result, already touched upon, and which is of great importance to our subsequent considerations: if the mean free path l is very small, λ_s becomes dependent upon it [7]. This conflicts with London's theory, where $\lambda_s = \lambda_L$ is a quantity that does not contain the "sensitive parameter" l (see eq. 12).

Measurements of surface resistance

There are two principal methods of measuring the surface resistance: the calorimetric method and the method of determining the bandwidth of resonant cavities. The surface resistance is proportional to the power absorption (II,14). In the calorimetric methods the absorbed power is measured directly as the heat generated per second; this heat appears in the form of a temperature difference across a known "heat leak" between the superconducting material and the helium bath. The second method makes use of the proportionality that exists (after correction for coupling-hole effects) between the line width of the resonance of the cavity and the power absorbed in the wall.

In discussions of the results at not too high frequencies it is usual to try and express the measured surface resistance as a constant plus the product of a function of the temperature only and a function of the frequency only:

$$R = R_n A(\omega) \Phi(\vartheta) + R_0 \dots (25)$$

(The fact that this does not hold at high frequencies is evident, for example, from the results of fig. 9). In (25), R is the measured surface resistance at a chosen

[12] D. Shoenberg, Proc. Roy. Soc. A 175, 49, 1940.
 [13] M. Désirant and D. Shoenberg, Proc. Phys. Soc. 60, 413, 1948.
 [14] J. M. Lock, Proc. Roy. Soc. A 208, 391, 1951.
 [15] H. B. G. Casimir, Physica 7, 887, 1940.
 [16] E. Laurmann and D. Shoenberg, Proc. Roy. Soc. A 198, 560, 1949.
 [17] W. L. McLean, Proc. VIIth Int. Conf. on low temperature physics, Toronto 1960, p. 330; Univ. of Toronto Press, 1961.
 [18] A. B. Pippard, Proc. Roy. Soc. A 191, 399, 1947.
 [19] E. A. Lynton, Superconductivity, Methuen, London 1962, p. 37.

$\vartheta < 1$, and R_n is the surface resistance of the normal metal just above the critical temperature. $\Phi(\vartheta)$ is a function of ϑ which is 0 for $\vartheta = 0$. R_0 is a "residual value" of R which depends to a great extent on the quality of the surface and which is generally assumed to be zero for an "ideal surface". Correcting for this (by extrapolating the measured results to $\vartheta = 0$), we find for the surface resistance ratio $q = R/R_n$ which, from (II,12), is equal to δ'/δ_n' :

$$q = A(\omega)\Phi(\vartheta). \quad \dots \quad (26)$$

In the GCL model in its simplest form, δ' is given by (23) and δ_n' by (II, 15, 26) ($\delta_n' = \lambda_p/\sqrt{2\omega\tau}$), so that the surface resistance ratio is found to be given by:

$$q = \delta'/\delta_n' = 2^{-1/2} (\omega\tau)^{3/2} \vartheta^4 (1 - \vartheta^4)^{-3/2}. \quad (27)$$

a) Temperature dependence

For the low-temperature region Pippard has proposed for $\Phi(\vartheta)$ the empirical expression:

$$\Phi_P(\vartheta) = \vartheta^4(1 - \vartheta^2)(1 - \vartheta^4)^{-2} \quad \dots \quad (28)$$

and in the temperature region $\vartheta < 0.8$ (i.e. $x < 0.4$ and $\Phi(\vartheta) < 0.4$) most investigators in fact find that the experimental values of q plotted against $\Phi_P(\vartheta)$ lie roughly on a straight line. The slope of this line then gives $A(\omega)$.

According to the GCL model the temperature dependence is given by (see 27):

$$\Phi_1(\vartheta) = \vartheta^4(1 - \vartheta^4)^{-3/2}. \quad \dots \quad (29)$$

Fig. 7 gives a plot of $\Phi_P(\vartheta)$ and $\Phi_1(\vartheta)$ versus ϑ .

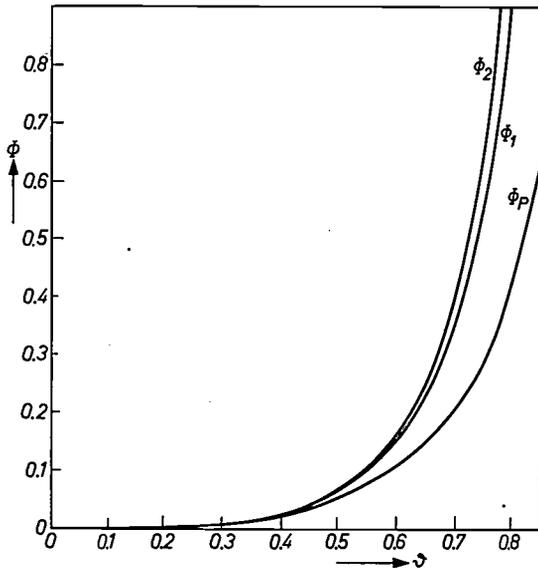


Fig. 7. The temperature dependence of the surface resistance ratio q according to the London theory (Φ_1 , eq. 29), the amended theory (Φ_2 , eq. 39) and Pippard's empirical formula (Φ_P , eq. 28). The functions are determined apart from a proportionality factor; they can be made to coincide at any chosen temperature. We have chosen $\Phi_1 = \Phi_2 = \Phi_P = \vartheta^4$ for $\vartheta \rightarrow 0$.

Qualitatively both curves have the same shape. If the curves for low values of ϑ are made to coincide (both then being $\propto \vartheta^4$) we see that at $\vartheta = 0.8$ there is a discrepancy of about a factor of 2.

b) Frequency dependence

Fig. 8 shows values of $A(\omega)$ as functions of $\hbar\omega/kT_c$, for various materials, as found by a number of investigators or derived from their results by the method described above (in some cases by using eq. (26) in a range where it does not really apply, see caption). Remarkably enough, all points lie more or less on one curve, whereas from (27) one would expect a considerable spread owing to the presence of τ in the expression.

Differences between the predictions of the frequency dependence of q arise from differences, not in δ' , but in δ_n' . With the GCL model, and also after the amendment to it which we shall presently discuss, it is found that δ' is proportional to ω . For δ_n' , however, one can choose between the classical limit ($\delta_n' = \lambda_p/\sqrt{2\omega\tau}$, cf. II,15,26) and the extreme anomalous limit ($\delta_n' = (\lambda_p^2 v/4b\omega)^{1/3}$, cf. II, 40, 41). In the first case $q \propto \omega^{3/2}$ (as in 27), and in the second case $q \propto \omega^{4/3}$. Provided the superconductors are not unduly "dirty" l is usually much greater than δ_n , so that in general we are concerned with the second case. In fig. 8 the slope 4/3 in the low-frequency region does in fact seem to fit the measured points better than the slope 3/2, although the difference is slight.

At higher frequencies the curve first flattens out, and then rises steeply.

c) Energy gap

The abrupt rise of $A(\omega)$ at $\hbar\omega/kT_c \approx 3$ is attributable to something that has not yet been touched upon, namely a quantum process. In very simple terms it may be said that an energy 2Δ is needed in order to raise the electrons from the superconducting to the normal state. At absolute zero this energy is roughly $3.5 kT_c$ (the BCS theory predicts $3.52 kT_c$), and it goes to zero when T goes to T_c . Now if the frequency becomes so high that $\hbar\omega \geq 2\Delta$, the absorption suddenly increases; quanta $\hbar\omega$, which raise the superconducting electrons to the normal state, are absorbed.

This behaviour is neatly illustrated by the measurements by Biondi and Garfunkel [24] on aluminium (fig. 9). Here again q is plotted against $\hbar\omega/kT_c$ (now

[20] C. J. Grebenkemper and J. P. Hagen, Phys. Rev. **86**, 673, 1952 (including some of Pippard's results quoted by these authors).
 [21] R. Kaplan, A. H. Nethercot, Jr., and H. A. Boorse, Phys. Rev. **116**, 270, 1959 (including results due to M. D. Sturge quoted by these authors).
 [22] C. J. Grebenkemper, Phys. Rev. **96**, 316, 1954.
 [23] M. S. Khaikin, Sov. Phys. JETP **7**, 961, 1958.
 [24] M. A. Biondi and M. P. Garfunkel, Phys. Rev. **116**, 853, 1959.

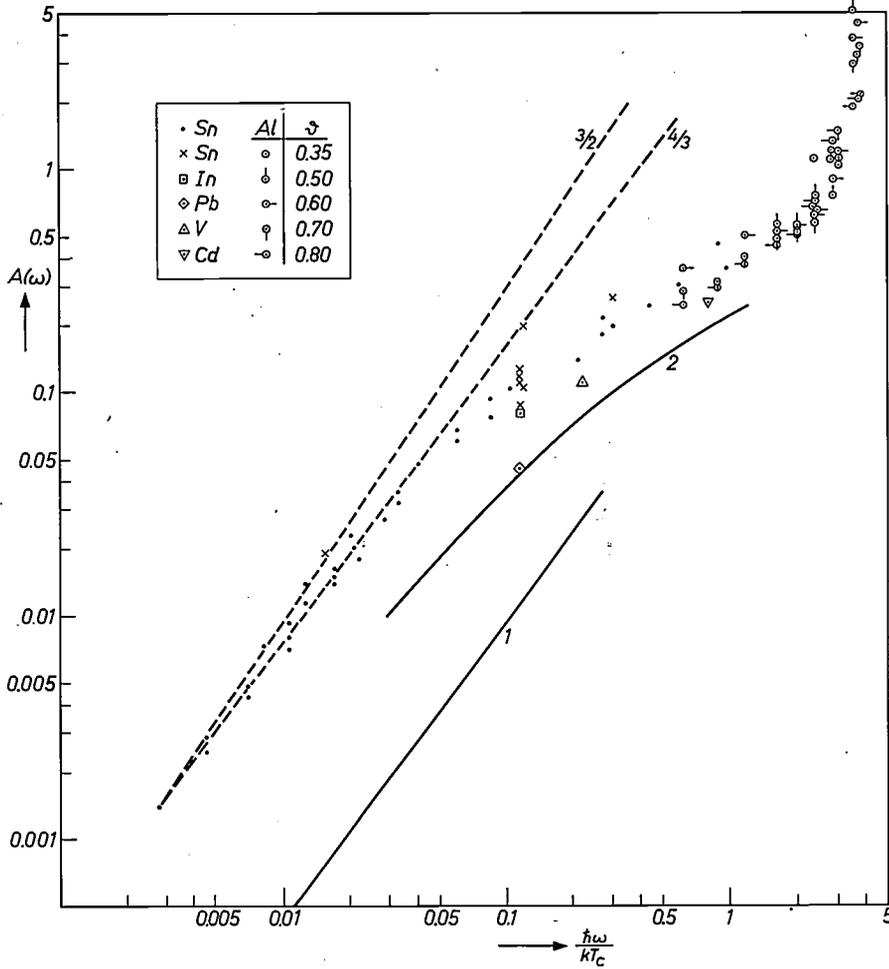


Fig. 8. The frequency-dependent factor $A(\omega)$ in the surface resistance ratio q as a function of $\hbar\omega/kT_c$ for Sn (crosses [20], dots [21]), In [20], Pb [20], V [22], Cd [23] and Al [24]. The plotted values of A are either the values reported by the authors themselves or the values derived from their results (the slope of the curve q vs Φ_F , eq. (28), in so far as this is straight; eq. (26)). A slightly different treatment has been applied to the results of Biondi and Garfunkel [24] (see fig. 9). Eq. (26) is evidently no longer applicable to these. Values for $A(\omega)$ have nevertheless been derived from these results (for $\vartheta \leq 0.8$) by putting $A = q/\Phi_F(\vartheta)$. The failure of (26) is reflected in a spread of the points obtained in this way for Al. Even so, these points link up to some extent with the others and also give an idea of the marked increase of q at the energy gap. The straight line 1 is the relation between $A(\omega)$ and $\hbar\omega/kT_c$ according to (40), the rough, amended London theory. Curve 2 is the relation given by the BCS theory in the extreme non-local limit for tin (Miller [29]). A line with slope 3/2 and a line with slope 4/3 have been drawn through the points at low frequencies (see text, p. 374).

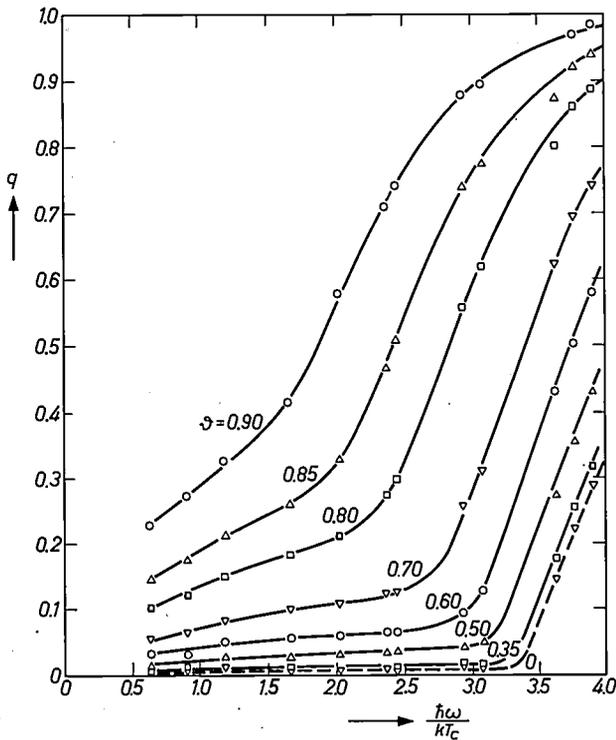


Fig. 9. Surface resistance ratio q as a function of $\hbar\omega/kT_c$ at different temperatures for Al, after Biondi and Garfunkel [24]. Each curve shows a kink, which is sharpest at the lowest temperatures. The kink is interpreted as the point where $\hbar\omega$ is equal to the energy gap.

on a linear scale) for various values of ϑ . The kink in the curve where the quantum energy is equal to the energy gap lies at lower frequencies the higher the temperatures. In fig. 10 the energy gap found from these experiments is plotted as a function of ϑ together with the theoretical curve predicted by the BCS theory. The points for aluminium in fig. 8 were derived from the results of fig. 9 for $\vartheta \leq 0.8$.

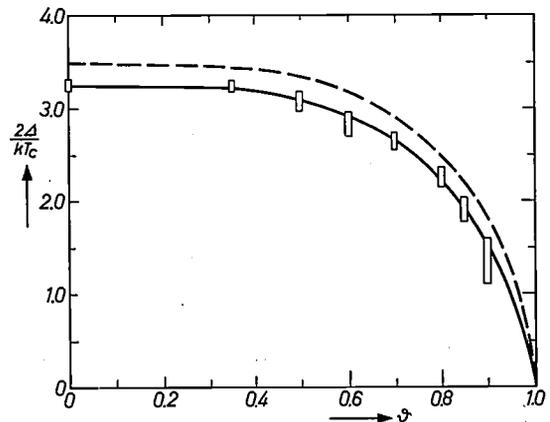


Fig. 10. The energy gap 2Δ (expressed in units of kT_c) for Al as a function of the temperature, after Biondi and Garfunkel [24]. The experimental values (rectangles) are derived from the results shown in fig. 9. The dashed curve is the result given by the BCS theory.

The experimental results and the GCL model

In considering the results of the measurements we have already made passing references to the GCL model by way of comparison. We shall now summarize this comparison in the following list.

- 1) The model gives a good prediction of the temperature dependence of the penetration depth λ_s (eq. 17).
- 2) The model predicts the order of magnitude of the penetration depth for $T = 0^\circ\text{K}$, λ_0 , but the theoretical value is systematically too small by a factor of between 2 to 5.
- 3) The experimental result that λ_s is dependent on the mean free path l is in contradiction with the theory.
- 4) The temperature dependence of the surface resistance ratio q is qualitatively predicted, but there are quantitative differences of a factor of 2 (fig. 7).
- 5) The theory gives a good prediction of the frequency dependence of q at low frequencies (fig. 8).
- 6) The results of the measurements indicate that the frequency-dependent factor in q , $A(\omega)$, as a function of $\hbar\omega/kT_c$, is not particularly susceptible to the "sensitive parameter" (l or τ) or to the choice of metal. This is not in agreement with the theory.

Item 3 was one of the main reasons for the postulation of the non-local theory. This theory will be briefly outlined in the following section. It will be shown that the non-local theory not only clarifies item 3 but also reduces the discrepancy under item 2.

The two amendments referred to (a new value for λ_s and a new value for τ) will then be applied to eq. (24), and we shall see that this gives some clue to the cause of the interesting result mentioned under item 6.

The non-local theory

In the London equation (11b)

$$\mathbf{J} = (-1/\mu\lambda_L^2)\mathbf{A} \dots \dots (30)$$

the coefficient $-1/\mu\lambda_L^2$ is governed by the number of superconducting electrons n_s . London showed that this equation might be derived by assuming that the superconducting electrons are in a very specific state ("the quantum state $p = 0$ ") in which they remain even when a magnetic field is applied. This may be freely interpreted by saying that the "order parameter" n_s is a rigid quantity that does not depend on the field and that has the same value throughout a superconductor at a particular temperature, even at the edge where the field penetrates slightly. Later, observations and theoretical considerations led to the assumption that this rigidity is not absolute: although such a thing exists as a "coherence length" within which the order parameter can undergo scarcely any change, variations are nevertheless possible over greater lengths. The remaining rigidity implies that the state at one given point

influences the state at another point at a distance less than the coherence length. Considerations of this nature led Pippard to propose that the current at a given point is determined not only by the vector potential at that point but also by the vector potential in an environment of the order of the coherence length. In other words, there exists a non-local relation between current and vector potential analogous to the non-local relation that exists between current and electric field in the theory of the anomalous skin effect (see II).

The analogy with the anomalous skin effect is very pronounced in the earlier mentioned experimental result found by Pippard [7]. As the mean free path l of the electrons in a superconductor decreases (due to an increasing concentration of impurities), the penetration depth λ_s at first remains the same, but begins to increase when l becomes equal to or less than λ_s . This increase is in contradiction with the London theory, where the penetration depth, according to (12), is independent of the "sensitive parameter" l . The behaviour runs parallel, however, with the skin effect in normal conductors: there too the skin depth is independent of l at large values of l (extreme anomalous region), whereas δ increases with decreasing l when $l \ll \delta$ (classical region).

By analogy with the skin effect, Pippard proposed that the London equation be replaced by the following non-local relation (cf. II,39):

$$\mathbf{J} = \frac{c_0}{\mu\lambda_L^2} \int \frac{\mathbf{r}(\mathbf{A} \cdot \mathbf{r}) \exp(-r/\xi)}{r^4} dV. \quad (31)$$

The quantity ξ , whose function here is similar to that of l in the skin effect, is the "coherence length" referred to in the foregoing, being a measure of the distance over which the current feels the effect of the vector potential. The coherence is disturbed by impurities: at small l the value ξ is also small. The coefficient outside the integral is proportional to the number of superconducting electrons ($1/\lambda_L^2$) and independent of ξ (in the same way that the coefficient $3\sigma/4\pi l$ in the skin effect is proportional to n_e but independent of l); c_0 is therefore independent of ξ .

In the *local limit* ($\xi \ll \lambda_s$) \mathbf{A} may be regarded as uniform. The integration is easily carried out and the result is: $\mathbf{J} = (4\pi/3)\xi(c_0/\mu\lambda_L^2)\mathbf{A}$. Comparison with (30) shows that c_0 has the dimension of a reciprocal length. Putting $c_0 = -3/4\pi\xi_0$, we find the *local relation*:

$$\mathbf{J} = -\frac{\xi}{\xi_0} \frac{1}{\mu\lambda_L^2} \mathbf{A} \dots \dots (32)$$

Just as (11b) leads to a penetration depth λ_L , so (32) gives for the penetration depth λ_s in the *local limit*:

$$\lambda_s^2 = (\xi_0/\xi)\lambda_L^2. \dots \dots (33)$$

The length ξ_0 , like c_0 , does not depend on the number of superconducting electrons or on the degree of impurity. In other words, ξ_0 is a characteristic length of a clean superconductor at $T = 0$. It now seems reasonable to assume — and this is justified by other considerations and results — that even in a clean superconductor the coherence is restricted to finite regions, whose dimensions may be related to ξ_0 . In the more general theory the ξ_0 introduced above turns out to be, in fact, the *coherence length* of the clean superconductor. (It would be wrong to conclude from (33) that for a clean superconductor, where $\xi = \xi_0$, the penetration depth is λ_L because, as we shall presently see, the relation between \mathbf{J} and \mathbf{A} for a clean superconductor is, in general, not local.)

The coherence length may be regarded as a measure of the spatial extension of the wave functions of the superconducting electrons. With the aid of the uncertainty principle ξ_0 can be expressed in terms of the energy gap $2\Delta(0)$. Only those electrons with energies within a band of width $\pm\Delta(0)$ about the Fermi energy play a role. The spread in the momentum is therefore limited to a value of roughly $\Delta p = 2\Delta(0)/v$, where v is the Fermi velocity. According to the uncertainty principle, the spatial extension of the wave function is then at least $\xi_0 = \hbar/\Delta p$, from which it follows that $\xi_0 = \hbar v/2\Delta(0)$. The BCS theory leads to almost the same result:

$$\xi_0 = \hbar v/\pi\Delta(0) = a\hbar v/kT_c, \dots (34)$$

where $a = (2/\pi)kT_c/2\Delta(0) = 2/\pi \times 3.52 = 0.18$.

Using this expression we can estimate the value of ξ_0 : for the "standard metal" ($v = 1.4 \times 10^6$ m/s) with $T_c = 1.18$ °K (aluminium) we find $\xi_0 \approx 1.6$ μm .

The argument presented is somewhat vague because we do not know what exactly we are to understand by "wave functions of the superconducting electrons". In the BCS theory, superconductivity comes about because electrons of opposite momentum and spin join up to form "Cooper pairs", thereby lowering their energy. In this theory ξ_0 is the average distance between the members of a Cooper pair.

The interpretation of ξ_0 as the coherence length of the clean superconductor means that ξ , which is small when l is small, is equal to ξ_0 when $l \gg \xi_0$. To obtain this situation we may, for example, write with Pippard:

$$1/\xi = 1/\xi_0 + 1/l. \dots (35)$$

If, now, we have a superconductor which is *clean* ($l \gg \xi_0$, so that $\xi = \xi_0$) and to which the *local limit* also applies ($\xi \ll \lambda_s$) we find for the penetration depth with the aid of (33):

$$\lambda_s = \lambda_L. \dots (36)$$

The description then ties up completely with that given by London: what we have is an *intrinsic London superconductor*. The inequality $\xi \ll \lambda_s$ is then equivalent to $\xi_0 \ll \lambda_L$. As we shall presently see, such superconductors are scarcely ever found.

If, on the other hand, we have a "dirty" superconductor ($l \ll \xi_0$, so that $\xi = l$) in the *local limit*, it follows from (33) that the penetration depth is:

$$\lambda_s^2 = (\xi_0/l)\lambda_L^2. \dots (37)$$

From the estimate of ξ_0 and the values of λ_0 in Table I it follows that the local limit does *not* apply to many metals, but that on the contrary $\xi_0 \gg \lambda_s$. The penetration depth to be expected in that case may be estimated as follows. The current density at the surface is, in Pippard's theory, calculated from (31). If \mathbf{A} were uniform, the integrand would extend roughly over a depth ξ , and the result would be (32). However, \mathbf{A} , and therefore the integrand, extend only over the penetration depth λ_s . The result is therefore reduced by a factor of approximately λ_s/ξ , so that $\mathbf{J} = -(\lambda_s/\mu\xi_0\lambda_L^2)\mathbf{A}$. For the penetration depth λ_s following from this we have $\lambda_s^2 = \xi_0\lambda_L^2/\lambda_s$, or $\lambda_s^3 = \xi_0\lambda_L^2$. The rigorous theory gives for the *Pippard limit* (the non-local limit, $\xi_0 \gg \lambda_L$) a penetration depth λ_s (see also II,40):

$$\lambda_s^3 = (2/b)\xi_0\lambda_L^2. \dots (38)$$

If we integrate simply over the volume of the superconductor, we find $b = 4\pi/\sqrt{3}$ (as in the anomalous skin effect with diffuse reflection).

Fig. 11 gives a log-log plot of the penetration depth λ_s in a clean superconductor as a function of ξ_0 . We note that the theoretical value of λ_s is always greater than the values in either of the two limits (local limit for $\xi_0 \ll \lambda_L$, non-local limit for $\xi_0 \gg \lambda_L$; see also Table II).

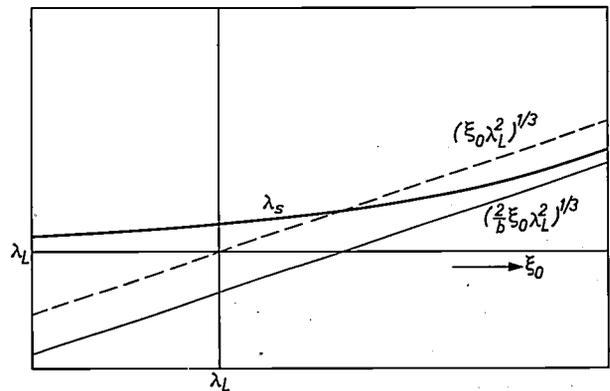


Fig. 11. The penetration depth λ_s in a clean superconductor as a function of the coherence length ξ_0 (schematic, logarithmic scales). For $\xi_0 \ll \lambda_L$ (left) λ_s approaches to λ_L (local limit); this is the "intrinsic London superconductor". For $\xi_0 \gg \lambda_L$ (right) λ_s approaches the Pippard limit, given by (38). Whatever the value of ξ_0 , the value of λ_s is greater than either of the corresponding limit values.

Of the experimental results giving general corroboration of the non-local theory, we shall mention only the following. First of all, as was to be expected, the theory provides a good description of the dependence of λ_s on l . For large l we must apply (38); λ_s is then independent of l .

As l becomes smaller we go more in the direction of (37) and λ_s increases as l decreases. Quantitatively too, Pippard found a good measure of agreement with a suitable choice of ξ_0 . Furthermore, the discrepancy between measured and predicted penetration depths in clean superconductors is considerably reduced. For our standard metal with $T_c = 1.18$ °K we found $\xi_0 = 1.6$ μ m. At $T = 0$ in the Pippard limit (38) this gives $\lambda_s = 59$ nm, a value that comes closer to the measured values than λ_L which, at $T = 0$, is equal to $\lambda_p = 22$ nm (II, Table I). As an illustration we give in Table II a few values for tin and aluminium. The success of the theory is particularly striking in the case of aluminium.

We shall not consider here the consequences of Pippard's theory on the temperature dependence of λ_s . We shall merely repeat that the prediction (17), $\lambda_s \propto (1 - \vartheta^4)^{-1/2}$ of the GCL model — only to be used as a guide for classifying the phenomena, having been superseded by more fundamental considerations — is on the whole a very reasonable prediction, not only qualitatively, but quantitatively as well.

Analogy with skin effect in normal metal; classification of superconductors

The approach described in the introduction is useful for briefly summarizing the theoretical results given above.

A superconductor is formally closely related to a normal metal in the relaxation limit. In a normal metal in the relaxation limit (B, E), v/ω is much smaller than l and δ and the screening $-1/\delta^2$ is $1/\lambda_p^2$. The modulus of the screening in the classical region A ($l \ll \delta$, v/ω) can be obtained by reducing the screening in B by a factor $l/(v/\omega)$. In the same way we obtain the modulus of the screening in region D of the anomalous skin effect ($\delta \ll l$, v/ω) by reducing the screening in B by a factor $\delta/(v/\omega)$.

Since Pippard's non-local theory for the superconductor runs formally parallel with the theory of the anomalous skin effect, it should be possible to find the equations for λ_s along similar lines. Apart from a few fine points such as the factor $2/b$ in (38), this is quite

Table II. The penetration depth in tin and aluminium as found experimentally and in various theoretical approximations; after Bardeen, Cooper and Schrieffer^[10]. λ_L is the London penetration depth, $\lambda_s(P)$ the penetration depth in the Pippard limit, $\lambda_s(th)$ the penetration depth predicted by the complete non-local theory and $\lambda_s(exp)$ the experimental value. The penetration depths and the coherence length, ξ_0 , are expressed in nm. v is the Fermi velocity, T_c the critical temperature.

	v	T_c	ξ_0	λ_L	$\lambda_s(P)$	$\lambda_s(th)$	$\lambda_s(exp)$
Sn	0.69×10^6 m/s	3.73 °K	250	35	44	57	51
Al	1.26×10^6 m/s	1.18 °K	1500	16	47	52	49

easily done. The London screening, $1/\lambda_L^2$, is to be expected in a superconductor that is not dirty and in which no non-local effects occur, in other words a superconductor where $\xi_0 \ll l$ and $\xi_0 \ll \lambda_s$ (the intrinsic London superconductor). For the screening in a "dirty" superconductor ($l \ll \xi_0$, λ_s) we find by applying the above procedure, now with reduction factor l/ξ_0 , that $1/\lambda_s^2 = (l/\xi_0) (1/\lambda_L^2)$, in agreement with (37). In a superconductor in the non-local limit ($\lambda_s \ll \xi_0$, l) we find $1/\lambda_s^2 = (\lambda_s/\xi_0) (1/\lambda_L^2)$, in agreement with (38), except for a factor $2/b$.

The formulae for the superconductor in the three limiting cases are obtained in detail from the corresponding equations for the skin effect in normally conducting metals by the substitutions:

$$\delta \rightarrow j\lambda_s, v/\omega \rightarrow j\xi_0, (l \rightarrow l).$$

The fact that ξ_0 in the superconductor corresponds to $-jv/\omega$ in the normal conductor can also be seen from a comparison of expression (31) with the corresponding expression (II,39), provided we make allowance in the latter expression for the "retardation" (see II, p. 310). To do so we must substitute for E in (II,39) the field at the time $t - r/v$, when calculating the current at the moment t . If current and field depend on time as $\exp(j\omega t)$, we then obtain a factor $\exp(-jr\omega/v)$ under the integral. The exponent of e under the integral is therefore $-r(1/l + j\omega/v)$ in the normal metal and $-r/l\xi = -r(1/l + 1/\xi_0)$ in the superconductor, which demonstrates the correspondence mentioned.

In fig. 12 we have classified superconductors according to their values of ξ_0 and l , by analogy with the ω - τ diagram in II, fig. 11 (reproduced here as fig. 1). We have plotted ξ_0 from right to left in order to bring out more clearly the analogy with fig. 11 in II. A' is the region of the "dirty" superconductors, B' that of the intrinsic London superconductors, and D' that of the superconductors in the Pippard limit. In the D' region $\xi > \lambda_s$; this corresponds to type I superconductors. The type II superconductors lie in A' en B' ($\lambda_s > \xi$).

Fig. 13 shows a number of superconductors in the ξ_0 - l diagram. Nearly all superconducting elements are to be found in D'. Examples are the elements Al and

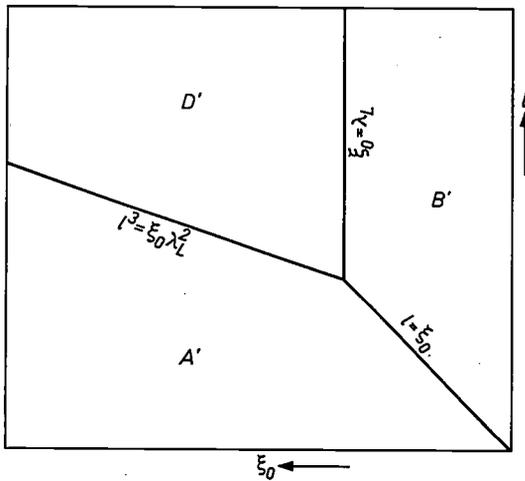


Fig. 12. The ξ_0 - l diagram for superconductors. In D' , well away from the boundaries, the Pippard limit applies (type I superconductors), in A' and B' the local limit (type II superconductors). A' is the region of "dirty" superconductors, B' the region of "intrinsic London superconductors". In A' the penetration depth λ_s is given by (37), in B' by (36) and in D' by (38). The figure is formally analogous with II, fig. 11, apart from region C in the latter diagram.

Sn indicated in the diagram. The superconducting alloys, which have proved to be much more important for practical applications, lie in A' . Intrinsic London superconductors are almost non-existent: the B' region is virtually empty. Interesting in this connection are the elements V and Nb, which are also indicated: they lie in the border regions between D' and B' . Using more rigorous criteria than we have employed here

(and which shift the boundary between D' and B' by a factor of $\sqrt{2}$ to the left) it can be shown to be a reasonable assumption that they really lie in B' [25]: they constitute the only examples known thus far of intrinsic London superconductors. To be such, however, they have to be of exceptional purity: an "ordinary" piece of material comes into region A' .

A striking difference between the ξ_0 - l and the ω - τ diagram (fig. 1) is that, for a given element, there is only one degree of freedom in the first case (l) but two in the second case (ω and τ). In the ξ_0 - l diagram one can only move vertically via the degree of impurity, whereas in the ω - τ diagram one can move horizontally as well (frequency of the field). Furthermore, the "filling" of the regions in the two cases is quite different. Admittedly A and A' are both to be regarded as the regions of "practical" materials for electrical engineering purposes. However, whereas with the normal metals it was possible to reach region D from region A only in carefully performed experiments, the first superconductors arrived in D' and it was not until later that the alloys in region A' began to be explored. And, while every shining metal surface bears witness to the reflection in the relaxation region B, E , region B' is practically empty.

Reversal of the sign of the field in the penetration layer

The non-local theory has a further remarkable consequence as regards the magnetic field in the superconductor. In the simple configuration shown by fig. 1 in part II (J // x -axis, H // y -axis, z -axis \perp surface) one can write for Maxwell's first relation (I,1):

$$\frac{1}{\mu} \frac{\partial^2 A}{\partial z^2} = -J,$$

that is to say, if A is plotted against z then the curvature of the curve is given by J . If, now, a local relation exists between J and A , e.g. of the type (11b) or (32), i.e. J is proportional to A but of opposite sign, the curvature is proportional to $+A$, so that the convex side is always directed towards the axis. This implies that, for a given A at the edge, there is only one value of dA/dz at the edge that shows the correct behaviour (see fig. 14a): if dA/dz is too large, A is deflected so quickly from the axis that it goes towards $+\infty$ for $z \rightarrow \infty$; if dA/dz is too small, it overshoots 0 and bends away from the axis on the negative side. The only acceptable solution is the exponential one.

If the relation between J and A is a non-local one, however, the curvature of A is proportional to the mean value of A over a region around the point under consideration. Close to the surface this will turn out to be greater than A at the point itself. Initially, therefore, there will be a greater curvature than in the first case, so that: 1) if A does not reach zero it will certainly bend away from the axis again, and 2) if A does go to zero there will still be some positive curvature left at the point where $A = 0$ (see fig. 14b). The acceptable solution, therefore, passes through zero. A consequence of the non-local relation is therefore that the vector potential A (and likewise the field H) undergoes a reversal of sign in the penetration layer.

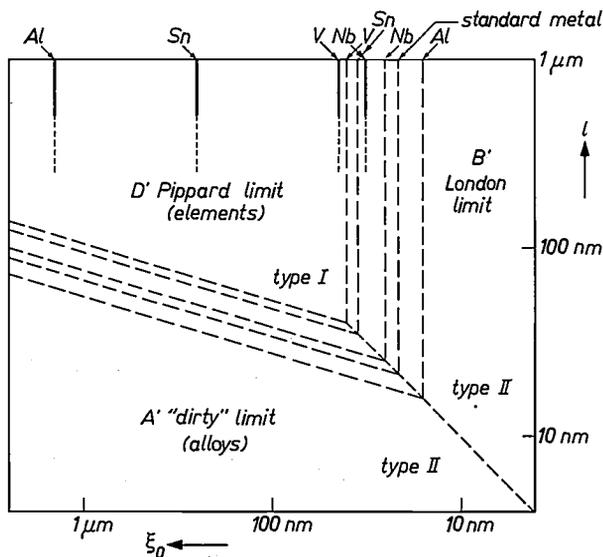


Fig. 13. The ξ_0 - l diagram for the standard metal and for Al, Sn, V and Nb. The boundaries (dashed lines) between areas A' , B' and D' are entirely governed by the value of $\lambda_L(0) = \lambda_p$ of the material. The bold vertical lines at the top represent the clean materials. As the metal becomes more impure it moves vertically downwards in the diagram. The values of $\lambda_L(0)$ and ξ_0 for Al and Sn are from Table I and for Nb and V from Radebaugh and Keesom [25] and Stromberg and Swenson [26]. Further details are explained in the text.

[25] R. Radebaugh and P. H. Keesom, Phys. Rev. 149, 217, 1966.

[26] T. F. Stromberg and C. A. Swenson, Phys. Rev. Letters 9, 370, 1962.

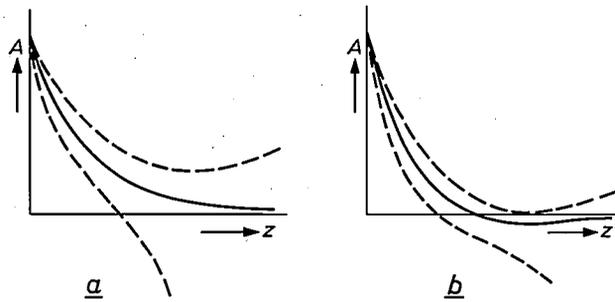


Fig. 14. The vector potential A as a function of the depth z in the superconductor, a) in the local theory, b) in the non-local theory.

An attempt to verify this, and with it the non-local theory in a fairly direct way, has been made by Drangeid and Sommerhalder [27], with positive results. In their experiment a hollow cylinder (a film of tin deposited on glass, thickness $1.87 \mu\text{m}$) was placed in an alternating field parallel to the axis of the cylinder. The intensity of the alternating field was very low (maximum 30 Oe) in order to avoid complications, and the frequency ($\approx 10^5$ Hz) was so low that the superconductor was effectively subject to a d.c. field (i.e. the first term in (20) is negligible). The field inside the cylinder was measured with a coil. This field, which was extremely weak (10^{-7} times the external field, in agreement with the theoretical prediction) was indeed found to reverse its phase below a certain temperature, and thus to have the opposite sign from that of the external field.

The amended theory of the surface resistance

We shall now apply the two amendments (a new value for λ_s and a new value for τ) to expression (24) for δ' , confining ourselves to superconductors of type I. Let the frequency be so high that the normal metal (just above the critical temperature) is in the anomalous region.

The new value for λ_s is given by (38). For τ we must substitute the time during which the field is in unperturbed interaction with the electron; in the present case this is the time in which the electron traverses the skin layer, and is roughly $b\lambda_s/v$. After substitution this gives:

$$\delta' = (b/2) (\omega/v) \lambda_s^4 / \lambda_n^2.$$

The factor $1/\lambda_n^2$ is proportional to the concentration of the normal electrons; according to (18) and (14), $1/\lambda_n^2 = \vartheta^4/\lambda_p^2$. Experimental evidence — and also the more elaborate theory — shows that the dependence of the penetration depth on temperature is reasonably represented by prediction (17): $\lambda_s \propto (1 - \vartheta^4)^{-1/2}$. Keeping to this temperature dependence, and using (38) and $\lambda_I(\vartheta = 0) = \lambda_p$, we find:

$$\lambda_s = (2\xi_0\lambda_p^2/b)^{1/3} (1 - \vartheta^4)^{-1/2}.$$

For δ' this gives:

$$\delta' = (\omega/v) (2\xi_0^4\lambda_p^2/b)^{1/3} \Phi_2(\vartheta),$$

where

$$\Phi_2(\vartheta) = \vartheta^4/(1 - \vartheta^4)^2. \dots (39)$$

The function $\Phi_2(\vartheta)$, along with $\Phi_P(\vartheta)$ and $\Phi_1(\vartheta)$, is plotted against ϑ in fig. 7. With

$$(26): q = \delta'/\delta_n' = A(\omega)\Phi(\vartheta),$$

$$(II,40-41): \delta_n' = (v\lambda_p^2/4b\omega)^{1/3},$$

$$(34): \xi_0 = ahv/kT_c$$

we find that

$$A(\omega) = 2(ah\omega/kT_c)^{4/3}. \dots (40)$$

This is a nice result, because we have now found an "explanation" for the fact that — according to fig. 8 — $A(\omega)$ as a function of the "reduced frequency" $h\omega/kT_c$ is not very dependent on the impurity of the metal or on the choice of metal. Not only the sensitive parameter l , but also such characteristic parameters of the metal as λ_p , v , ξ_0 , and even the parameter b , which depends on the manner in which an electron is reflected at the surface, have disappeared from the equation. According to (40), $A(\omega)$ is a universal function of $h\omega/kT_c$ [9]. In the BCS theory, a is a constant: $a = 0.18$ (eq. 34). We recall that the calculation given relates only to "extreme superconductors of type I" which, in the normal state, just above T_c , are situated in the extreme anomalous region.

Expression (40) is only a rough application of the non-local theory, and the agreement with experiment is quantitatively not particularly good. A plot of $A(\omega)$ in accordance with (40), for $a = 0.18$, is given in fig. 8 (line 1). Mattis and Bardeen [28] have presented a theoretical treatment of the anomalous skin effect for normal and superconducting metals on the basis of the BCS theory, and this theory has been worked out in more detail by Miller [29]. Curve 2 in fig. 8 is Miller's result for the non-local limit for tin. The discrepancy between curve 2 and the experimental results indicates that the non-local limit for most metals is not a very good approximation. Miller calculated the corrections to this limit for tin and aluminium, and his results largely explain the discrepancy.

At high frequencies the analysis can no longer be used. The hypothesis (26) that q may be written as a product of a function of ω only and a function of ϑ only is then no longer valid. This appears, for example, in the spread of the points for aluminium in fig. 8; this is not due to spread in the direct results of the measurements. According to the BCS theory, however, the sharp rise of q — reflected in the sharp rise of $A(\omega)$ in fig. 8 — at $T = 0$ does occur in all metals at the same value of $h\omega/kT_c$, namely at $h\omega/kT_c = 3.52$. This part of the curve too, therefore, has a universal character.

Superconductive resonant cavities

As regards the applications of superconductivity a great deal of interest is centred on type II supercon-

ductors in the mixed state (e.g. for producing strong magnetic fields or for the large-scale transport of electrical energy without losses [2]). These fall outside our subject because the field penetrates deep into the material. Applications in which the field penetration is limited to a skin layer are low-loss microwave cavity resonators and waveguides [30].

The construction of a high- Q cavity resonator involves the following considerations. A high Q is equivalent to low absorption in the wall, and therefore to a small surface resistance. It can be seen from fig. 7 that T should be small with respect to T_c , and from fig. 8 that ω should be small with respect to kT_c/h . If the T that can be reached is given, or if the ω is fixed by the problem, then a material with a large T_c must be chosen. According to (26) and (28), q should go to zero for $\phi \rightarrow 0$ but in practice, according to (25) the limit of the surface resistance is determined by the "residual value" R_0 . This depends to a very great extent on the state of the surface. Much of the research done on superconductive resonant cavities has therefore been directed towards finding the best method of surface treatment.

As an example of what has been achieved in this field we mention the work of Wilson [31] and that of Zimmer [32]. Wilson investigated copper cavities with a surface layer of tin or lead at a frequency of 2856 MHz. The best results were obtained with unpolished electrolytic films. The highest Q he found was 5×10^7 for tin and 2×10^8 for lead. These values were reached at a temperature of 1.8 °K. They are a factor of between 10^3 and 10^4 greater than the Q of an identical cavity of copper at room temperature — a remarkable improvement compared with the factor of merely 4 that would at the most be achieved by using high-purity copper at low temperature (see II, p. 315). Application of eq. (25) to the results showed the relative residual resistance for tin to be $R_0/R_n = 0.003$ and for lead $R_0/R_n = 0.0013$. Using a rectangular copper cavity with an electrolytically deposited layer of lead at 4.2 °K and 9375 MHz, Zimmer found a Q of more than 8×10^6 . A Q of 10^7 to 10^8 therefore seems representative of the best that can be achieved at the present time.

One possible application of such a resonant cavity

[27] K. E. Drangeid and R. Sommerhalder, Phys. Rev. Letters 8, 467, 1962.

[28] D. C. Mattis and J. Bardeen, Phys. Rev. 111, 412, 1958.

[29] P. B. Miller, Phys. Rev. 118, 928, 1960.

[30] See E. Maxwell, Superconducting resonant cavities, Progress in Cryogenics 4, 123-158, 1964.

[31] P. B. Wilson, Nucl. Instr. Meth. 20, 336, 1963.

[32] H. Zimmer, in: Philips; unsere Forschung in Deutschland, Philips Zentrallaboratorium GmbH, Aachen-Hamburg 1964, p. 134, 136.

[33] See J. C. B. Missel, Philips tech. Rev. 11, 145, 1949/50.

[34] See the articles on this subject in this journal: D. W. Fry, The linear electron accelerator, Philips tech. Rev. 14, 1-12, 1952/53; and C. F. Bareford and M. G. Kelliher, The 15 million electron-volt linear electron accelerator for Harwell, Philips tech. Rev. 15, 1-26, 1953/54.

is for frequency stabilization. The frequency stability that can be achieved using a resonant cavity as reference element is inversely proportional to the Q . For comparison we note that quartz crystals, widely employed for frequency stabilization — at lower frequencies, not exceeding about 10 MHz — have Q values of between 10^4 and 10^5 [33].

Other possible applications have been investigated at the Philips laboratory in Hamburg. With only moderate power input an intense microwave field can be excited in a superconducting resonant cavity: this field can be used to set up electron field emission, and as this effect is non-linear, it can be used for frequency multiplication [32]. This study, interesting in itself, has not, however, yielded results that can match those obtained with conventional methods of frequency multiplication.

A highly topical application is the use of superconductive resonant cavities or waveguides in linear particle accelerators, e.g. electron accelerators. The particles in such accelerators are passed through a series of resonant cavities which are synchronized in such a way that a group of particles is always subject to an accelerating field, or they are passed through a waveguide so shaped as to keep a group of particles "surf-riding" on a travelling wave [34]. For a given accelerating field, the high-frequency power required is greater the higher the losses in the wall. To keep the *average* power within technically and economically acceptable bounds, most existing accelerators are pulsed, so that high-energy particles are delivered only for, say, 1% of the time. Such considerations are governed by the capacity of the cooling system as well as the availability of high-frequency energy sources. By drastically cutting down the wall losses with the aid of superconductors it might well be possible, using a much smaller average power, to produce particles of the same energy continuously.

Summary. This last part of a series of articles on the skin effect deals with the penetration of an electromagnetic field in the surface of a superconductor. A simple model of the superconductor is the two-fluid model in which a fraction of the electrons is "superconducting" and the remainder "normal". Instead of Ohm's law, the London equations apply to the supercurrent; one of these states that the supercurrent is proportional to the vector potential. In this model the superconducting electrons are responsible for the screening effect, even at zero frequency, while the normal electrons in the skin layer thus formed are responsible for the absorption that occurs only at non-zero frequency. The "London penetration depth" is independent of frequency. Among the amendments to this model special attention is paid to that of Pippard, in which the London relation between current and vector potential is replaced by a non-local relation. This theory runs parallel with that of the anomalous skin effect, discussed in part II. A characteristic quantity occurring in Pippard's theory (and elsewhere) is the "coherence length". By comparing this with the penetration depth it is possible to classify superconductors into different types. Experimental results relating both to the penetration depth and the surface absorption support this refinement of the theory at several points. An application of this aspect of superconductivity, in which the field penetration is limited to a skin layer, is to be found in the fabrication of high- Q resonant cavities. Its applications in linear accelerators are at present under serious consideration.

Recent scientific publications

These publications are contributed by staff of laboratories and plants which form part of or co-operate with enterprises of the Philips group of companies, particularly by staff of the following research laboratories:

Philips Research Laboratories, Eindhoven, Netherlands	<i>E</i>
Mullard Research Laboratories, Redhill (Surrey), England	<i>M</i>
Laboratoires d'Electronique et de Physique Appliquée, Limeil-Brévannes (S.O.), France	<i>L</i>
Philips Zentrallaboratorium GmbH, Aachen laboratory, Weisshausstrasse, 51 Aachen, Germany	<i>A</i>
Philips Zentrallaboratorium GmbH, Hamburg laboratory, Vogt-Kölln-Strasse 30, 2 Hamburg-Stellingen, Germany	<i>H</i>
MBLE Laboratoire de Recherche, 2 avenue Van Becelaere, Brussels 17 (Boitsfort), Belgium.	<i>B</i>

Reprints of most of these publications will be available in the near future. Requests for reprints should be addressed to the respective laboratories (see the code letter) or to Philips Research Laboratories, Eindhoven, Netherlands.

- E. A. Aagaard:** Time division connection network for combined speech and large-signal transfer. Communications présentées au Colloque International de Commutation Electronique, Paris 1966, p. 289-300. *E*
- G. A. Acket:** Determination of the negative differential mobility of n-type gallium arsenide using 8 mm micro-waves. Physics Letters **24A**, 200-202, 1967 (No. 4). *E*
- W. Albers, G. van Aller & C. Haas:** Band structure and electrical properties of ternary chromium chalcogenides. Coll. int. Centre Nat. Recherche Scient. No. 157 (Dérivés semi-métalliques, Orsay 1965), 19-29, 1967. *E*
- L. K. H. van Beek, J. Helfferich, H. Jonker & Th. P. G. W. Thijssens:** Properties of diazosulfonates, Part I. The dissociation of methoxybenzenediazosulfonates. Rec. Trav. chim. Pays-Bas **86**, 405-409, 1967 (No. 4). *E*
- H. J. Bensiack & H. Severin:** Ebener Reflektor mit variablem Reflexionsfaktor für elektromagnetische Zentimeterwellen. Nachrichtentechn. Z. **20**, 280-286, 1967 (No. 5). *H*
- M. Berth & R. Petit:** Tube analyseur d'images à mosaïque de germanium. Acta electronica **10**, 123-135, 1966 (No. 2). *L*
- M. Berth & C. Venger:** Technologie des mosaïques de photodiodes. Acta electronica **10**, 157-180, 1966 (No. 2). *L*
- G. Blasse:** Concentration quenching of Eu^{3+} fluorescence. J. chem. Phys. **46**, 2583-2585, 1967 (No. 7). *E*
- G. Blasse:** Crystal structure and fluorescence of compositions ATiNbO_6 , ATiTaO_6 and ATiSbO_6 . Mat. Res. Bull. **2**, 497-502, 1967 (No. 5). *E*
- G. Blasse & A. Brill:** Crystal structure and fluorescence of some lanthanide gallium borates. J. inorg. nucl. Chem. **29**, 266-267, 1967 (No. 1). *E*
- G. Blasse & A. Brill:** An example of disagreement between site symmetry and the splitting of Eu^{3+} emission lines. Solid State Comm. **5**, 1-3, 1967 (No. 1). *E*
- G. Blasse & A. Brill:** Fluorescence of Eu^{3+} -activated garnets containing pentavalent vanadium. J. Electrochem. Soc. **114**, 250-252, 1967 (No. 3). *E*
- G. Blasse & A. Brill:** Fluorescence of Eu^{3+} -activated lanthanide oxyhalides LnOX . J. chem. Phys. **46**, 2579-2582, 1967 (No. 7). *E*
- G. Blasse & A. Brill:** Some observations on the Cr^{3+} fluorescence in the huntite structure. Phys. Stat. sol. **20**, 551-556, 1967 (No. 2). *E*
- R. Bleekrode:** Negative absorption in chemically reacting systems: flames and explosions. J. Chimie phys. **64**, 141-146, 1967 (No. 1). *E*
- P. F. Bongers & E. R. van Meurs:** Ferromagnetism in compounds with pyrochlore structure. J. appl. Phys. **38**, 944-945, 1967 (No. 3). *E*
- G. A. Bootsma & F. Meyer:** Ellipsometric investigation of the film growth at the germanium electrolyte interface. Surface Sci. **7**, 250-254, 1967 (No. 2). *E*
- G.-A. Boutry & P. Lebon:** Divers régimes d'une décharge électrique du type Penning aux très basses pressions. C.R. Acad. Sci. Paris **264B**, 519-522, 1967 (No. 7). *L*
- G. Bouwhuis:** Optische Methoden bei der experimentellen Untersuchung nichtlinearer Vorgänge im Gaslaser. Optik **25**, 294-298, 1967 (No. 3). *E*
- G. Brédart & Ph. van Bastelaer:** Les paramètres matriciels du transistor et leur emploi dans l'étude du comportement transitoire des circuits transistorisés (cas des petits signaux). 2e partie: Etude du comportement transitoire des circuits transistorisés. Rev. MBLÉ **10**, 33-42, 1967 (No. 1). *B*

- J. C. Brice:** The effect of arsenic pressure on crystal efficiency for injection luminescence in gallium arsenide. *Solid-State Electronics* **10**, 335-337, 1967 (No. 4). *M*
- J. C. Brice, J. A. Roberts & G. Smith:** Mass spectrometric studies of impurities in gallium arsenide crystals. *J. Mat. Sci.* **2**, 131-138, 1967 (No. 2). *M*
- J. C. Brice & P. A. C. Whiffin:** Solute striae in pulled crystals of zinc tungstate. *Brit. J. appl. Phys.* **18**, 581-585, 1967 (No. 5). *M*
- A. Broese van Groenou:** Low frequency magnetic relaxations in manganese ferrous ferrites at low temperatures. *J. Phys. Chem. Solids* **28**, 325-331, 1967 (No. 2). *E*
- A. Broese van Groenou, J. A. Schulkes & D. A. Annis:** Magnetic anisotropy of some nickel zinc ferrite crystals. *J. appl. Phys.* **38**, 1133-1134, 1967 (No. 3). *E, M*
- K. H. J. Buschow & W. A. J. J. Velge:** Phase relations and intermetallic compounds in the lanthanum-cobalt system. *J. less-common Met.* **13**, 11-17, 1967 (No. 1). *E*
- K. H. J. Buschow & J. H. N. van Vucht:** Systematic arrangement of the binary rare-earth-aluminium systems. *Philips Res. Repts.* **22**, 233-245, 1967 (No. 3). *E*
- H.-J. Butterweck:** Frequenzabhängige nichtlineare Übertragungssysteme. *Archiv elektr. Übertr.* **21**, 239-254, 1967 (No. 5). *E*
- T. D. Clark:** High energy structure in the d^2I/dV^2 versus V characteristic of metal-insulator-thallium junctions. *Physics Letters* **24A**, 459-460, 1967 (No. 9). *M*
- T. D. Clark & J. P. Baldwin:** Superconducting memory device using Josephson junctions. *Electronics Letters* **3**, 178-179, 1967 (No. 5). *M*
- J. P. Cosier & R. F. Pearson:** Low-temperature permeability measurements on ferrites. *Brit. J. appl. Phys.* **18**, 615-620, 1967 (No. 5). *M*
- A. J. Dekkers, A. van Duuren, F. A. Lootsma & J. Vlietstra:** Berekening van verkeerslichtenprogramma's met behulp van lineaire programmering. *Ingenieur* **79**, V 11-17, 1967 (No. 8). *E*
- A. van der Drift:** Evolutionary selection, a principle governing growth orientation in vapour-deposited layers. *Philips Res. Repts.* **22**, 267-288, 1967 (No. 3). *E*
- W. F. Druyvesteyn:** Collective modes in type II superconductors. *Physics Letters* **24A**, 415-416, 1967 (No. 8). *E*
- G. Engelsma:** Effect of cycloheximide on the inactivation of phenylalanine deaminase in gherkin seedlings. *Naturwiss.* **54**, 319-320, 1967 (No. 12). *E*
- A. J. Fox & J. R. Mansell:** Tunable microwave-frequency light modulator. *Proc. Instn. Electr. Engrs.* **114**, 741-744, 1967 (No. 6). *M*
- J. G. C. de Gast:** A new type of controlled restrictor (M.D.R.) for double film hydrostatic bearings and its application to high-precision machine tools. *Advances in machine tool design and research 1966* (Proc. 7th Int. M.T.D.R. Conf., Univ. Birmingham), p. 273-298; Pergamon Press, Oxford 1967. *E*
- P. Gerthsen, E. Kauer & H. G. Reik:** Halbleitung einiger Übergangsmetalloxide im Polaronenbild. *Festkörperprobleme* **5**, 1-56, 1966. *A*
- P. J. van Gerwen:** Application of pseudo-ternary codes for data transmission. *Progresso elettronico 1966* (Atti XIII Congr. per l'Elettronica, Roma), Vol. I, p. 463-477. *E*
- A. A. van der Giessen:** Magnetic properties of ultra-fine iron (III) oxide-hydrate particles prepared from iron (III) oxide-hydrate gels. *J. Phys. Chem. Solids* **28**, 343-346, 1967 (No. 2). *E*
- J.-M. Goethals:** Factorization of cyclic codes. *IEEE Trans. on information theory IT-13*, 242-246, 1967 (No. 2). *B*
- J. L. Goldstein:** Auditory nonlinearity. *J. Acoust. Soc. Amer.* **41**, 676-689, 1967 (No. 3). *E*
- A. H. Gomes de Mesquita & K. H. J. Buschow:** The crystal structure of so-called α -LaAl₄ (La₃Al₁₁). *Acta cryst.* **22**, 497-501, 1967 (No. 4). *E*
- H. C. de Graaff & J. A. van Nielen:** Temperature influence on the channel conductance of m.o.s. transistors. *Electronics Letters* **3**, 195-196, 1967 (No. 5). *E*
- C. A. A. J. Greebe:** Mechanical excitation of helicon waves. *Philips Res. Repts.* **22**, 133-141, 1967 (No. 2). *E*
- H. G. Grimmeiss:** Elektrolumineszenz in III-V-Verbindungen. *Festkörperprobleme* **5**, 221-248, 1966. *A*
- G. J. van Gorp:** Temperature dependence of the critical field in superconducting vanadium. *Phys. Stat. sol.* **19**, 173-176, 1967 (No. 1). *E*
- W. van Haeringen:** The effect of saturation on the polarization parameters of modes in anisotropic gas lasers. *Physics Letters* **24A**, 65-66, 1967 (No. 1). *E*
- J. Haisma:** Construction and properties of short stable gas lasers. *Thesis, Utrecht* 1966. *E*
- J. Haisma:** Gebruik van lasers. *Ingenieur* **79**, O 65-74, 1967 (No. 26). *E*
- N. Hansen:** Gerät zur automatischen Bestimmung der Oberflächengröße feinteiliger Substanzen. *Automatik* **11**, 253, 1966 (No. 7). *A*
- P. A. H. Hart:** Gas discharge as a source of incoherent radiation at millimetre and sub-millimetre wavelengths. *Philips Res. Repts.* **22**, 77-109, 1967 (No. 2). *E*
- H. F. van Heek:** Unusual electrode configuration for Hall measurements on thin films and field-effect devices. *Solid-State Electronics* **10**, 268-269, 1967 (No. 3). *E*
- J. C. M. Henning:** Covalency and hyperfine structure of $(3d)^5$ -ions in crystal fields. *Physics Letters* **24A**, 40-42, 1967 (No. 1). *E*

- J. C. M. Henning, J. Liebertz & R. P. van Stapele:** Evidence for Cr^{3+} in four-coordination: ESR and optical investigations of Cr-doped AlPO_4 crystals. *J. Phys. Chem. Solids* **28**, 1109-1114, 1967 (No. 7). *A, E*
- P. L. Holster:** Gaslagers met uitwendige drukbron, 1. Werking en eigenschappen van gaslagers met uitwendige drukbron, 2. Berekeningsmethode van aëro-statische lagers. *Polytechn. T. Werktuigbouw* **22**, 363-370, 415-425, 1967 (Nos. 9, 10). *E*
- G. J. van Hoytema, W. J. Oosterkamp, A. M. C. van den Broek & A. Druppers:** La neuroradiologie avec sous-traction utilisant la télévision et une mémoire d'image magnétique. *Neuro-Chirurgie* **12**, 801-809, 1966 (No. 7). *E*
- G. H. Jonker:** Kristalchemie en kristalfysica ten dienste van de elektrotechniek. *Chem. Weekblad* **63**, 81-86, 1967 (No. 8). *E*
- G. H. Jonker:** Optimale keramische Struktur für verschiedene Anwendungen von Bariumtitanat. *Ber. Dtsch. Keram. Ges.* **44**, 265-266, 1967 (No. 6). *E*
- W. Kischio:** Über Borphosphid. *Z. anorg. allgem. Chemie* **349**, 151-157, 1967 (No. 3/4). *A*
- R. J. Klein Wassink:** Wetting of solid-metal surfaces by molten metals. *J. Inst. Met.* **95**, 38-43, 1967 (No. 2). *E*
- M. Klerk & E. Roeder:** Der Elektronenstrahl-Mikro-analysator, ein analytisches Hilfsmittel für den Chemiker. *Chemie-Ing.-Technik* **39**, 567-574, 1967 (No. 9/10). *A, E*
- S. Kornblum & W. G. Koster:** The effect of signal intensity and training on simple reaction time. *Acta psychol.* **27**, 71-74, 1967. *E*
- W. G. Koster & J. A. M. Bekker:** Some experiments on refractoriness. *Acta psychol.* **27**, 64-70, 1967. *E*
- W. Kwestroo, J. de Jonge & P. H. G. M. Vromans:** Influence of impurities on the formation of red and yellow PbO . *J. inorg. nucl. Chem.* **29**, 39-44, 1967 (No. 1). *E*
- W. Kwestroo, C. Langereis & H. A. M. van Hal:** Basic lead nitrates. *J. inorg. nucl. Chem.* **29**, 33-38, 1967 (No. 1). *E*
- P. van der Laan & J. Oosterhoff:** Experimental determination of the power functions of the two-sample rank tests of Wilcoxon, Van der Waerden and Terry by Monte Carlo techniques, I. Normal parent distributions. *Statistica neerl.* **21**, 55-68, 1967 (No. 1). *E*
- H. de Lang:** Magnetic phenomena in gas lasers. *Physica* **33**, 163-173, 1967 (No. 1). *E*
- M. Lemke:** Wellentypen im ferritgefüllten Rechteckhohlleiter mit transversaler Magnetisierung parallel zur Hohlleiterbreite. *Archiv elektr. Übertr.* **21**, 440-446, 1967 (No. 8). *H*
- B. Lersmacher, H. Lydtin, W. F. Knippenberg & A. W. Moore:** Thermodynamische Betrachtungen zur Kohlenstoffabscheidung bei der Pyrolyse gasförmiger Kohlenstoffverbindungen. *Carbon* **5**, 205-217, 1967 (No. 3). *A, E*
- P. R. Locher:** Cu-NMR in paramagnetic and ferromagnetic CuCr_2Se_4 . *Solid State Comm.* **5**, 185-187, 1967 (No. 3). *E*
- C. Lockyer:** A new interferometric method for measuring the thickness of thick films. *J. sci. Instr.* **44**, 393-394, 1967 (No. 5). *M*
- F. A. Lootsma:** Logarithmic programming: a method of solving nonlinear-programming problems. *Philips Res. Repts.* **22**, 329-344, 1967 (No. 3). *E*
- F. K. Lotgering & R. P. van Stapele:** Magnetic and electrical properties of copper containing sulphides and selenides with spinel structure. *Solid State Comm.* **5**, 143-146, 1967 (No. 2). *E*
- H. Lydtin:** Eine einfache Methode zur Bestimmung des Bodenkörpertransportes über die Gasphase. *Z. Naturf.* **22a**, 571, 1967 (No. 4). *A*
- M. H. van Maaren & G. M. Schaeffer:** Some new superconducting group V^a dichalcogenides. *Physics Letters* **24A**, 645-646, 1967 (No. 12). *E*
- G. Marie:** Un nouveau dispositif de restitution d'images utilisant un effet électro-optique: le tube TITUS. *Philips Res. Repts.* **22**, 110-132, 1967 (No. 2). *L*
- R. Memming & G. Neumann:** Formation of surface states and radicals on Ge electrodes. *Physics Letters* **24A**, 19-21, 1967 (No. 1). *H*
- L. Merten:** Zur Ultrarot-Dispersion zweiachsiger und einachsiger Kristalle, I. *Z. Naturf.* **22a**, 359-364, 1967 (No. 3). *A*
- H. J. G. Meyer, G. Ahsmann, J. W. van der Laarse & E. H. A. M. van der Wee:** A new class of low-pressure arc columns with positive V - I characteristics. *Philips Res. Repts.* **22**, 209-218, 1967 (No. 3). *E*
- E. J. Millett:** The balanced substitution of mixed valency ions in oxide crystals. *Brit. J. appl. Phys.* **18**, 357-358, 1967 (No. 3). *M*
- B. J. Mulder:** Anisotropy of light absorption and exciton diffusion in anthracene crystals determined from the externally sensitized fluorescence. *Philips Res. Repts.* **22**, 142-149, 1967 (No. 2). *E*
- B. J. Mulder & J. de Jonge:** Electronic absorption spectrum of holes in anthracene. *Solid State Comm.* **5**, 203-205, 1967 (No. 4). *E*
- J. Neirynek & Ph. van Bastelaer:** La synthèse des filtres par factorisation de la matrice de transfert. *Rev. MBLE* **10**, 5-32, 1967 (No. 1). *B*
- J. Neirynek & J. P. Thiran:** Sensitivity analysis of lossless 2-ports by uniform variations of the element values. *Electronics Letters* **3**, 74-76, 1967 (No. 2). *B*
- D. J. van Ooijen:** Dynamics of the magnetization of a thin-walled superconducting cylinder in a parallel field. *Philips Res. Repts.* **22**, 219-232, 1967 (No. 3). *E*

- D. J. van Ooijen & A. S. van der Goot:** The internal friction of cold-worked niobium and tantalum containing both oxygen and nitrogen. Philips Res. Repts. **22**, 150-160, 1967 (No. 2). *E*
- J. J. C. Oomen:** On some electrical properties of colloidal systems. Thesis, Utrecht 1966. *E*
- G. W. van Oosterhout:** The transformation γ -FeO(OH) to α -FeO(OH). J. inorg. nucl. Chem. **29**, 1235-1238, 1967 (No. 5). *E*
- W. J. Oosterkamp:** A survey of scintillation camera systems. Progress in Radiology, XIth int. Congress, Rome 1965, Vol. II, p. 1150-1157; Excerpta Medica Foundation, Amsterdam 1967. *E*
- W. J. Oosterkamp, A. P. M. van 't Hof & W. J. L. Scheren:** Röntgenfarbbilder: neue Möglichkeiten der Subtraktion durch die Fernsehtechnik. Deutscher Röntgenkongress 1966, Berlin, Vol. A, p. 168-170; Thieme, Stuttgart 1967. *E*
- C. van Opdorp & H. K. J. Kanerva:** Current-voltage characteristics and capacitance of isotype heterojunctions. Solid-State Electronics **10**, 401-421, 1967 (No. 5). *E*
- A. Op het Veld & P. Bogers:** Eine Präparationsmethode zur Gefügeentwicklung von Hartmetallen (A preparation method for revealing the structure of hard metals). Prakt. Metallogr. **4**, 235-239 (240-242), 1967 (No. 5). *E*
- J. Th. G. Overbeek & G. W. Rathenau:** E. J. W. Verwey; bij zijn aftreden als directeur van het Philips Natuurkundig Laboratorium. Chem. Weekblad **63**, 3-7, 1967 (No. 1). *E*
- R. Petit:** Problèmes d'optique électronique dans les tubes analyseurs d'images infrarouges. Acta electronica **10**, 137-155, 1966 (No. 2). *L*
- L. J. van de Polder:** Target-stabilization effects in television pick-up tubes. Philips Res. Repts. **22**, 178-207, 1967 (No. 2). *E*
- J. C. M. A. Ponsioen & J. H. N. van Vucht:** The structure of β -TaCo₃ and the effect of the substitution of Ta and Co by related elements. Philips Res. Repts. **22**, 161-169, 1967 (No. 2). *E*
- A. Rabenau, E. Kauer & H. Klotz:** Sur l'hexaborure de lanthane LaB₆. Coll. int. Centre Nat. Recherche Scient. No. 157 (Dérivés semi-métalliques, Orsay 1965), 495-498, 1967. *A*
- A. Rabenau & H. Rau:** Kristallzüchtung mit Hilfe von chemischen Gleichgewichten zwischen zwei und mehr Phasen. Z. phys. Chemie Neue Folge **53**, 155-162, 1967 (No. 1-6). *A*
- H. Rau:** Defect equilibria in cubic high temperature copper sulfide (Digenite). J. Phys. Chem. Solids **28**, 903-916, 1967 (No. 6). *A*
- H. Rau & A. Rabenau:** Crystal syntheses and growth in strong acid solutions under hydrothermal conditions. Solid State Comm. **5**, 331-332, 1967 (No. 5). *A*
- H. G. Reik & D. Heese:** Frequency dependence of the electrical conductivity of small polarons for high and low temperatures. J. Phys. Chem. Solids **28**, 581-596, 1967 (No. 4). *A*
- H. G. Reik & R. Mühlstroh:** Phonon structures in the polaron enhanced light absorption at low temperatures. Solid State Comm. **5**, 105-108, 1967 (No. 2). *A*
- J. Revuz:** Variation du facteur de contraste dans les cibles à mosaïque de jonctions. Acta electronica **10**, 195-214, 1966 (No. 2). *L*
- C. Rooymans & W. Albers:** High pressure polymorphism of FeCr₂S₄ and related compounds. Coll. int. Centre Nat. Recherche Scient. No. 157 (Dérivés semi-métalliques, Orsay 1965), 63-66, 1967. *E*
- D. Saget:** Formation de mosaïques de jonctions de germanium *p-n* par épitaxie. Acta electronica **10**, 181-194, 1966 (No. 2). *L*
- J. G. van Santen:** Monolitische geïntegreerde schakelingen. Ingenieur **79**, E 37-42, 1967 (No. 10). *E*
- J. J. Scheer & J. van Laar:** Fermi level stabilization at cesiated semiconductor surfaces. Solid State Comm. **5**, 303-306, 1967 (No. 4). *E*
- H. Schemmann:** Wirkungsgrad von kleinen zweipoligen Gleichstrommotoren mit dauermagnetischem Läufer. Elektrotechn. Z. A **88**, 225-229, 1967 (No. 9). *A*
- W. Schilz:** Influence of the sample geometry on frequency and *Q*-value of the helicon resonance in PbTe. Solid State Comm. **5**, 503-507, 1967 (No. 7). *H*
- U. J. Schmidt:** Anwendungen und Stand der digitalen Lichtstrahlableitung. Int. elektron. Rdsch. **21**, 165-168, 1967 (No. 7). *H*
- K. J. Schmidt-Tiedemann:** On the realizability of traditors. Proc. IEEE **55**, 554-555, 1967 (No. 4). *H*
- K. J. Schmidt-Tiedemann:** Improved electronic half-shadow method for the measurement of birefringence and dichroism. Rev. sci. Instr. **38**, 625-627, 1967 (No. 5). *H*
- G. R. Schodder:** Akustische Verstärker. Funkschau **38**, 175-178, 1966. *A*
- J. F. Schouten & J. A. M. Bekker:** Reaction time and accuracy. Acta psychol. **27**, 143-153, 1967. *E*
- G. Schulten & J. P. Stoll:** A high-precision wide-band wavemeter for millimetre waves. Philips Res. Repts. **22**, 309-314, 1967 (No. 3). *H*
- E. Schwartz:** Streu- oder Betriebsmatrizen mit passiven und aktiven Bezugsimpedanzen. Archiv elektr. Übertr. **21**, 317-320, 1967 (No. 6). *A*
- E. Schwartz:** Über nichtreziproke, symmetrische Dreieck und Zirkulatoren. Nachrichtentechn. Z. **20**, 329-332, 1967 (No. 6). *A*
- A. M. J. H. Seuter:** The electrical resistivity of MnTe at elevated temperatures. Coll. int. Centre Nat. Recherche Scient. No. 157 (Dérivés semi-métalliques, Orsay 1965), 459-464, 1967. *E*

- M. J. Sparnaay:** Semiconductor surfaces and the electrical double layer.
Adv. Colloid Interface Sci. **1**, 277-333, 1967 (No. 3). *E*
- K. van Steensel:** Electrical conduction in island-structure films of gold and platinum on insulating substrates.
Philips Res. Repts. **22**, 246-266, 1967 (No. 3). *E*
- K. van Steensel, F. van de Burg & C. Kooy:** Thin-film switching elements of VO₂.
Philips Res. Repts. **22**, 170-177, 1967 (No. 2). *E*
- J. P. Stoll:** Vergleichende Messungen der komplexen skalaren Permeabilität an Kugeln und Stäben aus Ferriten mit hohen und niedrigen Verlusten bei Mikrowellen.
Z. angew. Physik **22**, 214-219, 1967 (No. 3). *H*
- T. L. Tansley & P. C. Newman:** Measurements of heterojunctions alloyed on to GaAs.
Solid-State Electronics **10**, 497-501, 1967 (No. 5). *M*
- N. C. de Troye:** De invloed van de micro-elektronica op de ontwikkeling van digitale schakelingen.
Ingenieur **79**, E 42-48, 1967 (No. 10). *E*
- J. Verweel:** Magnetic properties of ferrite single crystals with the Y structure.
J. appl. Phys. **38**, 1111-1117, 1967 (No. 3). *H*
- K. Walther:** Ultrasonic relaxation at the Néel temperature and nuclear acoustic resonance in MnTe.
Solid State Comm. **5**, 399-403, 1967 (No. 5). *H*
- W. L. Wanmaker, J. W. ter Vrugt & J. G. C. M. de Bres:** Luminescence of manganese-activated aluminium-substituted magnesium gallate.
Philips Res. Repts. **22**, 304-308, 1967 (No. 3).
- J. D. Wasscher:** Weak ferromagnetism in the NiAs structure and galvanomagnetic measurements on MnTe.
Coll. int. Centre Nat. Recherche Scient. No. 157 (Dérivés semi-métalliques, Orsay 1965), 465-471, 1967. *E*
- C. H. Weijnsfeld:** Yield, energy and angular distributions of sputtered atoms.
Thesis, Utrecht 1966. *E*
- H. O. Westermann:** Planning public lighting for X-town.
Int. Lighting Rev. **17**, 92-97, 1966 (No. 3). *A*
- M. V. Whelan:** On the nature of interface states in an SiO₂-Si system, and on the influence of heat treatments on oxide charge.
Philips Res. Repts. **22**, 289-303, 1967 (No. 3). *E*
- M. V. Whelan, A. H. Goemans & L. M. C. Goossens:** Residual stresses at an oxide-silicon interface.
Appl. Phys. Letters **10**, 262-264, 1967 (No. 10). *E*
- W. J. Witteman:** Dependence of radiation production and population densities on the thermal relaxation processes in a high power molecular laser.
J. Chimie phys. **64**, 107-110, 1967 (No. 1). *E*
- L. E. Zegers:** Error control in telephone channels by means of time diversity.
Philips Res. Repts. **22**, 315-328, 1967 (No. 3); Progresso elettronico 1966 (Atti XIII Congr. per l'Elettronica, Roma), Vol. I, p. 677-694. *E*
- A. L. Zijlstra & C. M. van der Burgt:** Isopaustic glasses for ultrasonic delay lines in colour television receivers and in digital applications.
Ultrasonics **5**, 29-38, 1967 (Jan.). *E*
- H. Zimmer:** Parametric amplification of microwaves in superconducting Josephson tunnel junctions.
Appl. Phys. Letters **10**, 193-195, 1967 (No. 7). *H*