PHILIPS TECHNICAL REVIEW

VOLUME 30

1969

PHILIPS TECHNICAL REVIEW

The Review is published by the Research Laboratories of N.V. Philips' Gloeilampenfabrieken, Eindhoven, Netherlands, and is devoted to the investigations, processes and products of the laboratories and plants which form part of or co-operate with enterprises of the Philips group of companies. In the articles the associated technical problems are treated along with their physical or chemical background. The Review covers a wide range of subjects, each article being intended not only for the specialist in the subject but also for the non-specialist reader with a general technical or scientific training.

The Review appears in three editions: English, Dutch and German, all identical in contents. There are twelve numbers per volume, each of about 32 pages. A yearly index is given for each volume and an index covering ten volumes appears every five years (the latest one is included in this volume).

Subject index, Volume 30, 1969

Author index, Volume 30, 1969

PHILIPS TECHNICAL REVIEW

VOLUME 30, 1969, No. 1

In Memoriam Ir. P. F. S. Otten

Ir. P. F. S. Otten, former President of N.V. Philips, died on 4th January, a few days after his 73rd birthday. His career with our Company began in 1924 when he joined as an electrical engineer, and ended at the Annual General Meeting in 1968, when he resigned as Chairman of the Supervisory Board. He was President from 1946 to 1961. The post-war recovery, reorganization and expansion of the Company took place under his leadership. Here it is appropriate to place on record our particular appreciation for the way in which Mr. Otten kept the interests of our research laboratories in mind and made sure that their development kept pace with the growth of the Company, even though he himself was not first and foremost an engineer but an entrepreneur.

The outside world knew his brilliant enthusiasm, at times bordering perhaps on over-exuberance, but can hardly have seen the true greatness of his leadership. Mr. Otten possessed the rare gift of being able to listen carefully and critically before taking his decisions — a quality surprising in a man of such restless energy. He was one from whom a rebuke was acceptable, because although expressed in no uncertain terms it was always objective. By the same token, a word of praise from him meant that much the more.

Mr. Otten will live in our memory as a man of contrasts: a man of impatient drive yet capable of calm analysis, sharply censorious yet just in his judgments, but always inspired by a sense of duty and dedication; in short, as a great man.

H. B. G. Casimir

A flywheel for stabilizing space vehicles

J. P. Reinhoudt

In the past huge "flywheel" devices have been used to stabilize ships at sea for the greater comfort of the passengers. A refined form of this stabilizing effect can now be applied, for rather different reasons, in unmanned space-craft.

A space vehicle in orbit is usually required to take up a definite attitude and to maintain it with a certain accuracy. This may be necessary, for example, to keep a telescope on board the vehicle accurately trained on a star, or to keep an aerial accurately trained on a transmitter or receiver stationed on the Earth, as in communication satellites.

Attitude stabilization can be brought about by means of an active attitude control system (passive control, using fields of force in space, will not be considered here). In this system a sensing system continuously observes the attitude of the vehicle; if the system detects an excessive deviation, caused by disturbing torques acting on the space vehicle, the attitude is corrected, for example by means of small control rockets.

Now although these outside torques generally fluctuate quite considerably, their mean values are relatively low. This makes it possible to keep the attitude of the vehicle practically constant with the aid of the gyro effect: a flywheel rotating in bearings rigidly attached to the vehicle gives the whole vehicle a resistance to changes in attitude which is proportional to the angular momentum of the wheel. This smooths out the attitude variations caused by the torque fluctuations, so that the control system only has to correct the effect of the mean disturbing torque. Another advantage is that the error in attitude will not become too great even though the sensing system is inoperative in certain parts of the orbit (e.g. when traversing the Earth's shadow).

When the direction of the axis of the flywheel is held constant in space, the vehicle can still be rotated about this axis. Continuous attitude correction will also be necessary for this degree of freedom, in order to keep a telescope or aerial properly directed. This correction can be effected by means of the reaction moments occurring when the angular velocity of the flywheel is increased or reduced. It is then left to a second, coarser system, using control rockets, to ensure that this angular velocity does not in the long run become too high or too low.

In what follows we shall not be concerned with the problem of stabilization as such [1], but with the design of the flywheel used for stabilization.

Closer examination of the flywheel design problem shows that everything turns — literally and figuratively — on the *bearings*. The advantages of an active attitude control system using a flywheel are particularly apparent in space vehicles that have to operate for a long period, but in such cases the flywheel must be exceptionally reliable in operation, and this depends critically upon the bearings. For communication satellites a working life of five years is already being contemplated. During this very long period the bearings must show no perceptible wear, and moreover they must be capable of withstanding the very large accelerations and vibration to which they are subject when the space vehicle is launched.

Extensive investigations carried out at Philips Research Laboratories in Eindhoven have shown that a *grease-lubricated hydrodynamic bearing system* is exceptionally well suited to this application. In hydrodynamically lubricated bearings (contactless bearings) the bearing surfaces are completely separated during rotation by a thin film of lubricant, so that no mechanical wear occurs and a very long life can be achieved [2]. The use of grease as lubricant has notable advantages over oil or gas in this respect. If gas is used there will be starting difficulties as there can be no question of the bearings being "contactless" when the wheel is first set in motion; and a gas-lubricated bearing may easily suffer damage when the vehicle is launched. If the lubricant used is oil, it is difficult to prevent the leakage of oil from the bearings before the wheel is set in motion (which is not until after the launching). Grease does not have these disadvantages: its consistency keeps it in the bearing, its buffer action protects the bearing from damage during launching, the starting torque

Ir. J. P. Reinhoudt is with Philips Research Laboratories, Eindhoven.

needed is not much greater than the small frictional torque during normal rotation, and the bearings can be designed so that the grease does not escape when the flywheel is spinning, which means that the bearings never need to be re-lubricated. This has been borne out by many experiments on greased bearings performed at these Laboratories [3].

In view of some of the factors quoted above it appears to us that *ball bearings* might not be very promising for this application. The risk of damage to ball bearings during launching is very real, and it is difficult to predict with any accuracy their useful life in uninterrupted operation. It would seem that a life of several years could only be obtained with ball bearings if the speed of revolution is relatively low, and that this condition would make the total mass of the flywheel system much greater than in the system we have designed.

The starting point in designing the flywheel is the angular momentum H, for which values of 10 to 30 Nms are required at the present time. From the equation

$$H = \text{const.} \times mR^2\omega \quad . \quad . \quad . \quad . \quad . \quad (1)$$

it can be seen that for a given H the mass m of the flywheel can be kept small — this is obviously desirable — by making the effective radius R large or the angular velocity ω high. However, as will presently be shown, both measures call for more power and hence for a larger energy source for driving the wheel. Moreover, a larger R entails a larger and heavier housing for the wheel. Both R and ω must therefore be chosen so as to minimize the *total* mass of flywheel, housing and driving gear and energy source.

Various forms of flywheel and bearings have been designed and built in our Laboratories. *Fig. 1* shows a simplified cross-section of a design for a wheel with an angular momentum of 20 Nms, which has an optimum speed of 1600 r.p.m. The wheel rotates on a shaft fixed at one end. The drive comes from an electric motor whose rotor forms a single assembly with the hub of the wheel. The wheel with bearing system and motor are contained in a housing filled with helium under low pressure (0.1 bar). Fixing the shaft at one end has the advantage that the housing can be made relatively light: it has no load-bearing function but simply provides the hermetic seal which is required when it is not desirable for the flywheel to operate in a vacuum for a long time.

The electric motor is of course also required to have a very long life, and in addition a sufficiently high starting torque, and a high efficiency (to limit the weight of the energy source needed). Its speed should

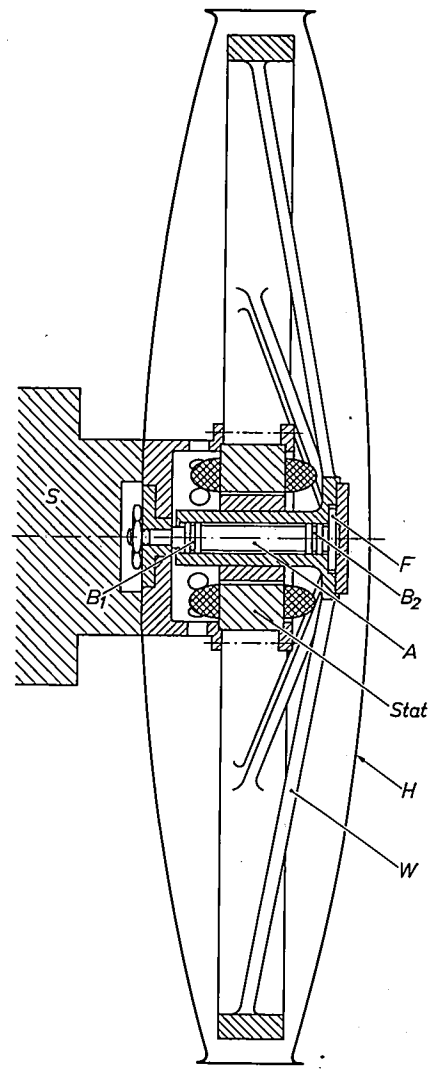


Fig. 1. Simplified cross-section of a flywheel, designed for an angular momentum of 20 Nms at an optimum speed of 1600 r.p.m. *A* shaft, fixed at one end to the body *S* of the space vehicle. *W* flywheel. *H* housing. *Stat* stator of electric motor. The flange *F* constitutes the thrust bearing for both directions of thrust, B_1 and B_2 are the journal bearings, all grease-lubricated.

be readily adjustable to give the variation in angular velocity, which, as we have seen, corrects for rotations, of the space vehicle around the axis of the flywheel. A motor which, goes a long way towards meeting these requirements is described in a separate article in this issue [4].

[1] A detailed treatment of this problem will be found in S. Glasstone, Source book on the space sciences, Van Nostrand, Princeton N.J. 1965, pp. 206-228.

[2] The principle of these bearings has been discussed by E. A. Muijderman, New forms of bearing: the gas and the spiral groove bearing, Philips tech. Rev. **25**, 253-274, 1963/64. See also E. A. Muijderman, Spiral groove bearings, thesis, Delft 1964.

[3] For a special design, see G. Remmers, Grease-lubricated spiral groove bearing for a straight-through shaft, Philips tech. Rev. **27**, 107-108, 1966.

[4] W. Radziwill, A highly efficient small brushless d.c. motor, Philips tech. Rev. **30**, 7-12, 1969 (No. 1).

We shall now take a closer look at the bearings of the flywheel.

To take up *axial* thrusts the end of the fixed shaft is provided with a flange (*F* in fig. 1), which has a pattern of spiral grooves on each face. The grooves supply the necessary load-carrying capacity through their hydrodynamic action [2]. These bearings can take up very large thrusts in both directions, which is necessary in order to be able to test the flywheel on the ground. The energy dissipation in this bearing is very low when the wheel is spinning under zero load (i.e. under conditions of weightlessness).

The *radial* thrust is taken up in cylindrical journal bearings. Special measures were found necessary here to avoid instability in the rotation of the flywheel. Owing to the state of weightlessness there will be hardly any load on the bearings when the flywheel is spinning, and it is well known that cylindrical journal bearings that are insufficiently loaded set up marked vibration in the system, known as "$\frac{1}{2}\omega$ whirl" [5]. Such vibration, which would be felt throughout the space vehicle, could of course be detrimental to the electronic equipment and might also cause damage to the bearings. The instability can be avoided by artificially pre-loading the journal bearings. We cannot go into much detail here, but we shall note that a stabilizing effect arises through the establishment during rotation of an equilibrium in which the shaft takes up a slightly *eccentric* position inside the cylindrical bearing (irrespective of which of the two is rotating). Conversely it can be said that such a position will result in a certain load, a "pre-load", on the bearing, which will equally occur in a state of weightlessness. This is exploited in our system as follows. The journal bearing on each end of the shaft is built up from *two* sections which possess a *built-in eccentricity*; see *fig. 2*. Each of these two bearing sections produces a pre-load between the shaft and the bearing bush which acts as a load on the other section, while no forces are experienced outside the bearing since the two eccentricities are opposite in sense. In this way we have managed to suppress completely the tendency of the system to vibrate.

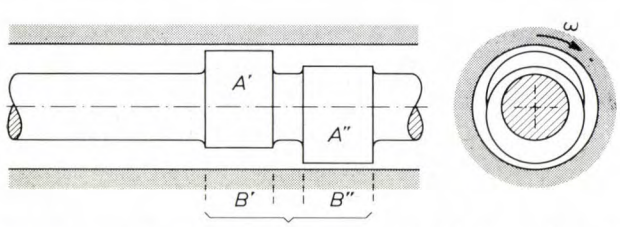


Fig. 2. Each of the journal bearings consists of two sections *B'*, *B''*, in which the corresponding part *A'*, *A''* of the fixed shaft is situated slightly off-centre. The eccentricity (greatly exaggerated in the drawing) is in the opposite direction for each section. This suppresses the tendency of the bearing to vibrate under conditions of weightlessness.

We have already remarked that when the wheel is stationary, the consistency of the grease prevents it from leaking out of the bearing. When the wheel is spinning, the grease behaves more like a fluid, which means that, at least at the end where the shaft is fixed, some lubricant might be lost; at the other end the bearing is completely enclosed (see fig. 1). This loss of lubricant is prevented by a kind of *centrifugal seal*. This is formed by making a step in the fixed end of the shaft, so that there is a cavity *k* between the hub of the rotating wheel and the shaft; see *fig. 3*. Grease in this chamber spins round with the wheel and is thrown away from the shaft by centrifugal action, thus preventing leakage through the gap *l*. The cavity *k* also acts as a grease reservoir for the bearing system: any momentary shortage of lubricant between the actual bearing surfaces is made good from *k*.

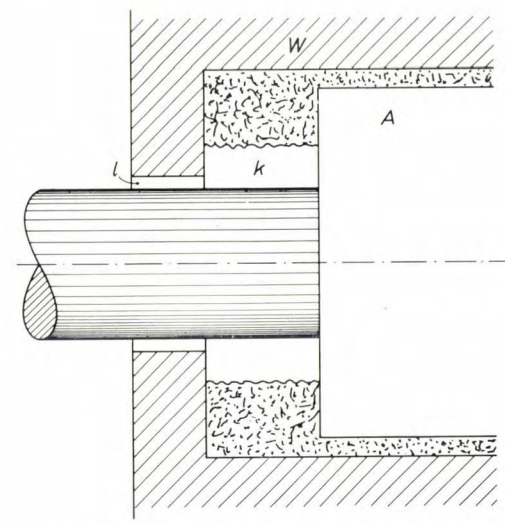


Fig. 3. Centrifugal seal of the cylindrical bearing. The cavity *k* prevents grease leaking from the gap *l* when the wheel *W* is rotating.

Another way in which grease might leak out of the bearings after years of operation might be by evaporation from the free surface of the lubricant in the cavity *k*. This will only happen if condensation can occur at a colder place in the hermetically sealed housing of the flywheel. However, since all heat generated during rotation (in the motor, the bearings and the gas atmosphere) is conducted via the fixed end of the shaft to the body of the space vehicle — in other words since the location of the shaft and the cavity *k* in the heat paths is close to the body which acts as a heat sink — it is possible to ensure that this chamber is also the coldest place in the whole system.

We can now consider in somewhat more detail the design calculations and the minimum condition mentioned earlier. To compensate for the energy losses in

the bearings and in the gas atmosphere a total power P_t is required. This can for example be obtained from solar batteries, possibly combined with a storage battery. The mass of this energy source may be taken as proportional to P_t', i.e. as αP_t, where α is of the order of 0.5 kg/W. If m_h is the mass of the housing and m the mass of the flywheel (see eq. 1), then we have to minimize the sum:

$$m + m_h + \alpha P_t \quad . \quad . \quad . \quad . \quad . \quad . \quad (2)$$

by means of an appropriate choice of parameters, in particular the speed of revolution (ω) and the wheel diameter (R). To determine this minimum we must know the way in which the losses in motor, bearings and gas atmosphere depend upon the parameters. Because there are so many equations relating the different quantities, the actual calculations can only be done with a computer.

Since the frictional losses in the hydrodynamically lubricated bearings are so low, the losses due to gas friction (even with a low pressure in the housing) are relatively important. These losses increase rapidly with the wheel diameter. Direct measurement of these losses with the flywheel itself would be cumbersome and time-consuming. It has been found possible, however, to determine the gas friction losses as a function of diameter and speed of revolution by applying model theory to the results of measurements made with a wheel rotating *in oil* [6]. *Fig. 4* shows the gas friction losses and bearing losses found in this way for the flywheel of fig. 1, with given angular momentum H and effective radius R, as a function of the speed of revolution (the efficiency of the electric motor used [4] varies only slightly with the speed of revolution, so that the motor losses can be accounted for in the constant α). The curves also show the wheel mass found from eq. (1) and the sum (2) based on $\alpha = 0.5$ kg/W. It can be seen that the optimum speed is at 1800 r.p.m. (In actual fact α can be taken somewhat higher, in which case we arrive at the value of 1600 r.p.m. quoted earlier; the optimum value of R found from the computer calculations has been used for fig. 4.)

In the course of the investigations we have made and tested a number of bearing systems. In the tests the flywheel was temporarily replaced by a cylindrical rotor of the same mass (to obtain the same load on the bearings). The rotor was driven by an air turbine. Two such experimental arrangements, with the fixed shaft vertical, can be seen in *fig. 5*. In the arrangement on

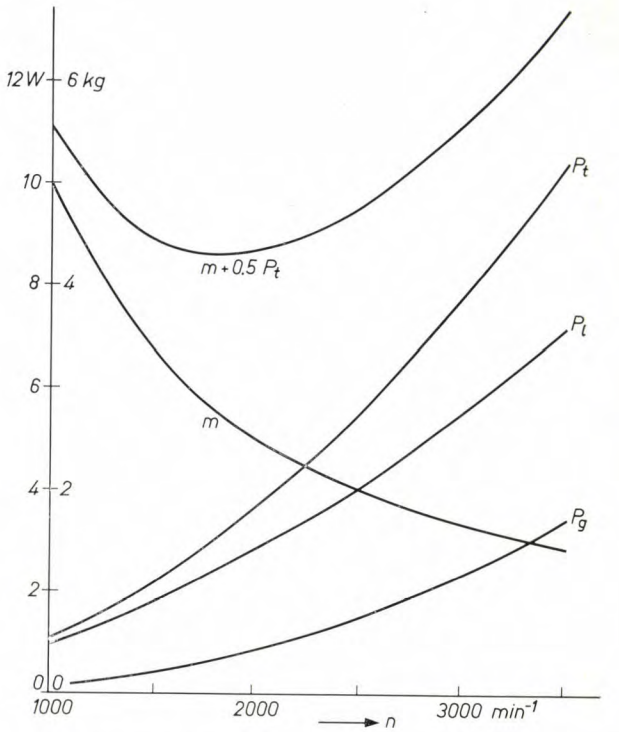


Fig. 4. The curves show the required mass m of the flywheel, the bearing losses P_l, the gas friction losses P_g, their sum P_t, and the sum of the flywheel mass and the mass αP_t of the energy source ($\alpha = \frac{1}{2}$ kg/W) as a function of the speed n for the flywheel described, which has an angular momentum $H = 20$ Nms and an effective radius of about 20 cm. The curve for the sum $m + \alpha P_t$ reaches a minimum at $n = 1800$ r.p.m.

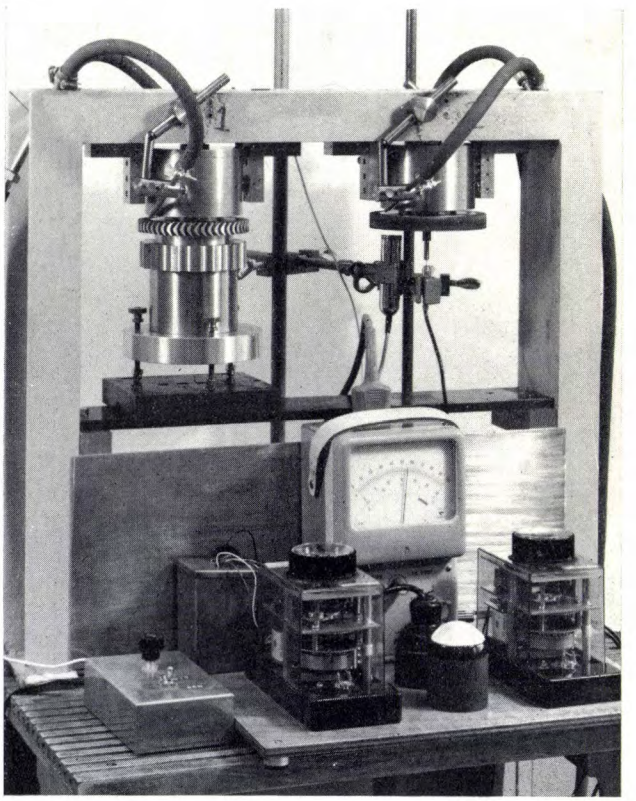


Fig. 5. One of the experimental arrangements with two bearing systems. A state of weightlessness can be simulated with the system on the left.

[5] For a treatment of these vibrations, see B. Sternlicht and N. F. Rieger, Rotor stability, in: Lubrication and Wear (Proc. Instn. Mech. Engrs. **182**, Part 3A), pp. 82-99, 1967/68.

[6] An account of this work, which has been carried out by Ir. J. van der Hilst and H. J. W. M. Volman of these Laboratories, will be published elsewhere.

the left the whole weight of the rotor can be carried by a separate air bearing under the shaft, for simulating the behaviour of our bearing in conditions of weightlessness. The fixed shaft is shown in *fig. 6*. The spiral-

Fig. 6. The fixed shaft with thrust bearing (flange which has a spiral-groove pattern on both sides) and two journal bearings, each consisting of two sections. Helical grooves on the left and right keep the lubricant in the bearings at the correct pressure.

groove patterns of the thrust bearing can be seen on the flange at the end; the two journal bearings, each consisting of two sections (the eccentricity is too slight to be seen with the naked eye), lie between two sets of helical grooves, whose pumping action keeps the lubricant in the bearing at the right pressure.

This bearing, suitable for a flywheel with an angular momentum up to 30 Nms, was found to dissipate 2 to 4.5 watts under no load, depending on the viscosity of the grease used, and 3 to 6 watts under load (i.e. without compensation of the gravitational field). These figures were in good agreement with the calculations.

To check the stability of the rotation, the displacement of the centre line of the rotor was measured in two directions perpendicular to the centre line of the shaft. The electrical resistance between the rotor and the shaft was measured at the same time to check for unwanted metallic contact between the bearing surfaces; this contact, which is present to some extent in the stationary state, was found to be completely broken above about 10% of the normal speed. Many starts and stops were made during the experiments; after stripping down the bearings were found to show no signs of wear at all.

A more advanced flywheel designed on the same principles is at present under construction.

Summary. The resistance of a rotating flywheel to changes in the attitude of its axis of rotation can be put to good use in stabilizing the attitude of a space vehicle, for example to keep the aerials of a communication satellite correctly trained. A grease-lubricated hydrodynamic bearing system is particularly suitable for a flywheel of this kind, which is required to give years of reliable operation. This bearing system has very low friction losses, is free from wear, requires no high starting torque, can be designed so that no subsequent lubrication is required, and can withstand the high accelerations and vibrations to which it is subject when the space vehicle is launched. At Philips Research Laboratories in Eindhoven a vibration-free bearing system has been developed on these lines for a flywheel with an angular momentum of 30 Nms and for an optimum speed, calculated by the computer, of 1600 r.p.m.

A highly efficient small brushless d.c. motor

W. Radziwill

The application of electronic devices has not only led to new techniques, such as radio, television, and the like, which have left their mark on our entire civilization, but their influence is now beginning to change the face of classical electrical engineering as well. Converters, motors and other types of equipment can be made much more versatile with the aid of electronics and often their performance can be considerably improved.

The commutator of a conventional d.c. motor can be regarded as a system of controlled switches which reverse the voltage across the armature coils at certain positions of the rotor. Since the commutator segments have a fixed position in relation to the coils of the rotating armature, and the brushes have a fixed position in relation to the poles of the stator, the switching is always bound to occur at the right moments.

In small, fast-running machines the wear caused by friction and sparking limits the life of the commutator to a few thousand hours under normal conditions, and possibly to a few minutes at very low pressure. A longer life without maintenance becomes possible if electronic switching elements are used instead of the mechanical commutator. To control these switches in the correct phase, however, the motor must be provided with a device which senses the position of the rotor, without using mechanical contacts, and which gives a position signal from which a suitable control signal is then obtained.

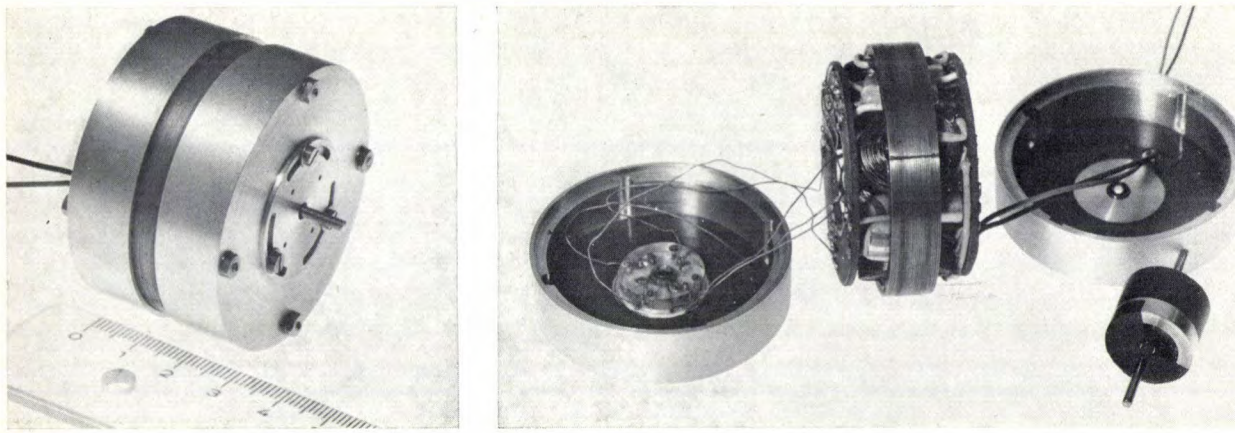
Switching systems of this kind, known as electronic commutators, and based for example on optical or magnetic methods of sensing the position of the rotor, have been known for some time [1]. In all such systems there is a clear distinction between two functions: position sensing with information processing, and the actual power switching by the commutator. This division of function has the fundamental consequence that the power switching can be carried out instead by a position sensor, i.e. depending on the position of the rotor, by a switching device whose operation can be controlled by any variable that one may choose The special problems related to this approach have been studied intensively in the last six years at the Philips laboratory at Aachen. We shall confine ourselves here to electronic commutation that depends on the position of the rotor. This case corresponds with the basic concept of the classical d.c. motor, whose good speed-control capability and high starting torque have given it considerable importance in electrical engineering.

Our investigations concerned with the commutator circuit and the design of the motor were mainly directed towards achieving the highest possible efficiency. We were able to obtain values considerably in excess of the usual 30 to 50% obtained with normal motors (with powers of the order of a few watts). One of the results of our work is the motor described below, which can be designed for a rated power of up to about 20 watts and whose overall efficiency can reach 80%. *Fig. 1* shows an experimental model of such a motor with its commutation system.

The principal applications of these motors are to be found (for the present at least) in air and space travel, where great reliability and an exceptionally long life without maintenance are very important. The advantage of the very high overall efficiency may also be of vital significance in this field, since it permits a considerable saving of weight in the energy source; the weight of the motor with its control system is not appreciably greater than that of its predecessor with brushes. An interesting application in space techniques is for driving a "flywheel" with speed control for stabilizing a space

Dipl.-Ing. W. Radziwill is with the Aachen laboratory of Philips Zentrallaboratorium GmbH.

[1] See for example Philip A. Studer, Development of a brushless d.c. motor for satellite application, NASA Technical Note D-2108, 1964, and W. Yates and R. E. Skamfer, Brushless d.c. torque motor, NASA-CR-59771, 1964.



a *b*

Fig. 1. *a*) Experimental version of the brushless d.c. motor. *b*) The motor before assembly. The stator is at the centre; the plates fixed at each end of it carry all the electronic circuits (figs. 4, 6 and 8). The rotor is at the front on the right.

vehicle, as described in another article in this issue [2]. The rotor can be made integral with the flywheel, thus considerably reducing weight and size.

The model of the motor shown in fig. 1 corresponds to a conventional type of four-phase synchronous motor with two-pole permanent-magnet rotor. This is seen in the schematic of *fig. 2*. It will be noted that the exciting field rotates while the armature with its coils remains stationary, whereas in the classical d.c. motor the situation is exactly opposite. The magnetic material used is an anisotropic ceramic material, ferroxdure 330 k, which has a relatively high remanence and — even more important — a very high coercivity. This makes it insensitive to the strong demagnetizing fields of the armature which occur during starting. The rotor is magnetized along a diameter, giving an approximately sinusoidal distribution of flux density along the circumference. The armature coils *1* and *3* are connected in series, and so are *2* and *4*, and these two pairs are connected to two separate switching units (fig. 2*b*), which form the actual commutator circuit. This will be discussed in more detail below. We shall now first consider the principle of our position sensor, which is illustrated in *fig. 3*.

Fitted in a holder, which is rigidly fixed to the stator, are four small toroidal ferrite cores mounted at angular distances of 90° around the shaft; each core is wound with a coil consisting of a few turns. Fixed to the shaft is a cylindrical sector of ferroxdure 330 k, which is magnetized *axially*. Depending on the position of the rotor, the cores are either exposed or not exposed to the magnetic field of the sector. The *B-H* curve of the ferrite material of the cores has a sharp bend at saturation. In the magnetic field of the cylinder sector the cores

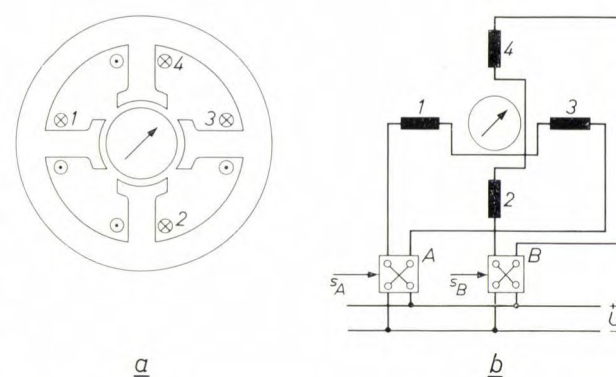


Fig. 2. *a*) Schematic arrangement of the laminated stator and the permanently magnetized rotor, showing the direction in which the armature coils *1* . . . *4* are wound. The arrow indicates the direction in which the rotor is magnetized.
b) Connection of the armature coils to two switching units *A*, *B* (schematic). By means of signals s_A, s_B, coming from the position sensor (or from outside the motor) the switching units are controlled in such a way that they apply the battery voltage U_b to the coils in the required switching sequence.

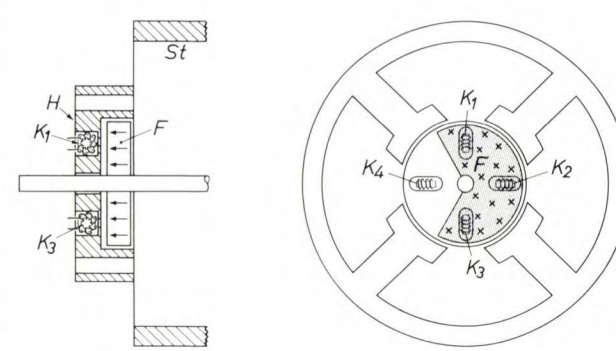


Fig. 3. The position sensor. A plastic holder *H* fixed to the stator *St* carries four wirewound toroidal ferrite cores K_1 . . . K_4. An axially magnetized ferroxdure sector *F* fixed to the motor shaft rotates in front of the cores.

are saturated, making the differential permeability μ_d very small, whereas outside this magnetic field μ_d assumes a very high value. The inductance of the coil around the core therefore varies between a very low and a very high value, depending on the position of the rotor. The transition between the two states is exceptionally sharp.

The state of each core is scanned about 18 000 times a second. The circuit used for this is shown in *fig. 4*. In each period in which the unijunction transistor is non-conducting (and which lasts about 50 microseconds) the four capacitors are charged up through the relatively large resistance R (*fig. 5a*). The inductance of the coils has virtually no effect on this. Once the peak voltage of the transistor has been reached, its emitter diode becomes conducting for about 5 μs, so that during this short time the full capacitor voltage appears across the coil, unless the core is saturated. In that case the coil behaves like a short circuit of very low resistance, and the voltage therefore remains zero. As long as the core is not saturated, the 5 μs voltage peak (with an oscillatory transient which is not significant here) appears across the coil after every non-conducting period of the transistor, i.e. 18 000 times per second. As the rotor rotates pulses of high-frequency signal are thus obtained at the different coils (fig. 5*b*), the relative length of the pulses with respect to one revolution of the rotor depending on the angle occupied by the magnet sector fixed to it. Demodulation of the signals gives four control signals consisting of a series of steep-sided rectangular pulses, and spaced by a quarter of a period. The instantaneous values of these signals depend solely on the instantaneous position of the rotor.

As compared with other known position sensors, our system offers several advantages. No great accuracy of construction is required; the air gap is not critical and can be as large as 1 mm. The pulse edges of the low-frequency control signal are exceptionally steep, so that no further pulse-shaping stages are needed. The design is simple, insensitive to dirt or mechanical shock, requires no maintenance at all and has a power consumption of only about 40 mW.

We shall now turn to the commutator circuit proper, which transfers the power to the armature coils and is represented symbolically in fig. 2*b* by the two switching units *A* and *B*.

Each of these units is made up from two *NPN* and two *PNP* transistors together with a pair of diodes connected in parallel opposition, as shown in *fig. 6*. The

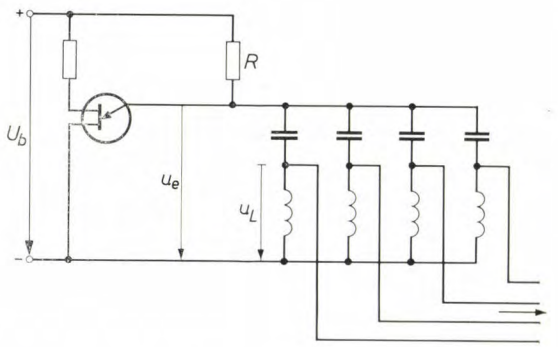
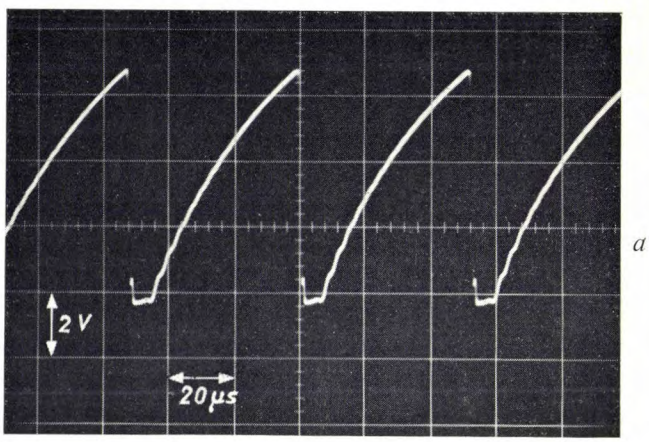
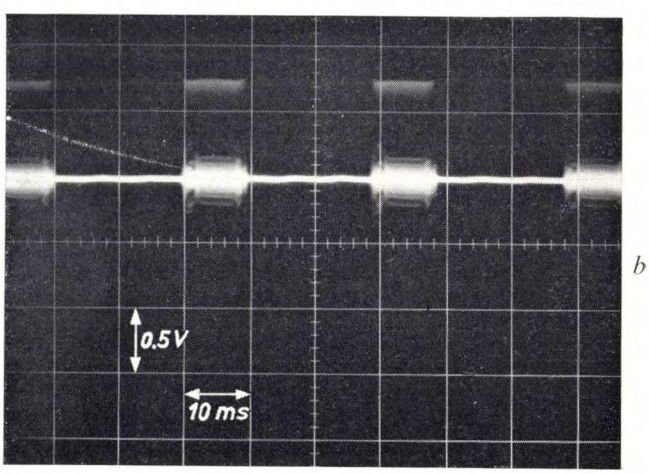


Fig. 4. Oscillator circuit for scanning the magnetic state of the cores.



a



b

Fig. 5. *a*) Oscillogram of the high-frequency variation of the emitter voltage u_e in fig. 4.
b) Oscillogram of the voltage u_L across one ferrite-core coil when the motor is running. As can be seen, the core is saturated during two thirds of every revolution ($u_L = 0$).

[2] J. P. Reinhoudt, A flywheel for stabilizing space vehicles, Philips tech. Rev. **30**, 2-6, 1969 (No. 1).

switching sequence we have adopted and which has been found to be the most suitable for both theoretical and practical reasons, is given in *Table I*. The total angle 2π is divided into eight parts δ_i, not all of the same size; the position sensor has to tell these parts from each other and make the eight transistors T_1T_1'... T_4T_4' conducting (0) or non-conducting (1) as shown for each angular division in Table I. The effect of this is a kind of stepped rotation of the armature energization,

Table I.

	T_1	T_1'	T_2	T_2'	T_3	T_3'	T_4	T_4'
δ_1	0	1	1	0	1	0	0	1
δ_2	0	1	1	1	1	0	1	1
δ_3	0	1	0	1	1	0	1	0
δ_4	1	1	0	1	1	1	1	0
δ_5	1	0	0	1	0	1	1	0
δ_6	1	0	1	1	0	1	1	1
δ_7	1	0	1	0	0	1	0	1
δ_8	1	1	1	0	1	1	0	1

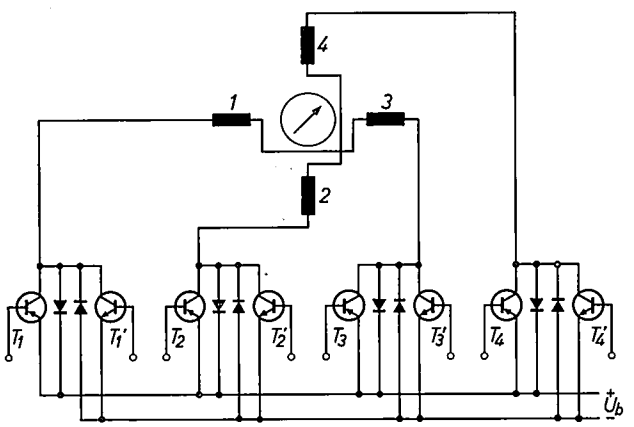


Fig. 6. Balanced arrangement of the switching units shown as *A* and *B* in fig. 2*b*. They include *PNP* and *NPN* transistor stages $T_1T_1' \ldots T_4T_4'$, and diodes connected in parallel opposition.

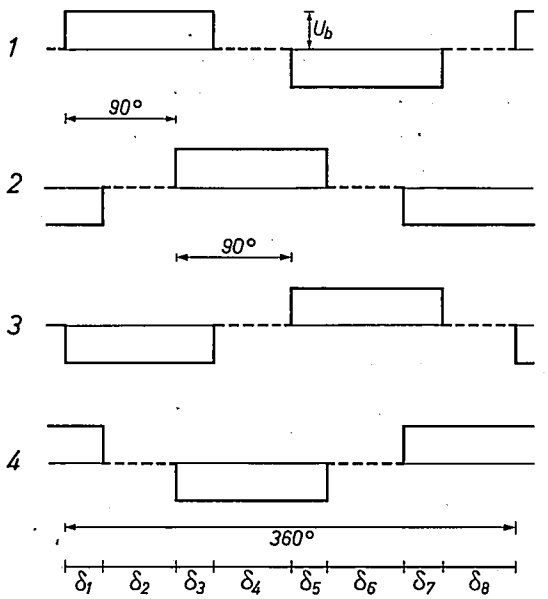


Fig. 7. The voltage waveform applied to the armature coils *1* . . . *4* by the switching units. Each time after switching off and before switching on again with the opposite polarity, the applied voltage is not defined. The windings are connected (fig. 2*a*) in such a way that the voltages across *1* and *3* are 180° out of phase with each other, as are also the voltages across *2* and *4*.

which by interaction with the rotating field of the rotor produces a torque that is always positive and fluctuates only slightly with the position of the rotor. The angular divisions δ_i are assigned the values $\delta_1 = \delta_3 = \delta_5 = \delta_7 = \pi/6$ and $\delta_2 = \delta_4 = \delta_6 = \delta_8 = \pi/3$. This corresponds to the design of the position sensor indicated in fig. 3, with a magnet sector that covers 240°. From the columns of Table I we can now see how the applied armature voltage varies with time. Thus for example coils *1* and *3* are connected via T_1 and T_3' to the battery voltage U_b in one polarity during 120° of rotation of the rotor, are then switched off for the next 60° and are then connected via T_1' and T_3 to the voltage U_b in the opposite polarity for another 120°, and so on. *Fig. 7* shows the waveform of the voltage across the coils *1* . . . *4*.

We shall see below that this waveform is particularly suitable. The switching pattern adopted to give this waveform and shown in Table I also has the special advantage that the transistor pairs T_1 and T_3', T_2 and T_4', etc. are always switched on or off simultaneously: each pair can therefore be controlled by a single driving stage (*fig. 8*). In the circuit shown here, which has been kept deliberately simple, the high-frequency signal is demodulated by the emitter-base diode of the appropriate

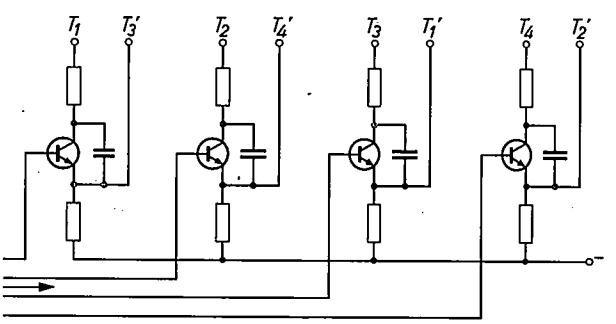


Fig. 8. Circuit of the combined demodulator and control stages. The pattern adopted for switching the coils allows each pair of power transistors $T_1T_1' \ldots T_4T_4'$ (see fig. 6) to be controlled by a common driving stage.

driving transistor. For smoothing the low-frequency control current an external capacitance is used between emitter and collector, which at the same time can make the edges of the driving signal less steep if necessary to avoid radio interference from the power-switching transistors. Further filtering of the residual high-frequency components is not necessary since the armature coil itself makes a very good low-pass filter.

The waveform in fig. 7, which is applied by the electronic commutator to the armature coils, is particularly suitable because the current contains no harmonics of order divisible by 3. The following shows that this is very important if high efficiency is desired.

Of the current i and the induced voltage e in each of the m armature coils ($m = 4$), only the harmonics with *the same frequency* can contribute to the power delivered by the motor. This power, at a speed of $n = \omega/2\pi$, is:

$$P_0 = \frac{m}{2\pi} \int_0^{2\pi} ei \, \mathrm{d}(\omega t), \quad \ldots \quad (1)$$

and the above statement is at once evident if the integrand is expressed as a series. Since the flux density is distributed sinusoidally along the circumference of the rotor, the induced voltage at a constant angular velocity ω is given by:

$$e = \omega N \Phi \sin(\omega t + \psi). \quad \ldots \quad (2)$$

Here N is the number of turns of one coil, Φ the maximum enclosed flux from the rotor, and ψ the phase lead of e with respect to the instant of switching-on of the voltage applied to the coil by the commutator circuit. Since the induced voltage contains only the fundamental, all higher harmonics in the current are attributable only to the applied voltage — provided there is no saturation in the iron of the stator — and, as equation (1) shows, can therefore only make a contribution to the losses. If, instead of the waveform shown in fig. 7, a square wave were applied (with energization during 180° in each half-period), then the contributions from the third, ninth, and fifteenth harmonics, etc. alone would account for about 50% of these losses.

We should add for completeness that the current waveform that must be known for equation (1) cannot yet be derived when only the waveform of the applied voltage (fig. 7) and the opposing induced voltage (equation 2) are known. During the periodic switching-off of the voltage across them the coils remain connected to the battery voltage for a short time via the parallel-connected diodes. The polarity depends on the direction of the current at the moment of switching-off, and the time that the connection lasts is determined by the magnitude of the current at that moment (and by other factors). Depending on the loading of the motor (full load, no load, braking, starting) very different situations are thus found, and the behaviour of the system cannot be described with simple equations. The theoretical torque M and the power consumption P_1 have therefore been calculated with an electronic computer as a function of speed, for various values of the characteristic parameters. Some of the resultant characteristics are shown in *fig. 9*. *Fig. 10* gives measured values for M and the efficiency η. For comparison the corresponding theoretical curves are shown; unlike the curves in fig. 9,

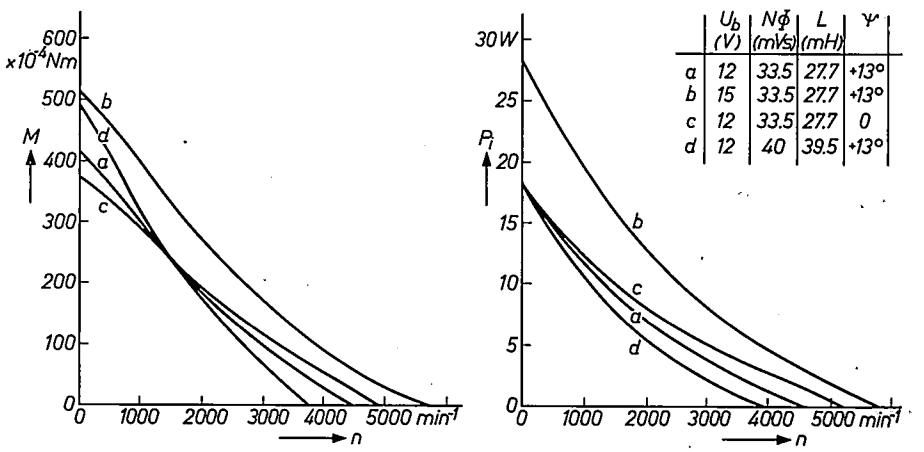


Fig. 9. Theoretical curves of the torque M and the power consumption P_1 (not considering the electronic components) as a function of the speed n for various values of some of the characteristic parameters. L is the inductance of the series arrangement of coils *1* and *3* or *2* and *4*. The resistance R of this series arrangement was 10.5 Ω in all cases.

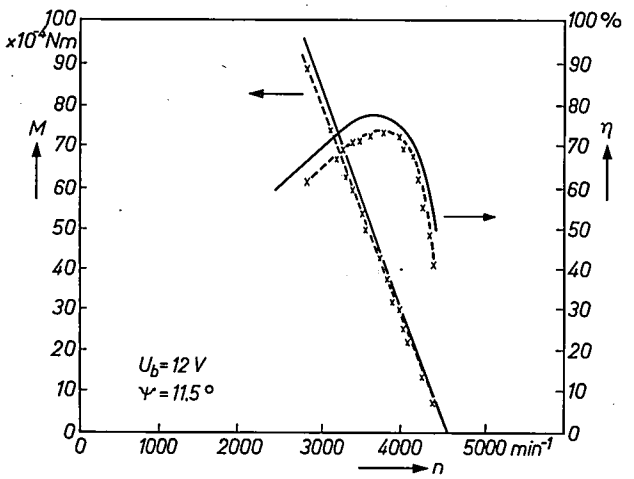


Fig. 10. The points joined by dashed lines give measured values of the torque M and the efficiency η, the solid curves give the corresponding theoretical values. Efficiencies up to 75% were achieved in this case.

these curves include the iron losses and the power consumption of the position sensor. Only the friction losses, which are not readily reproducible, have not been taken into account. It can be seen that in this particular case an efficiency of 75% was obtained at the optimum speed.

Summary. The function of the conventional commutator in a d.c. motor can be performed instead by an electronic device, giving a motor that will operate for a very long time without maintenance. A motor is described that can be built for a rated power of up to about 20 W and in which the commutation is controlled by a magnetic position sensor that takes very little power. The state of the sensor is scanned at a high frequency. The armature coils are energized by means of power transistors, which are controlled by the position sensor in a switching pattern that gives virtually no harmonics of order divisible by 3 in the armature current. As a result of this and other features of the design the overall efficiency of such a motor can reach 80%.

α-particle spectrometry with semiconductor detectors

W. K. Hofker, K. Nienhuis and J. C. Post

The low ionization energy of silicon and germanium makes semiconductor diodes of these materials very suitable as detectors of the ionizing radiation emitted by radioactive sources. Semiconductor diodes are therefore in general use for detecting the various forms of this radiation. Some time ago the operation and distinctive features of semiconductor detectors were dealt with in this journal. The article below deals in greater detail with their use for the detection and energy measurement of particle radiation: α-particle spectrometry is a useful example for demonstrating the practical requirements to be met by semiconductor detectors.

Semiconductor detectors, with their various features that make them specially suitable for the detection of ionizing radiation, have largely superseded the older types of detector like the ionization chamber and the scintillation counter. This was discussed at some length in a previous article in this journal [1]. The applications of semiconductor detectors today include the detection and energy measurement of charged particles (α- and β-radiation, protons, deuterons, tritons and heavier particles) as well as of electromagnetic radiation (γ-radiation). As an example of an application for charged-particle studies we shall describe here the use of a semiconductor detector for measuring the energy spectrum of α-radiation, showing in particular how the detector is adapted to meet the requirements of this application. At the end of this article a complete α-particle spectrometer will be discussed.

Sources of α-radiation are radioactive nuclides of high atomic number. Some of these nuclides occur naturally, but a greater number are obtained artificially by the irradiation of atomic nuclei. Since the energy of the emitted α-particles is characteristic of the emitting nuclide, it is possible to find out which nuclides are present in the radiation source by measuring the energy of the α-particles. The importance of this in fundamental nuclear research need hardly be stressed; in addition, however, there are more immediately practical applications such as the investigation of fuel elements for nuclear reactors and, in laboratories where radioactive substances are used, the regular examination of persons and workrooms for contamination by α-active substances, which are a serious health hazard. In the latter case an analysis of the α-radiation present gives an indication of the origin of the contamination. This also applies to the monitoring of air, surface water and soil for α-contamination which, since it concerns public health, is a matter for the public authorities.

A technical application of α-radiation is in the measurement of foil thickness, the energy difference between the incident and the emergent radiation being a measure of this thickness. Another application is in the investigation of a material surface. When α-particles strike such a surface, some of them are scattered backwards with an energy that depends upon the mass number of the atoms present in the surface layer. The composition of the surface can thus be deduced from the energy spectrum of the back-scatter. This method was used, for example, in the spacecraft Surveyor V for investigating the surface of the moon.

Principle of α-particle spectrometry

Basically, a semiconductor detector is a semiconductor diode with a reverse bias across it. The ionization caused by a nuclear particle penetrating into the depletion layer frees pairs of charge carriers (electrons

Ir. W. K. Hofker, of Philips Research Laboratories, is with the Institute for Nuclear Physics Research (IKO) in Amsterdam; Dr. K. Nienhuis is with the Philips Electronic Components and Materials Division (ELCOMA), Eindhoven; Drs. J. C. Post is with the Institute for Nuclear Physics Research (IKO) in Amsterdam. *The contribution of Drs. J. C. Post to this article is based on work undertaken as a part of the programme of the Foundation for Fundamental Research on Matter (F.O.M.), which is supported by the Netherlands Organization for the Advancement of Pure Research (Z.W.O.)*

[1] W. K. Hofker, Semiconductor detectors for ionizing radiation, Philips tech. Rev. 27, 323-336, 1966.

and holes), which travel towards the two electrodes under the influence of the applied field, thus producing a current pulse. If the whole of the energy of the incident particle (or a constant fraction of it) is used for the freeing of charge carriers, then the number of the charge carriers and hence the height of the current pulse will be proportional to the energy of the particle.

The pulses are amplified and fed to a pulse-height analyser (kick-sorter). This has several hundreds or thousands of channels, into which the incoming pulses are sorted by height. The pulses in each channel are counted, giving the number of particles with the energy value corresponding to each channel. *Fig. 1* gives an example of an α-spectrum obtained with a 400-channel analyser. The figure shows that it is possible to discriminate between α-particles of different energy only if the peaks are sufficiently narrow. The energy resolution of the measuring arrangement is expressed by the *line width*, defined as the half-height width of a peak. The figure gives an example showing that for the highest peak of ^{241}Am the line width is 18.5 keV.

Considering that there are some 1000 different α-spectrum lines known [2], originating from about 40 natural and about 240 artificial α-nuclides, and that more than 90% of these are between 4.5 and 8 MeV, it is evident that a small line width is of great importance in α-particle spectrometry. In many cases a line width of 20 keV proves to be sufficient in α-particle spectrometry and can be achieved with semiconductor detectors. A prerequisite, however — we shall return to this on p. 19 — is that the radioactive source must be very thin (in practice the source is a layer, about 10 nm thick, on a substrate whose surface roughness must be less than 1 μm).

A second important requirement to be met in α-particle spectrometry is a low background level, so that weak lines can also be detected. This spectrum background has many and various causes, which will be dealt with shortly in some detail.

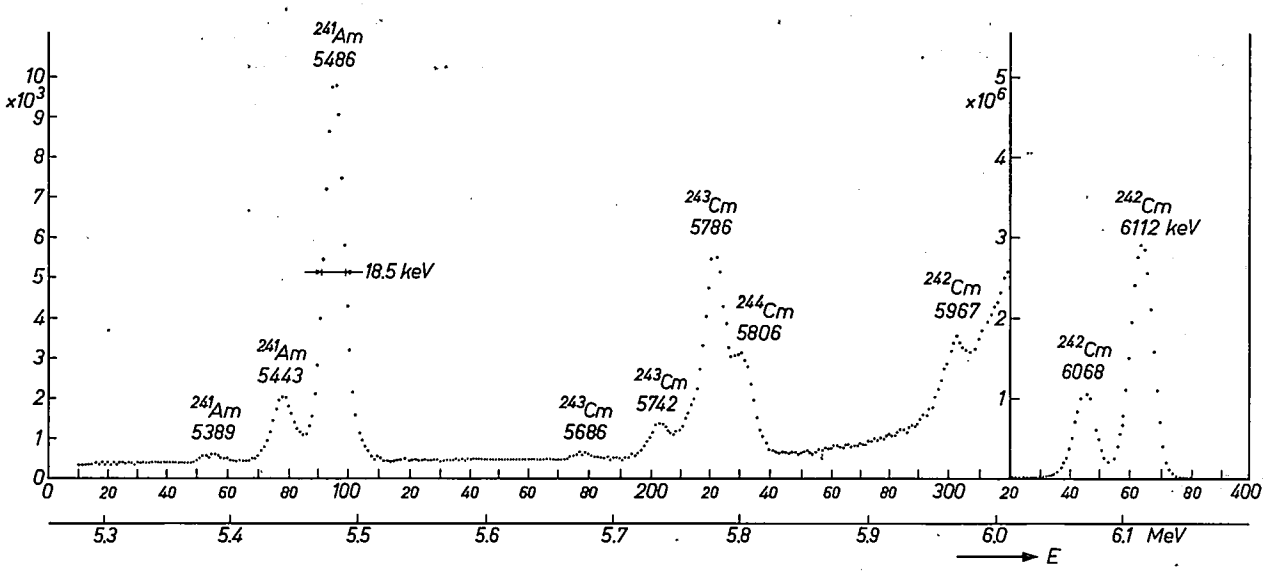


Fig. 1. Record of α-spectrum of a ^{242}Cm specimen, obtained with a 400-channel pulse-height analyser. The number of particles per channel is plotted vertically. In the right-hand part of the spectrum the vertical scale is compressed 500 times to display the main peak of ^{242}Cm. The ^{242}Cm was obtained by irradiating ^{241}Am with neutrons and then chemically separating the curium and americium. The spectrum shows that there are still traces of Am present in the Cm, and small amounts of ^{243}Cm and ^{244}Cm have also been formed due to multiple neutron trapping.

Both in resolution and spectrum background the semiconductor detector is clearly superior to the ionization chamber formerly used for α-particle detection. The ionization chamber, however, has the advantage of a larger sensitive volume. If it is desired to measure the radiation from sources with a very long half-life ($>10^{10}$ years), and thus a very low intensity, it is necessary to have a source with a large surface area to obtain a sufficient number of pulses per unit time. In such cases the small dimensions of the solid-state detector are a disadvantage, and the ionization chamber can still be used.

The surface-barrier detector

The *P-N* junction required in the semiconductor detector can be obtained by the familiar method of diffusing *N*-type impurities into one side of a wafer of *P*-type

silicon. This gives a surface layer of N-type silicon. The α-particles penetrate through this layer (which need be only a few tenths of a micron thick) and reach the depletion layer below it, produced by the application of a reverse bias. A disadvantage of this detector, however, is that the preparation by diffusion can only take place at a relatively high temperature (about 950 °C). At this high temperature, recombination centres are formed in the monocrystalline silicon and these trap some of the charge carriers released by the α-particles. Owing to its statistical character this process has a detrimental effect on the energy resolution of the detector.

Detectors for α-particles are therefore preferably made by a process that takes place at low temperatures. Now it has been found that after N-type silicon has been given a certain oxidizing treatment, a depletion layer is formed right at the surface, particularly if a gold electrode has been evaporated on to the surface. During this evaporation the temperature of the substrate does not rise above about 50 °C. A surface-barrier detector produced in this way possesses the optimum characteristics for α-particle detection [3].

The surface barrier

We can only give an approximate picture of the way in which the surface barrier forms, since some aspects of the process still remain to be explained. Even without the presence of gold, a depletion layer can form at a silicon surface. The crystal lattice of silicon is abruptly interrupted at the surface. Because of this, and also because foreign atoms can easily adhere to the surface, acceptors and donors at many different energy levels are created. Which atoms are involved depends on the atmosphere to which the silicon surface is exposed during its formation; the control of this atmosphere is therefore of the utmost importance in the preparation of a good surface barrier. To obtain a depletion layer in N-type silicon, the distribution of acceptors and donors over the energy levels must be such that, if all levels below the Fermi level E_F are occupied by electrons, a negative surface charge results. This charge drives the conduction-band electrons out of the crystalline layer immediately below the surface, so that a depletion layer is formed there. In this layer there is a positive space charge of ionized donors, which, together with the negative surface charge, makes the silicon outwardly electrically neutral. If we draw the energy-band diagram of the surface barrier (*fig. 2a*), we see that, as a result of the space charge and the local electric field which it causes, the energy bands near the crystal surface are bent. Here b is the depth of the depletion layer and V_d the potential difference across the barrier.

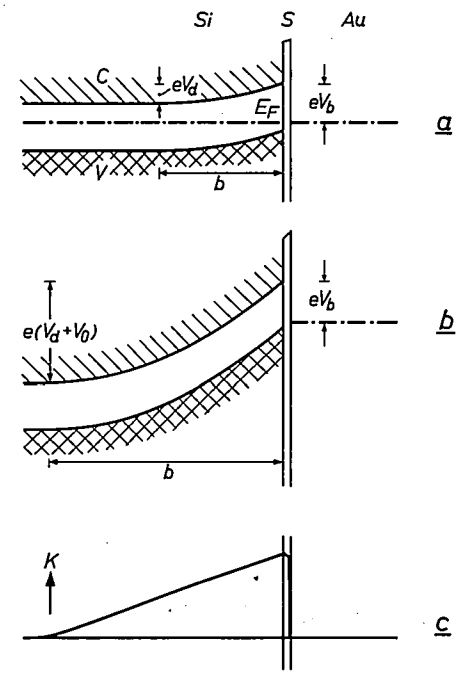


Fig. 2. Energy diagram of the barrier at the surface of N-type silicon. C conduction band. V valence band. *Si* silicon. S surface film. *Au* gold electrode.
a) The location of the Fermi level E_F is such that a negative charge is formed at the silicon surface. This creates a depletion layer of depth b, immediately under the surface. In the depletion layer a positive space charge is formed which neutralizes the negative surface charge. The existing electrical field leads to a potential difference V_d, the diffusion potential, and is the cause of the indicated band bending. eV_b is the height of the barrier.
b) The application of a reverse bias voltage V_0 increases the depletion depth b.
c) The strength of the electrical field K as a function of position in the depletion layer.

If we now provide the crystal with an electrical contact by depositing a gold film *Au* on to the surface layer S (this layer is so thin — 5 to 12 nm — that charge transport through it can take place by means of the tunnel effect) the Fermi levels in the metal and the semiconductor assume the same value. In this way the gold film affects the band bending. It has been found experimentally that a deposited layer of gold supports the formation of a surface barrier, whereas aluminium, say, almost completely destroys it and makes the contact with the substrate more or less ohmic.

[2] Many of these are included in: J. C. Post, Table of alpha-emitting nuclides according to alpha energy, Actinides Rev. **1**, 55-69, 1967.

[3] Another method which can be used for producing a P-N junction at lower temperatures than those required for diffusion is ion implantation. In this method accelerated donor ions are injected into a P-type semiconductor; these form an N-type surface layer that can be very thin. Subsequent thermal treatment is necessary.

The atoms of the surface layer S have to be bound so firmly to the silicon that this layer, which is essential to the operation of the surface barrier, undergoes no change under vacuum. (It is necessary to measure α-particles under vacuum in order to avoid energy losses caused by collisions with gas molecules.) It is now possible to make surface barriers that are completely stable in a vacuum as low as 10^{-7} torr.

The thickness of the depletion layer

When an external voltage V_0 is applied to the detector in the reverse direction, the band bending becomes more pronounced and b, the thickness of the depletion layer, increases (fig. 2*b*). This thickness must be sufficient for an incident α-particle to lose all its energy inside the depletion layer. *Fig. 3* shows the range of α-particles in silicon as a function of their energy [4]. It can be seen that a depletion depth of 100 μm is amply sufficient for the detection of α-radiation up to 11 MeV, that is to say of all known α-spectrum lines. The way

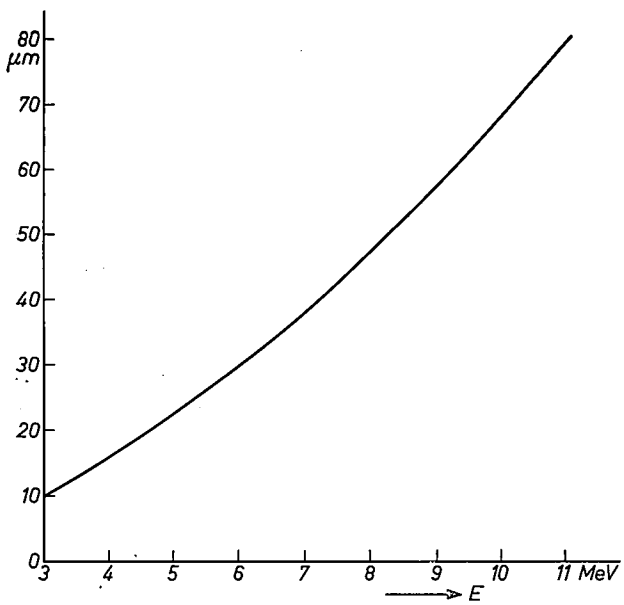


Fig. 3. The range of α-particles in silicon as a function of their energy E.

in which the depletion depth b (in microns) depends on the applied voltage V_0 (in volts) is given by the relation:

$$b = 5.3\sqrt{\varrho(V_d + V_0)},$$

where ϱ is the resistivity of the silicon (in Ωm). This relation is given graphically in *fig. 4* for a number of values of ϱ. We deduce from this that an applied voltage of 10 V is needed for a depletion depth of 100 μm in material with a resistivity of 50 Ωm.

When thicker depletion layers are required, as for example in the detection of β-particles, which have a range of a few millimetres, a limit is imposed by the breakdown strength of silicon.

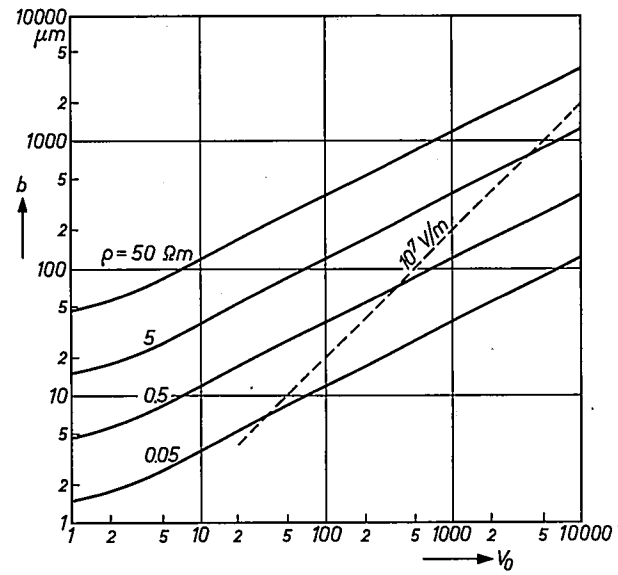


Fig. 4. The depth b of the depletion layer as a function of the applied voltage V_0 at different values of the resistivity of N-type silicon. The points where the breakdown field strength of silicon is reached (10^7 V/m) are indicated by a dashed line.

The dashed line in fig. 4 indicates the operating points where the breakdown field strength of silicon is reached, at about 10^7 V/m. The figure takes into account the fact, which can be seen from fig. 2*c*, that the maximum field strength occurs at the surface layer S and is twice the average field strength in the depletion layer. To keep the field strength below the breakdown value it is necessary to use silicon of very high resistivity. Even then, high voltages are required (above 1000 V). By choosing a very homogeneous basic material and taking care in manufacture it is now possible, using silicon with a resistivity of 100 to 200 Ωm, to produce detectors with a depletion depth of 2.5 mm at an applied voltage of more than 2000 V. These detectors are made in a form in which the depletion layer extends up to the rear electrode, enabling them to be stacked for the measurement of particles with a range greater than 2.5 mm.

We ought to note in passing that very thin surface-barrier detectors with a depletion layer extending up to the rear electrode are being made for use in particle identification systems (see the treatment of the dE/dx detector on page 334 in reference [11]). These can now be made with a thickness of only 4 μm, which means that an α-particle loses only a small part of its total energy in it. This is illustrated in *fig. 5*, where the peak on the right represents the total energy of α-particles emitted by ^{241}Am, and the peak on the left the energy which these α-particles lose in a 4 μm thick detector.

The reverse current

Even when there is no ionizing radiation present, a small reverse current will still flow in the detector. This "dark current" is the sum of a number of components, two of which in particular cause electrical noise and therefore have a detrimental effect on the energy resolution. The first of these two components is due to those electrons in the gold electrode that possess sufficient energy to cross the potential barrier and reach the conduction band. The potential barrier is indicated by

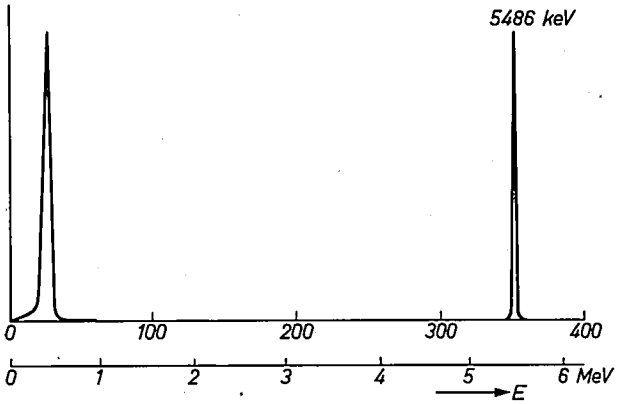


Fig. 5. Energy absorption for the penetration of a surface-barrier detector with a total thickness of 4 μm by α-particles (emitted by ^{241}Am). The left-hand peak in the spectrum is due to pulses given by the detector when α-particles pass through it, and thus corresponds to the energy dissipated in the detector. The right-hand peak is due to a calibration pulse and indicates the total energy of these α-particles.

the potential difference V_b in fig. 2*a* and *b*. The second component is due to the thermal generation of free charge carriers inside the depletion layer. Both components increase with rising temperature. Since both of them give rise to shot noise in the detector signal, they must be kept as small as possible. This calls for a high potential barrier and a minimum of crystal defects and other charge-generation centres in the silicon basic material. Moreover, the temperature of the detector should preferably not rise above room temperature.

The reverse-current characteristic of a surface-barrier detector is given in *fig. 6*. In addition to the components mentioned, the total reverse current includes the leakage current along the surface of the detector. Unlike the first two components, this is not so important for the energy resolution, since the frequency spectrum of the noise it causes enables us to filter most of it out.

Besides the electrical noise referred to, there are also

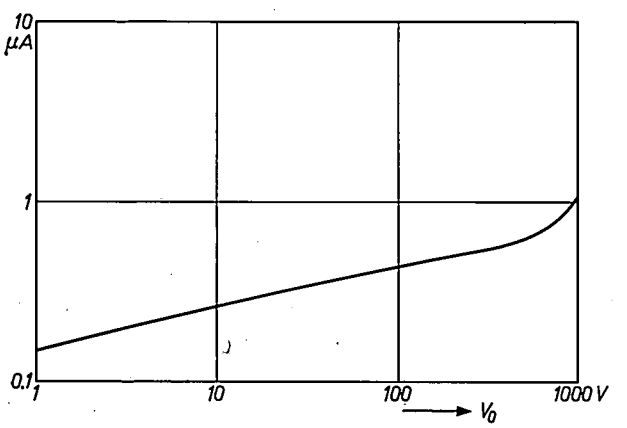


Fig. 6. Reverse-current characteristic of a surface-barrier detector with a surface area of 2 cm².

statistical fluctuations in the actual ionizing process which adversely affect the resolution. The contributions of several processes to the line width can readily be taken into account in an approximate calculation, which is given below.

Calculation of energy resolution

A primary cause of peak broadening is the energy absorption in the window through which the α-particles reach the depletion layer. In the first place, there is a spread in the energy absorption because the particles pass through the window at different angles. This can be minimized by placing the source, whose dimensions are assumed to be approximately equal to those of the counter (e.g. 1 cm²), at a certain distance (2 to 3 cm) from the counter. After this precaution has been taken, the fact remains that there is a statistical spread in the energy absorption in the window, even though the particles pass through the window at normal incidence.

For a gold window we can deduce that for α-particles the contribution B_g to the line width (in keV) is:

$$B_g = 5.09 \times 10^4 \sqrt{d},$$

where d is the thickness of the window in metres. For $d = 20$ nm, for example, $B_g = 7.2$ keV.

This expression was obtained by using the Bohr equation for the variance σ^2 of the energy absorption per unit path length [5]:

$$\sigma^2 = 4\pi z^2 e^4 NZ,$$

where z is the charge of the particle in terms of the elementary charge, e the charge of the electron, N the number of atoms per unit volume of the absorption layer, and Z the atomic number of the atoms in the absorption layer. This equation can be used to give the accuracy required here because the average number of collisions of an α-particle in the window is sufficiently large.

Another cause of energy dispersion is to be found in the ionization process. In the first instance we have a statistical spread in the number of hole-electron pairs formed by ionization. This is because the energy of the incoming particle is only partly used for ionization processes. For the standard deviation in the number of hole-electron pairs N we can write:

$$\Delta N = \sqrt{FN_0},$$

where N_0 is the average number of hole-electron pairs and F is the Fano factor [6]. The contribution B_f (again in keV) which this fluctuation makes to the line width is:

$$B_f = 2.36\, w\sqrt{FN_0} = 2.36\sqrt{FEw},$$

[4] This figure is based on calculations by D. J. Skyrme, Nucl. Instr. Meth. **57**, 61, 1967.

[5] R. D. Evans, The atomic nucleus, McGraw-Hill, New York 1967, p. 661.

[6] U. Fano, Phys. Rev. **72**, 26, 1947.

where w is the average ionization energy of silicon, expressed in eV, and E is the energy of the α-particle in MeV. For an α-particle with $E = 6$ MeV and for $w = 3.6$ eV, assuming $F \doteq 0.15$ [7], we find $B_f = 4.25$ keV.

It has been found experimentally that the fluctuations of the accumulated charge in the detection of heavy particles are substantially greater than may be deduced from the foregoing. A possible cause is the occurrence of recombinations of electrons and holes in the ionization track as a result of the high charge density which exists there. Moreover, additional fluctuations are caused by heavy particles transferring part of their energy to atomic nuclei that have a poor ionization yield. The probability of ionization already begins to decrease steeply when the energy of the atomic nuclei is still fairly high. Although the total residual energy lost by all the atomic nuclei struck by an α-particle is fairly small, being about 12 keV for an α-particle of 6 MeV, its fluctuations are considerable. The contribution B_l which these fluctuations make to the line width may be calculated approximately [8] from:

$$B_l = 0.47\, z^{1/2} A^{4/3},$$

where z is the charge of the ionizing particle in terms of the elementary charge and A the mass number. For α-particles B_l is about 4.5 keV.

Another significant cause of line broadening is the electrical noise of the counter and the amplifier; this noise produces fluctuations in the measured pulse height. To minimize the noise of the preamplifier a junction-type field-effect transistor is often used in the input stage. Further improvement in the signal-to-noise ratio can also be obtained by including in the amplifier a bandpass filter with a response such that the pulses are transmitted without significant attenuation but the noise is attenuated. A simple RC filter, as shown in *fig. 7*, is frequently used for this purpose. It can be shown that, to obtain an optimum signal-to-noise ratio, the time constants of the two sections should be made identical [9]. If we assume that the only sources of noise are the shot-noise component I_d in the detector current and the equivalent noise resistance R_{eq} of the input stage, we can show that the noise contribution to the line width is a minimum at a particular value of the filter time constant τ; this minimum contribution B_n (in keV) follows from the expression:

$$B_n = 10^{-3}\, e\, \frac{w}{e} \left(C^2 R_{eq} I_d \frac{kTe}{2} \right)^{1/4}.$$

Here e is the base of the natural logarithms, e the charge of the electron, C the capacitance of the amplifier input (mainly the detector capacitance), and T the absolute temperature (all in SI units).

The line width increases with increasing capacitance C since an α-particle frees a given quantity of charge. Over a larger capacitance this charge means a smaller voltage, giving a lower signal-to-noise ratio.

When the applied voltage V_0 is raised the depletion depth increases and the capacitance C decreases, resulting in a lower value of B_n. If V_0 is raised too far, however, the reverse current I_d increases rapidly (fig. 6) and so too does B_n. There is therefore an optimum value of V_0 and of the related depletion depth giving minimum line width. The location of this optimum is also affected by the filter time constant τ, because the frequency composition of the noise is not the same in all cases. These effects are illustrated in *fig. 8*, where the line width measured for a detector is shown in a three-dimensional diagram as a function of depletion depth and filter time constant. It can be seen from this diagram that the smallest line width is measured at a specific value of τ, which is the optimum value mentioned earlier.

For an approximate calculation of B_n we assume a capacitance C of 100 pF (this is the capacitance of a detector with an area of 1 cm^2 and a depletion depth of 100 μm), an equivalent noise resistance $R_{eq} = 200\ \Omega$ and a reverse current $I_d = 0.1\ \mu$A. In this case, at the then applicable optimum time constant $\tau \approx 1\ \mu$s, we have $B_n = 5.5$ keV. The total line width, due to the various causes mentioned, is:

$$B_{tot} = \sqrt{B_g^2 + B_f^2 + B_l^2 + B_n^2}.$$

Substituting in this the values from the examples, we obtain:

$$B_{tot} = 11 \text{ keV}.$$

This line width corresponds to the value actually found in optimum cases. As a rule it is not as good as this, for various reasons. One of these seems to be the inhomogeneous distribution of recombination and trapping centres in the depletion volume, another is the occurrence of very small breakdown effects ("microplasmas") when a high reverse bias is applied to the detector. These microplasmas, which are prone to occur at the edge of the detector, are a source of electrical noise whose energy content per Hz bandwidth is inversely proportional to the frequency f ($1/f$ noise). Although the greater part of this noise falls outside the passband of the filter, the contribution of the micro-

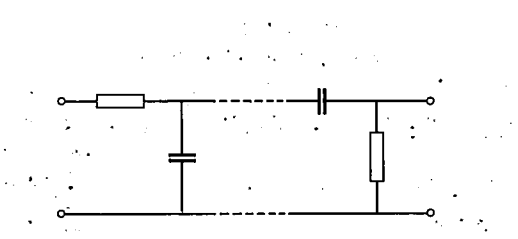


Fig. 7. Schematic representation of an RC bandpass filter as used in the electronic measuring circuit. The filter consists of an integrating section and a differentiating section; for optimum operation the time constants of both sections should be identical.

plasmas is not negligible at high operating voltages.

Finally, it is found in practice that the quality of the radiation source also has some effect on the line width. This comes about because the α-particles can lose energy by collisions inside the source itself. Such collisions cause a spread in the energy of the α-particles which arrive at the detector, and a corresponding increase in the line width. This is why the source has to be as thin as possible, as we noted earlier.

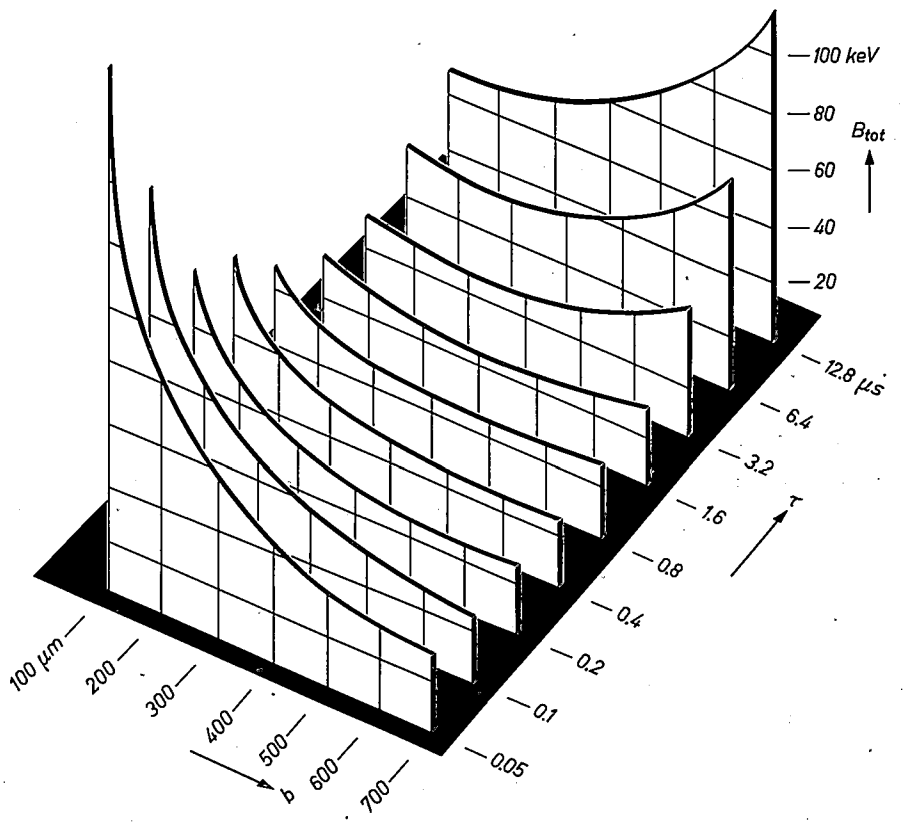


Fig. 8. Peak width B_{tot} of a silicon surface-barrier detector (surface area 2 cm²), measured as a function of the depletion depth b at different filter time constants τ.

Practical design of the detector

The gold electrode vapour-deposited on the silicon wafer must be extremely thin (see page 17) in order to minimize the absorption in the gold window. The gold layer is 20 nm thick, and only a few keV of the 6 MeV energy of α-particles are absorbed in it. Contact is made with the back face of the silicon wafer by evaporating an aluminium electrode on to it.

The silicon wafer (*Si* in the cross-sectional drawing of the detector shown in *fig. 9*) is cemented between two glass rings *R* with an epoxy resin *E*. The effect of the epoxy resin on the silicon surface is a marked reduction of the edge current. The coefficient of expansion of the glass rings is matched to that of the silicon, so that no unwanted mechanical stresses occur in the wafer even when the detector is cooled to −75 °C.

The detector holder is mounted on a standard coaxi-

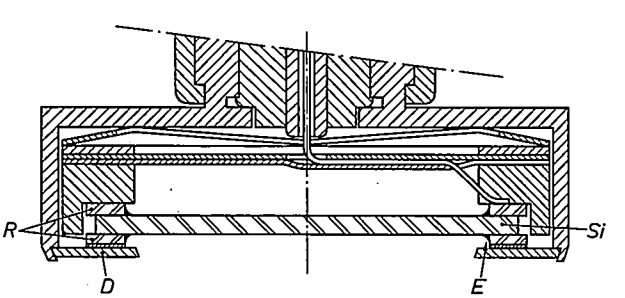


Fig. 9. Cross-sectional sketch of mounted detector. *Si* silicon wafer. *R* glass rings. *E* epoxy resin. *D* diaphragm.

[7] This value was found experimentally for high-energy γ-radiation. Whether it also applies to α-particles has not been verified.

[8] J. Lindhard and V. Nielsen, Nuclear collisions and ionisation fluctuations in charged particle detectors, Physics Letters **2**, 209-211, 1962.

[9] J. A. W. van der Does de Bye, Signal-to-noise ratio of a *P-N*-junction radiation counter, Philips Res. Repts. **16**, 85-95, 1961.

al plug (omitted in fig. 9); its centre conductor is connected to the aluminium electrode of the silicon wafer. A diaphragm D is mounted in front of the detector window.

The spectrum background

A measured α-spectrum always contains a background level that makes it impossible to distinguish very weak peaks. There are two effects contributing to this background. In the first place, it can happen that radiation is measured before a source is brought near to the detector. In this case the detector or its environment (the wall of the vacuum chamber) is contaminated by α-activity from a previous experiment. In the measurement of relatively strong sources this contamination can hardly be avoided, since such sources also emit recoil daughter nuclei, which may reach the counter. Sometimes a slight air pressure is admitted to the vacuum chamber and a retarding electric potential is applied between source and detector, in order to prevent these recoils from reaching the detector without stopping the α-particles. With an air pressure of 30 torr and a retarding voltage of 225 V at a distance of 1 cm between source and detector it has been found possible to reduce the contamination by a factor of 50 [10]. This reduces the resolution, however, and in experiments where a low background level is important we therefore prefer to use separate detectors, which should not be used with strong sources.

When these and other precautions are taken, the spontaneous detection is found to be very low; we recorded one pulse per cm^2 of detector surface per 24 hours in the energy range of natural α-radiation. This shows that the radioactivity of the counter material is negligibly low. It is also known that ambient γ-radiation and cosmic radiation give pulses that are a great deal smaller than the pulses from α-particles. In theory, then, the semiconductor detector is suitable for the detection of α-rays of very low intensity.

If there is a very high peak in the α-spectrum, however, this detection becomes difficult. Here — and this is the second contribution to the spectrum background — a high peak of this nature is always accompanied by a tail at the side of the lower energies. This tail can be seen in fig. 1 on the left of the main peak of ^{242}Cm. The tail appears because a certain fraction of the emitted α-particles lose a substantial amount of energy — varying, however, in magnitude from one particle to another — in various collision processes outside the sensitive volume of the detector. Collisions may take place inside the source itself, and outside the source there may be collisions with objects in the environment of the detector or with particles of dust on the detector surface. Another cause of an α-particle being counted in a channel with lower energy is the existence of local concentrations of recombination centres. The existence of these appears to prevent the whole of the free charge from arriving at the electrodes. In any case the tail involves a higher background level in the energy range below the energy of the main peak.

How this affects the detection of α-lines will now be illustrated with reference to fig. 1. In general we can say that the background level of 100 to 250 keV below the energy of the main peak is of importance for the observation of secondary peaks of low relative intensity. An example of this is the secondary peak of ^{242}Cm with an energy of 5967 keV, i.e. 145 keV below the main peak, and a relative intensity of 0.035%. The spectrum background from 200 to 450 keV below a main peak is also important in connection with the detection of related nuclides with low isotope ratio and a relatively long half-life; see for example the isotopes ^{243}Cm and ^{244}Cm in fig. 1, which are about 300 keV below the main peak. The height of the background in this energy range, and at energies farther removed from the main peak, also determines the relative concentration of impurities still measurable. The background level is expressed as a single number by giving the ratio of the height of the main peak to that of the background, in terms of the energy distance to the main peak. The value of this ratio depends on the resolution, but the advantage using it is that it indicates at once the relative intensities that can be detected above the background. Thus for the configuration used for detecting the spectrum shown in fig. 1, for example, we find a ratio of 2900:1 between the main peak and the spectrum background 200 keV below the main peak, while at 400 keV and 800 keV below the main peak the ratios are 6000:1 and 8000:1 respectively.

Although still not ideal, these ratios compare very favourably with what can be achieved with ionization chambers. In spite of the most rigorous precautions, the corresponding ratio for an ionization chamber is 250:1 at 1 MeV below the main peak [11].

Experiments show that all kinds of small details can affect the height of the spectrum background. In a very recent experiment, using a current form of detector, it was found possible to improve considerably on the above-mentioned ratios, the values measured being 21 100:1 at 400 keV and 39 700:1 at 800 keV below the main peak.

An α-particle spectrometer

A photograph of an α-particle spectrometer is shown in *fig. 10*. When this is combined with a kick-sorter (and a vacuum pump) we have a complete equipment for α-spectrometry. A block diagram of the equipment is shown in *fig. 11*. The pulses from the detector are fed to a preamplifier. Owing to the use of capacitive feedback, this preamplifier delivers voltage pulses that are

proportional to the charges supplied to the input. The magnitudes of these charges depend only on the energy of the incident particle and, within certain limits, not on detector capacitance or voltage, variations in which therefore do not affect operation. These variations may for example be caused by temperature fluctuations, which bring about variations in the reverse current; because of the high series resistance (e.g. 50 MΩ) such

Fig. 10. An α-particle spectrometer in use. The vacuum cylinder has been removed and a radiation source is being placed on the specimen holder. The height of the specimen holder can be adjusted under vacuum by means of the calibrated knob in front of the instrument. The detector is seen above the specimen holder. The meter on the left of the instrument can be used to indicate the pressure in the vacuum chamber.

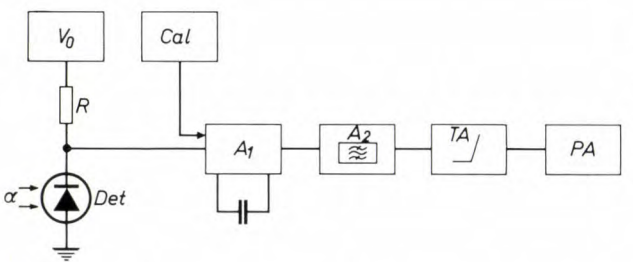


Fig. 11. Block diagram of a complete equipment for α-particle spectrometry. *Det* detector. V_0 voltage source. *R* series resistance. *Cal* calibrating pulse generator. A_1 preamplifier. A_2 main amplifier with bandpass filter. *TA* threshold amplifier. *PA* pulse-height analyser.

fluctuations can give rise to variations in the detector voltage. The voltage pulses are next fed to a main amplifier, which brings them to the required level for pulse-height analysis. The main amplifier also incorporates the band-pass filter mentioned earlier (page 18).

If it is required to use all of the channels of the available pulse-height analyser for finer analysis of part of

[10] A. Chetham-Strode, J. R. Tarrant and R. J. Silva, The application of silicon detectors to alpha particle spectroscopy, IEEE Trans. on nuclear science **NS-8**, No. 1, 59-63, 1961.
[11] R. Benoit, G. Bertolini and G. B. Restelli, Nucl. Instr. Meth. **29**, 149, 1964.

the spectrum a threshold amplifier is included. This suppresses pulses below a certain height and amplifies the pulses above it. The gain characteristic of the threshold amplifier is shown in *fig. 12*. A threshold amplifier was used, for example, for measuring the spectrum given in fig. 1; in this case all 400 available channels were employed for recording the spectrum between 5.2 and 6.2 MeV, giving a sufficient number of points for accurate display of the individual peaks.

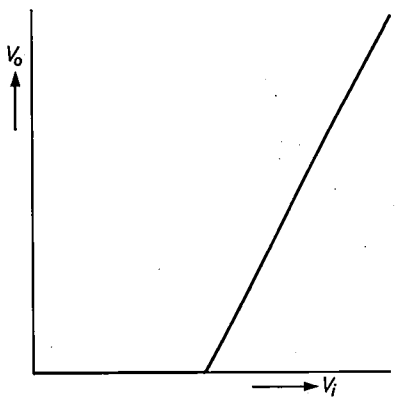


Fig. 12. Gain characteristic of a threshold amplifier. V_i input voltage. V_o output voltage.

The α-particle spectrometer in fig. 10 embodies all the functions mentioned here. The spectrometer is shown with the vacuum cylinder removed; a radiation source is being placed on a table that can be moved up and down by means of a screw-thread. The detector is situated above this table. The distance between source and detector can be adjusted between 2 and 12 cm with a calibrated knob outside the vacuum chamber, and this makes it possible to set the number of pulses per unit time to a suitable value. The pressure in the vacuum chamber should be lower than 0.1 torr during the measurement, to prevent energy absorption and line broadening. The preamplifier is mounted directly on the vacuum chamber. The threshold amplifier incorporated in this spectrometer permits a tenfold spread of part of the spectrum. The spectrometer also contains a pulse generator which can supply constant pulses with an accuracy of $\pm 15 \times 10^{-6}$, with a magnitude calibrated in MeV, to the input of the measuring circuit; these are used for calibrating the energy scale of a measured spectrum. The peak on the right of the spectrum in fig. 5 comes from this pulse generator.

Summary. The various special features of semiconductor detectors have led to their general use for the detection and energy measurement of charged particles (α- and β-particles, protons, etc.) and electromagnetic radiation (γ-rays); α-particle spectrometry is an appropriate example for illustrating the measurement technique. In α-particle spectrometry a good energy resolution and a low spectrum background are of paramount importance. In this respect the surface-barrier detector is the most suitable. Some special designs are described. It is shown that an approximate calculation of the energy resolution can be made. The spectrum background, which is attributable to many and various causes, is particularly high near to a high peak and makes it difficult to measure very low peaks near it. Special measures which can be used to improve the peak-to-background ratio considerably are described. At 400 keV below a high peak, for example, it is now possible to obtain a peak-to-background ratio of 6000:1 (a ratio as high as 20 000:1 has been obtained with a special arrangement). The electronic measuring circuits required for α-particle spectrometry are combined with the vacuum chamber containing the radiation source and detector to form a single piece of equipment, the α-particle spectrometer. This is described at the end of the article.

A muscle stimulator for hemiplegic patients

Certain nervous disorders or damage to the brain can have the result that the muscles used in lifting the foot during walking are not activated, or are only partly activated. A patient with this disability (sometimes called drop-foot) is in danger of stumbling over his own foot, and when walking therefore tends to remain standing on the sound leg until the affected one is safely and securely on the ground. This gives him a dragging gait, and even with a walking-stick he moves only with difficulty.

It has been found that this disability can often be largely overcome if an electrical stimulus is applied at the appropriate instant to the nerve that causes the muscles involved to contract. A portable electronic muscle stimulator for this purpose has been developed and tested by the Institute for Perception Research in cooperation with the Stichting Revalidatie (Centre for the Rehabilitation of the Physically Disabled), Eindhoven.

The main problem was to *synchronize* the periodically supplied stimuli with the patient's walk. In earlier investigations in this field experiments were carried out with a mechanical switch placed under the heel of the affected foot [1]. The disadvantage of this arrangement was that the stimuli were switched on and off at times which came too soon with respect to the desired cycle of lifting and lowering the foot. A better place for the switch was under the *ball* of the affected foot, but owing to the vulnerability and rapid wear of mechanical switches placed outside the shoe (and since these only work properly on hard, level ground) the solution was not satisfactory.

We have now obtained very good results with a pneumatic switch which is operated by a small air chamber in a rubber insole fitted in the shoe on the *unaffected* foot. The air chamber is situated under the ball of the foot; the instants at which the ball of this foot touches and leaves the ground or, to be more exact, the instants at which the weight of the body is placed on the insole and removed again coincide very closely with the start and the end of the swing of the affected foot. The instants at which the muscles of this foot contract and relax are thus controlled correctly from the movement of the patient as he walks.

Fig. 1 shows the appliance in use, and *fig. 2* the insole with the switch. The generator, i.e. the stimulator proper, is carried on a belt around the patient's waist, and delivers a rectangular pulse at a repetition rate of 50 Hz, 0.6 ms duration and a maximum amplitude of 100 V (rated current 6 mA). The amplitude of the pulse can be varied by means of a potentiometer. An elastic bandage immediately under the knee of the affected leg carries the two electrodes which are energized by the pulses. The positive (passive) electrode is made of electrically conducting rubber and is in direct contact with the skin. The other (active) electrode, to which a maximum of 100 V is applied, is made of stainless steel and makes contact with the skin through a damp sponge to avoid a high contact resistance.

The system has proved to be very reliable and the

Fig. 1. The muscle stimulator in use. A rubber insole containing a small air chamber, connected by a flexible tube to a pneumatic switch, is fitted in the shoe on the unaffected foot. When the patient walks, the pneumatic switch (attached to the lower leg by means of an elastic bandage) is operated and switches the stimulator on and off. The stimulator, which is carried on a belt around the waist, applies pulses to an electrode in the elastic bandage around the affected leg.

[1] W. T. Liberson, H. J. Holmquest, D. Scot and M. Dow, Functional electrotherapy: stimulation of the peroneal nerve synchronized with the swing phase of the gait of hemiplegic patients, Arch. phys. Med. Rehabil. **42**, 101-105, 1961.

switch operates on any kind of floor covering or paved surface without getting damaged. This is a very important point, since it gives the patient confidence in the appliance.

For some considerable time now the appliance has been under test at the Eindhoven Centre for the Rehabilitation of the Physically Disabled. As a result of using the appliance many patients have developed a practically normal gait. An important contributory factor was found to be that the affected leg is better directed when the foot is placed on the ground and therefore stands more securely in the "standing" phase. It is interesting to note that after some time many patients retain the improvement in their walk even *without* further use of the appliance. Out of 60 patients over 40 years of age who underwent the treatment for three to six weeks, 50 recovered completely or show considerable improvement. The other patients responded well while walking with the appliance, but showed no lasting improvement. Two of these patients with no autonomous muscle control have been using the appliance constantly for more than six months, and are walking confidently.

The appliance is also being used on children with a drop-foot disability due to a brain defect contracted before or during birth. These children were not able to learn to walk normally. Some of these children who are paralysed on one side only (hemiplegia) have been under treatment for several months and have now begun to develop a more normal gait. For children paralysed on both sides (diplegia) an appliance with two separately adjustable outputs is now being tested; the insole here is placed under the less-affected foot. This investigation is still in progress.

Fig. 2. Insole with flexible tube and pneumatic switch.

The results so far obtained with this appliance have demonstrated its helpful effect in the cases we have mentioned. It appears likely that the muscle stimulator will also be a useful aid for spastic children, and possibly for certain polio patients.

H. J. van Leeuwen
J. Vredenbregt

H. J. van Leeuwen is with the Stichting Revalidatie, Eindhoven; J. Vredenbregt is with the Institute for Perception Research, 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 *E*
Mullard Research Laboratories, Redhill (Surrey), England *M*
Laboratoires d'Electronique et de Physique Appliquée, Limeil-Brévannes (Val-de-Marne), France *L*
Philips Zentrallaboratorium GmbH, Aachen laboratory, Weisshausstrasse, 51 Aachen, Germany *A*
Philips Zentrallaboratorium GmbH, Hamburg laboratory, Vogt-Kölln-Strasse 30, 2 Hamburg-Stellingen, Germany *H*
MBLE Laboratoire de Recherches, 2 avenue Van Becelaere, Brussels 17 (Boitsfort), Belgium. *B*

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 applied to the measurement of the speed of sputter ion pumps.
Mass Spectrometry, Proc. Symp. Enfield College of Technology 1967, p. 129-137; 1968. *M*

G. Arlt & H. Schweppe: Paratellurite, a new piezoelectric material.
Solid State Comm. **6**, 783-784, 1968 (No. 11). *A*

L. K. H. van Beek, J. Boven & J. Helfferich: Properties of diazosulfonates, Part V. The effect of substituents on spectra and dissociation of 4-dimethylaminobenzenediazosulfonate.
Rec. Trav. chim. Pays-Bas **87**, 737-745, 1968 (No. 8). *E*

V. Belevitch & Ph. Piret: Théorie de l'oscillateur à deux fréquences d'accord dont le rapport est voisin de 3.
Rev. HF **7**, 171-176, 1968 (No. 6). *B*

G. Blasse: Influence of crystal structure on luminescence.
Mat. Res. Bull. **3**, 807-815, 1968 (No. 10). *E*

G. Blasse & A. Bril: Photoluminescent efficiency of phosphors with electronic transitions in localized centers.
J. Electrochem. Soc. **115**, 1067-1075, 1968 (No. 10). *E*

G. Blasse, A. Bril & W. C. Nieuwpoort: On the Eu^{3+} fluorescence in several mixed metal oxides.
Proc. int. Conf. on Luminescence, Budapest 1966, Vol. 2, p. 1646-1650; 1968. *E*

G. Blasse, A. Bril & J. de Vries: Fluorescence of Eu^{2+}-activated barium octaborate.
J. Electrochem. Soc. **115**, 977-978, 1968 (No. 9). *E*

R. Bleekrode: Contribution to the study of "abnormal" rotational energy distributions in the $C^3\Pi_u$-$B^3\Pi_g$ bands of $^{14}N_2$.
J. chem. Phys. **49**, 951-953, 1968 (No. 2). *E*

R. Bleekrode: A spectroscopic method for the determination of temperatures in hot gaseous systems.
Physics Letters **27A**, 673-674, 1968 (No. 10). *E*

A. J. van Bommel & F. Meyer: Reply to the comment of H. K. Lintz concerning "LEED study of a nickel induced surface structure on silicon (111)".
Surface Sci. **12**, 391-392, 1968 (No. 2). *E*

H. Bosma: A general model for junction circulators; choice of magnetization and bias field.
IEEE Trans. **MAG-4**, 587-596, 1968 (No. 3). *E*

G.-A. Boutry, H. Dormont, J.-C. Richard & P. Saget: Etude de films épitaxiés de potassium et de rubidium par diffraction d'électrons de faible énergie.
C.R. Acad. Sci. Paris **267B**, 255-258, 1968 (No. 4). *L*

G. Bouwhuis: On the anomalous circular polarization of the 1.523 μ He-Ne laser line.
Physics Letters **27A**, 693, 1968 (No. 10). *E*

A. Bril: Report on application of luminescence.
Proc. int. Conf. on Luminescence, Budapest 1966, Vol. 2, p. 2112-2117; 1968. *E*

A. Bril, W. C. Nieuwpoort, W. L. Wanmaker (Philips Lighting Division, Eindhoven), **G. Blasse & C. D. J. C. de Laat:** On the transfer of energy in europium-activated phosphors.
Proc. int. Conf. on Luminescence, Budapest 1966, Vol. 2, p. 1689-1693; 1968. *E*

C. F. Brockelsby: Mullard Research Laboratories, Redhill, Surrey.
N.E.C. Rev. (National Electronics Council) **4**, 117-124, 1968 (No. 4). *M*

J. J. van den Broek: Determination of ternary diffusion coefficients.
Scripta metall. **2**, 561-564, 1968 (No. 10). *E*

C. M. van der Burgt: Enkele toepassingen van ultrageluid, IV. Piëzomagnetische en piëzo-elektrische transducenten voor diverse ultrasone toepassingen.
Ingenieur **80**, O 168-181, 1968 (No. 45). *E*

K. H. J. Buschow, J. F. Fast & A. S. van der Goot: Magnetic properties of some Co-rich erbium cobalt intermetallic compounds.
Phys. Stat. sol. **29**, 825-829, 1968 (No. 2). *E*

H. B. G. Casimir: Chimistes, physiciens et ingénieurs, leur collaboration dans la recherche industrielle.
Proc. 36ème Congrès int. de chimie industrielle, Bruxelles 1966, p. 64-66; 1968. *E*

H. B. G. Casimir: On electromagnetic units.
Helv. phys. Acta **41**, 741-742, 1968 (No. 6/7). *E*

E. Constant (Faculté des Sciences de Lille) **& A. Semichon:** Diodes semiconductrices en régime d'avalanche. Application aux hyperfréquences.
Onde électr. **48**, 703-721, 1968 (No. 496/497). *L*

C. D. Corbey, R. Davies & B. H. Newton: Varactor harmonic-generator chain as a pump source for a parametric amplifier.
Electronics Letters **4**, 397-398, 1968 (No. 19). *M*

C. Crevecoeur & H. J. de Wit: Dipole relaxation in MnO, doped with Li.
Solid State Comm. **6**, 843-845, 1968 (No. 12). *E*

H. J. van Daal: The static dielectric constant of SnO_2.
J. appl. Phys. **39**, 4467-4469, 1968 (No. 9). *E*

J. Dieleman: On the utilization of electron spin resonance results: illustration with the aid of the (3*d*) transition elements in some II-VI compounds.
Proc. int. Conf. on Luminescence, Budapest 1966, Vol. 2, p. 2131-2141; 1968. *E*

A. M. van Diepen (Natuurkundig Laboratorium der Universiteit van Amsterdam), **H. W. de Wijn** (N.L. Univ. Amst). **& K. H. J. Buschow:** Nuclear magnetic resonance and susceptibility of equiatomic rare-earth-aluminum compounds.
Phys. Stat. sol. **29**, 189-192, 1968 (No. 1). *E*

C. Z. van Doorn: Recombination mechanisms for the "edge" emission in cadmium sulphide.
Proc. int. Conf. on Luminescence, Budapest 1966, Vol. 1, p. 1124-1130; 1968. *E*

C. Z. van Doorn & J. H. Haanstra: Simple improvement of Hinteregger-lamp continua by a titanium getter.
Appl. Optics **7**, 1655-1656, 1968 (No. 8). *E*

J. Durieu & Y. Genin: Determination of an orbit from only three angular measurements.
AIAA J. **6**, 1432-1433, 1968 (No. 7). *B*

G. Engelsma: Zeitmessung in Pflanzen.
Umschau in Wiss. u. Technik **68**, 727-728, 1968 (No. 23). *E*

U. Enz: Der photomagnetische Effekt.
Helv. phys. Acta **41**, 781-784, 1968 (No. 6/7). *E*

I. Flinn: Extent of the $1/f$ noise spectrum.
Nature **219**, 1356-1357, 1968 (No. 5161). *M*

N. V. Franssen: Über die Frequenzkurven elektroakustisch gekoppelter Räume und ihre Bedeutung für die künstliche Nachhallverlängerung.
Acustica **20**, 108-109, 1968 (No. 2). *E*

N. J. Freedman & L. K. Brundle: Nonlinear behaviour of magnetostatic surface waves.
Electronics Letters **4**, 427-428, 1968 (No. 20). *M*

Y. Genin: Gaussian estimates and modern linear discrete filtering methods.
Rev. A **10**, 70-75, 1968 (No. 2). *B*

A. H. Gomes de Mesquita: The direct structure determination of a silicon carbide crystal of type 120 *R*.
Acta cryst. **B 24**, 1461-1466, 1968 (No. 11). *E*

E. E. Havinga: W-like dependence of critical temperature on number of valence electrons in non-transition metal Cu_3Au-type alloys.
Physics Letters **28A**, 350-351, 1968 (No. 5). *E*

J. C. M. Henning & H. van den Boom: ESR studies on divalent iron in Cs_3ZnCl_5.
Physics Letters **27A**, 685-686, 1968 (No. 10). *E*

W. Hermann: Internal friction in III-V-compounds below room temperature.
Solid State Comm. **6**, 641-644, 1968 (No. 9). *A*

A. H. Hoekstra (Philips Lighting Division, Eindhoven): The reaction of calcium chlorine apatite with antimony trioxide.
Proc. int. Conf. on Luminescence, Budapest 1966, Vol. 2, p. 1337-1341; 1968.

Th. Holtwijk & W. Lems: The iron curtain, a batch-fabricated storage array.
IEEE Trans. **MAG-4**, 561-564, 1968 (No. 3). *E*

H. Jonker, Th. P. G. W. Thijssens & L. K. H. van Beek: Properties of diazosulfonates, Part VI. Quantum yields for the photolysis of 2-methoxybenzenediazonium and for the photo-isomerization of 2-methoxybenzene-*trans*-diazosulfonate.
Rec. Trav. chim. Pays-Bas **87**, 997-1005, 1968 (No. 10). *E*

H. Kalis & J. Lemmrich: Frequenzanaloge Drehzahlregelungen, I, II.
Regelungstechnik **16**, 497-502, 555-562, 1968 (Nos. 11, 12). *H*

D. J. Kroon: Opening tweede hoogbouw Philips Natuurkundig Laboratorium.
Ned. T. Natuurk. **34**, 344-346, 1968 (No. 11). *E*

K. Lagemann: Ein Vorschlag zur Darstellung asynchron betriebener JK-Flipflops.
Elektron. Rechenanl. **10**, 171-176, 1968 (No. 4). *H*

H. Leich (Faculté Polytechnique de Mons) **& Ph. van Bastelaer:** La synthèse des quadripôles non dissipatifs au moyen de capacités et de gyrateurs déséquilibrés.
Rev. MBLE **11**, 31-48, 1968 (No. 2). *B*

F. A. Lootsma: Constrained optimization via penalty functions.
Philips Res. Repts. **23**, 408-423, 1968 (No. 5). *E*

F. A. Lootsma: Constrained optimization via parameter-free penalty functions.
Philips Res. Repts. **23**, 424-437, 1968 (No. 5). *E*

F. K. Lotgering: Spin canting in $MnCr_2S_4$.
J. Phys. Chem. Solids **29**, 2193-2197, 1968 (No. 12). *E*

R. J. Meijer: Der Philips-Stirlingmotor.
Motortechn. Z. **29**, 284-298, 1968 (No. 7). *E*

C. H. de Minjer: Bepaling van de gevoeligheid van putvormige corrosie van roestvrij staal.
Polytechn. T. Procestechniek **23**, 587-595, 1968 (No. 17). *E*

P. Nagels (Solid State Physics Dept. SCK, Mol, Belgium), **M. Denayer** (S. S. Phys. Dept. SCK, Mol), **H. J. de Wit & C. Crevecoeur:** Comments on "The Hall mobility of electrons and holes in MnO at high temperature" by Gvishi *et al.*
Solid State Comm. **6**, 695-696, 1968 (No. 10). *E*

A. G. van Nie: Noise level and zero-drift of broad-band electrometers used for measuring small currents.
Electronic Engng. **40**, 520-523, 1968 (No. 487). *E*

M. Noé: Problèmes d'interconnexion optimale (1re partie).
Rev. MBLE **11**, 1-12, 1968 (No. 1). *B*

G. Piétri: The transfer method, a new technique for the sensitization of photoelectric tubes.
IEEE Trans. **NS-15**, No. 3, 171-178, 1968. *L*

L. J. van de Polder: Courbe de réponse spectrale désirée du tube „Plumbicon" et du système optique dans une caméra de télévision en couleur.
Onde électr. **48**, 880-883, 1968 (No. 499). *E*

K. Reiber: Absorption von Ultraschall in Tellur.
Solid State Comm. **6**, 779-781, 1968 (No. 11). *A*

J. Roos & J. Koorneef: Thin bilaminar piezodisks used as microphone and telephone membranes.
35th Conv. Audio Engng. Soc., New York 1968, preprint No. 589 (J-5), 3 pp. *E*

W. Schilz: Anisotropy of high-field helicon propagation in PbTe.
Phys. Stat. sol. **29**, 559-568, 1968 (No. 2). *H*

J. G. Siekman: Cutting of thin metal films with a CO_2-gas laser beam.
Microelectronics and Reliability **7**, 305-311, 1968 (No. 4). *E*

M. J. Sparnaay: On the attraction between two plates of semiconducting materials which are electrically connected.
Proc. Kon. Ned. Akad. Wetensch. **B 71**, 387-395, 1968 (No. 5). *E*

J. M. Stevels: Struktur und elektrische Eigenschaften von Gläsern.
Haus der Technik — Vortragsveröffentlichungen No. 188, 6-16, 1968. *E*

H. J. L. Trap: Anwendungsbereiche neuerer glasartiger Stoffe in der Elektronik.
Haus der Technik — Vortragsveröffentlichungen No. 188, 64-76, 1968. *E*

H. J. L. Trap: Elektronenleitung in Glas.
Elektro-Anzeiger, Ausg. ges. Industrie, **21**, 317-322, 1968 (No. 15). *E*

M. T. Vlaardingerbroek, G. A. Acket, K. Hofmann & P. M. Boers: Reduced build-up of domains in sheet-type gallium-arsenide Gunn oscillators.
Physics Letters **28A**, 97, 1968 (No. 2). *E*

M. T. Vlaardingerbroek, W. Kuypers & G. A. Acket: Energy relaxation time of hot electrons in GaAs.
Physics Letters **28A**, 155-156, 1968 (No. 2). *E*

G. Vonk: A new type of compact heat exchanger with a high thermal efficiency.
Adv. cryog. Engng. **13**, 582-589, 1968. *E*

J. Vredenbregt & W. G. Koster (Institute for Perception Research, Eindhoven): Measurements on electrical and mechanical activity of the elbow flexors.
Biomechanics I, 1st Int. Seminar, Zürich 1967, p. 102-105; 1968.

K. Walther: Anisotropy of magnetoacoustic attenuation and deformation potential in bismuth.
Phys. Rev. **174**, 782-790, 1968 (No. 3). *H*

W. L. Wanmaker, J. G. C. M. de Bres & J. W. ter Vrugt (Philips Lighting Division, Eindhoven): Fluorescence of some Ga and In containing alkaline earth orthophosphates.
Proc. int. Conf. on Luminescence, Budapest 1966, Vol. 2, p. 1370-1373; 1968.

W. L. Wanmaker (Philips Lighting Division, Eindhoven), **A. Bril, J. W. ter Vrugt** (Philips L.D.) **& J. Broos** (Philips Electronic Components and Materials Division, Eindhoven): Luminescent properties of europium-activated phosphors of the type $A^{III}B^{V}O_4$.
Proc. int. Conf. on Luminescence, Budapest 1966, Vol. 2, p. 1633-1638; 1968. *E*

C. H. Weijsenfeld: A core model explaining Hall effect and resistivity in the mixed state of type II superconductors.
Physics Letters **28A**, 362-363, 1968 (No. 5). *E*

K. Weiss: Thermische Ausdehnung von reinem und dotiertem NaCl.
Z. phys. Chemie Neue Folge **61**, 257-261, 1968 (No. 5/6). *E*

H. W. Werner & H. A. M. de Grefte: Improvements in noise level and zero drift of a broadband electrometer used for ion detection in a mass spectrometer.
Electronic Engng. **40**, 493-496, 1968 (No. 487). *E*

H. W. de Wijn (Natuurkundig Laboratorium der Universiteit van Amsterdam), **K. H. J. Buschow & A. M. van Diepen** (N.L. Univ. Amst.): Experimental evidence for interband mixing in rare-earth intermetallic compounds.
Phys. Stat. sol. **30**, 759-766, 1968 (No. 2). *E*

H. C. Wright, R. J. Downey & J. R. Canning: Conductivity storage in CdS.
Brit. J. appl. Phys. (J. Physics D), ser. 2, **1**, 1593-1601, 1968 (No. 12). *M*

G. Zanmarchi: Free and trapped electron recombination luminescence of the aluminium centre in cubic SiC.
Proc. int. Conf. on Luminescence, Budapest 1966, Vol. 2, p. 1482-1489; 1968. *E*

G. Zanmarchi: Luminescence of the aluminium centre in cubic SiC: dependence of the recombination rates on the intensity of the light excitation.
J. Phys. Chem. Solids **29**, 1727-1736, 1968 (No. 10). *E*

G. Zanmarchi & C. Haas: Magnon drag.
Philips Res. Repts. **23**, 389-407, 1968 (No. 5). *E*

A. L. Zijlstra & A. J. Burggraaf (Philips Glass Division, Eindhoven): Fracture phenomena and strength properties of chemically and physically strengthened glass, I. General survey of strength and fracture behaviour of strengthened glass.
J. non-cryst. Solids **1**, 49-68, 1968 (No. 1).

Contents of Philips Telecommunication Review **28**, No. 1, 1968:

H. J. Kramer: The SR 600 series of transistorized radio-link equipment (p. 1-20).
A. E. Serrure: Test results on 6 MHz telephone line equipment on small-bore coaxial cables on a trial route between Brussels and Dendermonde (p. 21-30).
P. W. L. van Iterson: 5 kW HF communication ISB transmitter with transistorized driving equipment series RZ 510 (p. 31-43).

Contents of Mullard Technical Communications **10**, No. 95, 1968:

S. Sijtstra: Vertical deflection amplifier for 150 MHz oscilloscope (p. 158-166).
K. Hart & F. G. Oude Moleman: Horizontal deflection amplifier for 150 MHz oscilloscope (p. 167-174).
P. A. Neeteson: Schmitt trigger as level detector: analysis and performance of FCL 101 (p. 175-181).
F. A. Sowan: Light units in SI (p. 182-184).

Contents of Mullard Technical Communications **10**, No. 96, 1968:

J. Catherall: Automatic magnetiser and flux-setter for d.c. permanent magnet motors (p. 186-189).
A. Ciuciura: Protection against flashover in television picture tubes (p. 190-192).
G. W. Broekema: C.r.t. grid drive circuit using d.c. restoration of the unblanking signal (p. 193-202).
M. C. Gander: Transistor video amplifier design for monochrome tv receivers (p. 203-208).

Contents of Valvo Berichte **14**, No. 4, 1968:

P. Gerdsen & F. Weitzsch: Grundlagen für die Weißpunktänderung bei Farbfernsehempfängern (p. 121-135).
H. Achterberg & J. Wölber: Die für eine gute Schirmbildqualität erforderlichen Eigenschaften des PAL-Decoders einschließlich der Verzögerungsleitung, I (p. 137-168).

PHILIPS TECHNICAL REVIEW

VOLUME 30, 1969, No. 2

Automatic drawing of masks for integrated circuits

C. Niessen and H. E. J. Wulms

Numerical control, which is already widely used in metalworking, is also a valuable aid in the manufacture of integrated circuits. A computer supplied with the circuit data can be used to create a control tape for a numerically controlled drawing machine, which then automatically draws or cuts the scribing-film masks required for producing the circuits. In a recently developed method the circuit data can be presented to the computer in a very abbreviated form as a coded description of the necessary components (e.g. the type of transistors, the geometry of a resistor, etc.), their location and the connections to be made between them.

Solid circuits, the most important type of integrated circuit, are made by forming zones of *P*-type and *N*-type material at certain places in a single crystal of silicon. In this way resistors, transistors and diodes can be created in the crystal, and any desired circuit can be obtained by connecting up the various components with aluminium strips as required.

The zones of *P*- or *N*-type material are formed by a series of diffusion processes, which are carried out by using a kind of photographic technique. The starting point is a substrate of *P*-type silicon on which an *N*-type layer has been grown. The surface of this layer is oxidized in a suitable atmosphere to form silicon dioxide. The oxide layer is next coated with a photosensitive lacquer and the lacquer is then illuminated through a *mask* (*photomask*) wich does not transmit light at the places where a *P*-type zone is required. The unexposed lacquer is then dissolved and removed. An etching process then follows, in which the SiO_2 is only etched away at places where no photosensitive lacquer is present (i.e. where the *P*-zone is required). Finally the exposed lacquer is removed, and the crystal is heated in an atmosphere containing boron. At the places where the SiO_2 has been etched away the boron diffuses into the *N*-type silicon, forming zones with *P*-type impurities, which can be made to penetrate to the depth required.

Since the various components of a circuit must be electrically insulated from each other, the first step is always an isolation diffusion, carried out with an *isolation mask*. In this process *P*-type impurities are diffused through the *N*-type layer, down to the substrate, in a pattern of channels: A number of islands of *N*-type material are thus formed, each surrounded by *P*-type silicon.

One of the *N*-type islands can now be formed into, say, a resistor. This is done by oxidizing the silicon surface again, and, in a rather similar process, performing a resistance diffusion through a *resistance mask* to form a *P*-type zone in the *N*-type island. This *P*-type zone is the resistor. After further oxidation, contact holes are etched into the SiO_2 at the ends of the resistor. The mask used for this third stage is referred to as the *contact-hole mask*. By covering the contact holes with aluminium the resistor can be connected with other components. In a fourth stage these connections are etched out of a homogeneously evaporated layer of aluminium with the aid of the *aluminium mask*. The resistor thus obtained is shown in cross-section in *fig. 1a*, and the four masks used are shown in fig. 1*b*.

Diodes and transistors are made in a similar way; the same masks used in forming a resistor are also used in the corresponding diffusions for diodes and

Ir. C. Niessen and Ir. H. E. J. Wulms are with Philips Research Laboratories, Eindhoven.

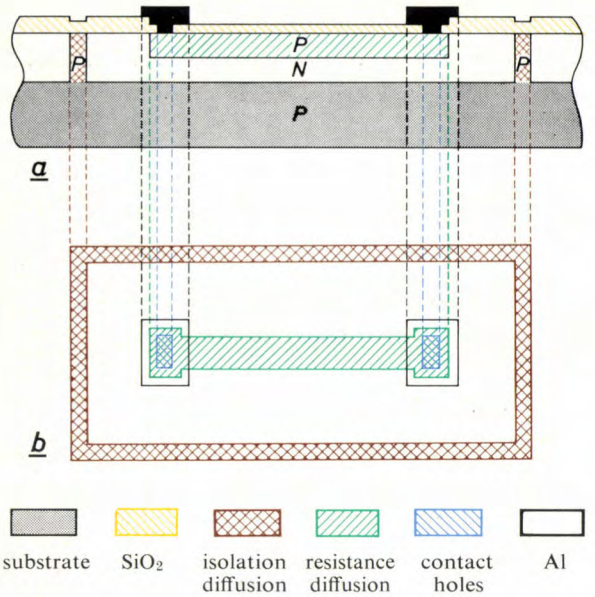


Fig. 1. *a*) Cross-section through a resistor in an integrated circuit. To illustrate the correspondence between the masks and the diffusion patterns, the patterns are shown in a simplified form.
b) Set of masks for making this resistor. The masks are drawn in different colours and placed one above the other.

transistors. We shall not go any further into the subject here, as the manufacture of such circuits has already been dealt with at some length in this journal[1].

Making the photomasks

In each stage of the process outlined above a photomask is needed for selective exposure. One circuit therefore requires the use of *several* masks.

The starting point in making the masks is the electrical circuit diagram. A circuit is designed that covers the smallest possible area and in which the components can be joined without crossovers by a pattern of aluminium strips. The characteristics of the components are determined by the dimensions of the various diffusion patterns. The designer must therefore accurately indicate the contours of the different masks on a single drawing. This drawing is usually made to a scale of 200:1, and a separate colour is used for each mask to aid identification. The designer also has to stick to certain rules: the distances between individual diffusion zones must not be smaller than a known minimum value and there is also a minimum value for the dimensions of some of the contours. The result is a very detailed layout drawing which contains all the data needed for making the circuit.

Next, with the aid of this layout, the contours have to be cut on the same scale in a sheet of scribing film, a plastic sheet consisting of two layers, one transparent and the other opaque.

The *scribing-film mask* is made by stripping away the opaque material from inside the cut contours. The final versions of the masks are now made by means of a photographic reduction and repetition process.

Formerly the drawing of the layout and the cutting of the scribing-film masks were done by hand. It was time-consuming work involving a considerable risk of error, and moreover checking was made difficult by the complexity of the pattern obtained when the plastic masks were placed one above the other.

The design contains many standard components, however, and also components that require no further specification of certain details: the contact holes of a resistor, for example, have to be located at a fixed distance from the diffusion boundary. In addition the circuits needed today tend increasingly to contain repetitive patterns. Automation is therefore obviously indicated. In this article an account will be given of the work done in this direction at Philips Research Laboratories in Eindhoven. The mask workshop set up here is equipped with a drawing machine, which can be numerically controlled with the aid of the COBRA[2], a small computer specially designed for the numerical control of machine tools. Here we shall briefly describe the method developed for automatically drawing and cutting the scribing-film masks with this machine.

Numerical control

Various articles on numerical control have appeared in this journal[3]. A numerically controlled machine has a control unit which is fed by a control tape. This is usually a paper, plastic or metal tape punched with the data which the control unit needs to make the tool (in this case a pen or an engraving tool) follow the desired path. Usually a considerable amount of data is needed to define this path. From very early on, computers were therefore used to produce the control tape from a much smaller quantity of *simply coded* input data. The computer programs developed for this purpose have done a great deal to ensure that numerical control really does ease the task of the designer. Much effort has gone into designing the program in such a way that a wide variety of details on the workpiece to be made are "known" to the computer. This has been done by adding a "library" of data to the program; and by means of a few appropriate instructions the

[1] A. Schmitz, Solid circuits, Philips tech. Rev. **27**, 192-199, 1966.
[2] R. Ch. van Ommering and G. C. M. Schoenaker, The "COBRA", a small digital computer for numerical control of machine tools, Philips tech. Rev. **27**, 285-297, 1966.
[3] See in particular: J. Vlietstra, The APT programming language for the numerical control of machine tools, Philips tech. Rev. **28**, 329-335, 1967.
[4] Art work is the name widely used in the American literature to describe the graphic design of an integrated circuit.

workpiece designer can get the data he requires out of the library and punched on the control tape. What is even more important is that the designer can have all kinds of *calculations* carried out. If he wants the tool to describe a circular path, for example, he only has to supply the computer with the word CIRCLE, followed by the coordinates of the required centre-point and the length of the radius. The computer then carries out the correct subroutines or procedures for calculating the data required for controlling the tool and punches this information on the control tape.

The program we have developed here for drawing and cutting scribing-film masks has an "input system" that enables the user to give a simple definition of the *components* of a required circuit, their *locations* and *connections* between them. The input code used, which is discussed below, is called DRAW (from Drawing of Art Work [4]).

The first step, designing an integrated circuit from the information contained in the circuit diagram, cannot yet be included in the program. It has proved exceptionally difficult to establish fixed rules for this work. The most compact possible layout of the components with no crossovers in the wiring is made by hand, and it remains a trial-and-error procedure based on intuition and experience.

Coding of the components

Now that the DRAW system is available, the circuit designer has only to code a list of the components and of their positions in the design. If the circuit contains repeated sub-circuits, the list does not even have to be complete. By means of a copying instruction the designer can arrange for a specified group of components to be repeated a number of times. The computer successively calculates all the detailed data needed for controlling the pen or engraving tool.

We shall now discuss at somewhat greater length the coding of a resistor. We can deal quite briefly with the coding of other components. After discussing the copying instruction, we shall conclude with a short account of the operation of the DRAW computer program.

Coding of a resistor

In our system a resistor can be coded with a single instruction provided the following conditions are satisfied: the resistance track must be of constant width and consist of one or more sections running alternately in the X- and Y-directions; the connecting strips must be chosen from a collection of standard types. The instruction to produce such a resistor is given by

$$\text{RES/X, Y/D, } F_1, F_2, \text{ØR/}A_1, A_2, \ldots, A_n$$

The parameters used in this instruction have the following meanings:

RES indicates that it is a resistor;

X and Y are the coordinates of the reference point, which fixes the position of the resistor in the design in accordance with a particular convention;

D is the (constant) width of the resistor track;

F_1 and F_2 are the code numbers of the resistor end-connection strips;

ØR is the orientation of the first section of the resistor track; an H or a V must be filled in here to indicate whether the first section is horizontal or vertical, i.e. parallel or perpendicular to the X-axis of the coordinate system being used;

$A_1, A_2, \ldots, A_n$ are the lengths along the centre lines of the different sections of the track; they are given a sign to indicate the direction.

To illustrate the significance of these groups of symbols, *fig. 2* shows a resistor obtained from the instruction:

RES/7.5, 20/2.5, 4, 2, V/—13, 3(7.5, 13, 7.5, —13)17.5

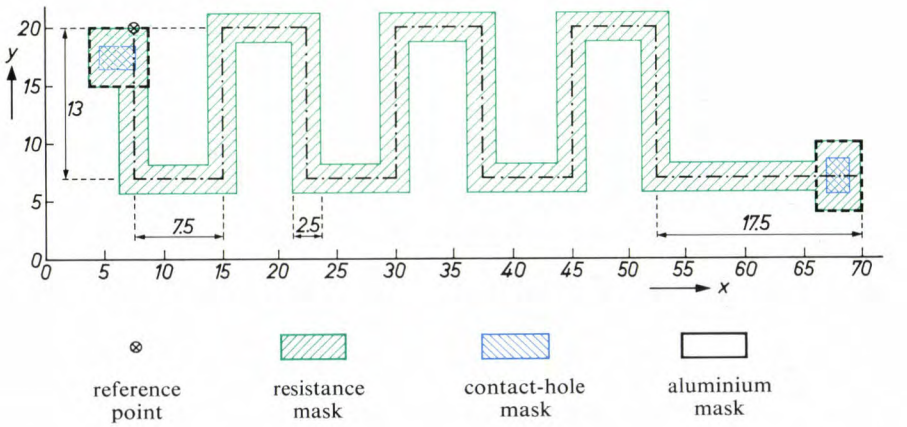


Fig. 2. Resistor with a "meander" track. The instruction for this resistor in the DRAW code is given in the text.

The instruction includes a row of numbers between brackets. The factor in front of the brackets indicates the number of times the sections described by these numbers have to be drawn; in this case it is three times. The final result of the instruction is that the contours are drawn or cut in *three* masks: a "meander" track in the resistance mask, contact holes in the contact-hole mask and connecting strips in the aluminium mask.

No data for the isolation mask are calculated in this resistor instruction. Several resistors may be placed in one island. A separate instruction is then needed to ensure that the appropriate island is drawn in the isolation mask of the integrated circuit.

It is also possible to code resistors that do *not* satisfy the conditions we have stated. This generally requires more than one instruction. However, a full account of all the things that can be done with DRAW lies outside the scope of this article.

Coding of a transistor

The designer can choose from 182 standard transistor types. Their dimensions, and hence their electrical characteristics, are the same as those of the "bread-board" transistors. These separate transistors are made in the same way as integrated circuits and the designer can connect them up to form an experimental circuit to check the electrical characteristics of his design. The bread-board transistors are characterized by a "name" beginning with a T followed by three letters and a figure. These refer to the library in the DRAW computer program, which includes the mask of all standard transistors.

An instruction to the computer to draw transistor masks would look like

TLLLN/X, Y/L, SP, ALPHA

The parameters used in this instruction have the following meanings:

TLLLN is the "name" designating the data to be extracted from the library in the DRAW computer program;

X and Y are the coordinates of the reference point; the convention for a rectangular transistor is that the reference point is the bottom left-hand corner;

L is the length of the emitter;

SP indicates an instruction for a reflection in one of the axes; SPX indicates that the transistor has to be reflected in the X-axis and SPY indicates a reflection in the Y-axis; if nothing is written here then there is no reflection;

ALPHA is the angle through which the transistor has to be rotated about the reference point; if nothing is written here then there is no rotation.

Copying instructions

A digital circuit is often built up from a number of identical basic circuits, and can therefore be described by repeating the coding of the components of the basic circuit with different coordinates. The CØPY instruction provides a more elegant solution: the basic circuit to be copied is preceded by the instruction BEGINCØPY and followed by ENDCØPY. The copying data are added after ENDCØPY. An example of a complete ENDCØPY instruction is

ENDCØPY/SP, ALPHA/DX, DY/N

where the parameters have the following meanings:

SP is again an instruction for a reflection; SPX is written for a reflection of the basic circuit in the X-axis and SPY for a reflection in the Y-axis of the coordinate system in use; if nothing is written here then there is no reflection;

ALPHA gives the angle through which the basic circuit must be rotated about the origin of the coordinate system; if nothing is written here then there is no rotation;

DX and DY are the distances by which the basic circuit is displaced parallel to the X- or Y-axis;

N is the number of copies required.

Example of a DRAW program

Fig. 3 shows the circuit diagram of an integrated circuit, whose complete coding will be given here as an example. The list of DRAW instructions for making

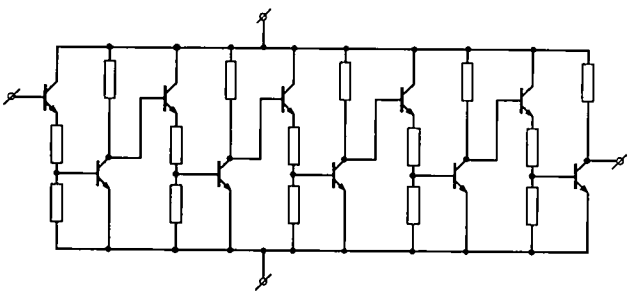


Fig. 3. Electrical diagram of a circuit taken here as an example.

the control tape for this circuit is shown in *fig. 4*, and *fig. 5* shows the drawing made with the aid of the control tape produced.

The circuit to be programmed consists of five identical basic circuits, each made up from two transistors and three resistors. It is sufficient to code this basic circuit once only: with two transistor instructions (TABT0), three resistor instructions (RES), one RHK instruction, which produces a rectangular contour in the isolation mask to enclose the three resistors in one island, and a number of ALU instructions for produc-

Fig. 4. Program in DRAW code for the integrated circuit corresponding to the diagram in fig. 3.

```
CARDNO      LINENUMBER
     1           0      BEGINCOPY
     2           1      TAB*0/10,118/12,
     3           2      TAB*0/10,48/12,SPX,
     4           3      RHK/11.5,51/3/31.5,64
     5           4      RES/18.75,78/3.5,1,1,V/32.5
     6           5      RES/33.75,78/3.5,1,1,V/32.5
     7           6      RES/16,59.25/5.5,4,,H/18.75,10.75,-16,7.75
     8           7      ALU/38.5,124/7,4/,-8/-5.5,-7.5
     9           8      ALU/27,125.5/7,3/,-2.5/.5,-1,4/,-4/-8,-10
    10           9      ALU/27,40.5/7,3/,4/-.5,1,4/,14/-6.5,
    11          10      ALU/32.25,42/7,3.5/,27/-15,11
    12          11      ALU/38.5,42/7,4/,41
    13          12      ALU/43.5,11/7,4/-35.5,/,33.5/17,9
    14          13      ALU/38.5,83/7,3.5/,13/7,
    15          14      ALU/6,96/7,4/5.25,/,17.75/10.25,7.25/,3
    16          15      ALU/40.5,147.5/7,4/10.5,
    17          16      ENDCOPY//37.5,/4
    18          17      RHK/7,20/3/190,126
    19          18      ALU/8,94.5/7,4/,-2/-8,-18/BV
    20          19      ALU/201,147.5/7,4/2,/8,-13/BV
    21          20      ALU/195,96/7,4/2,/14,-6/BV
    22          21      ALU/193.5,11/7,4/2,/15.5,8/BV
    23          22      URK/201,60/
    24          23      KRB/-20,/250,160
    25          24  EINDE
```

ing the pattern of aluminium strips for connections between these components. This group of instructions is preceded by BEGINCØPY and is followed by

ENDCØPY//37.5,/4

to indicate that the basic circuit is to be copied four times, and displaced each time by a distance of 37.5 mm in the X-direction. This is followed by some instructions that do not have to be repeated. The regions in the isolation mask that lie outside the islands formed by the transistor instructions and the RHK instruction constitute an almost complete

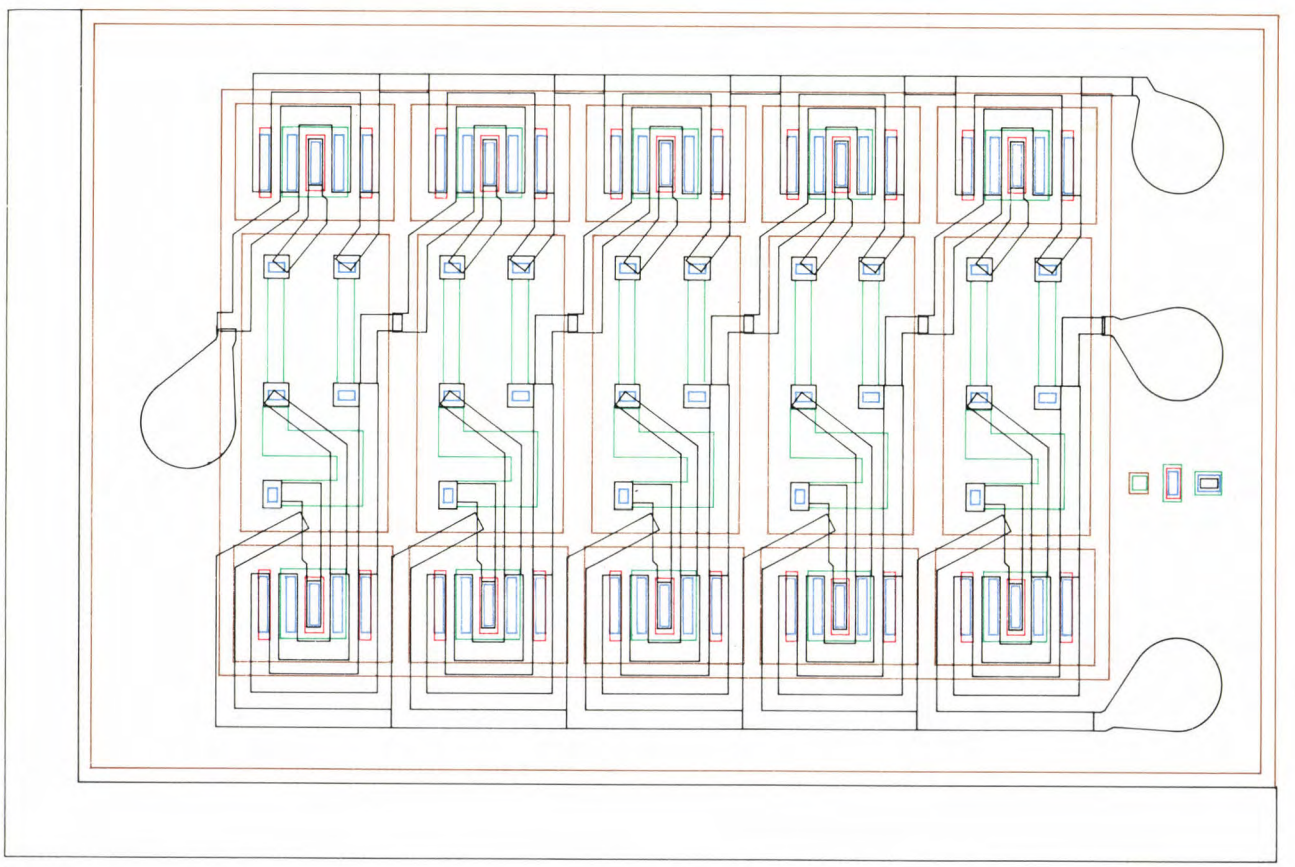


Fig. 5. A drawing made automatically with the control tape produced by the computer from the DRAW program given in fig. 4. A drawing of this kind can be used for checking the control tape.

pattern of channels. To complete this pattern a rectangular contour is needed around the islands already formed, and this is produced with a further RHK instruction. Also included in the program are some ALU instructions, for the connecting strips of the circuit; the KRB instruction, which provides a break line for the separation of the circuit from the other circuits produced in the crystal during the manufacturing process; and finally the URK instruction, which provides for the alignment marks for adjusting the mask on the silicon wafer.

Processing by the computer

The instructions, wich are punched on cards or on paper tape, are read by the computer and first checked for coding errors.

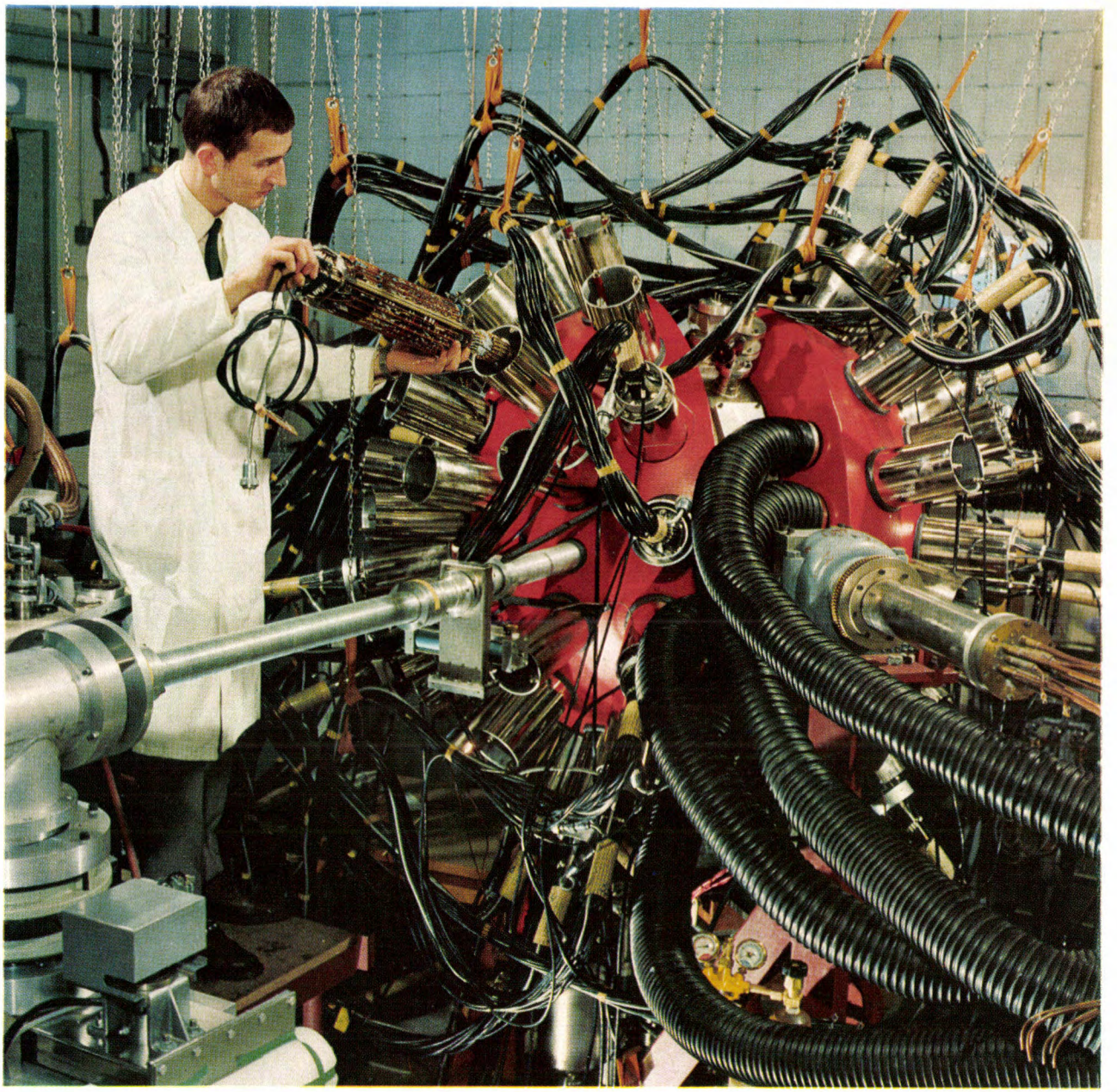
If no such errors are found, the computer starts to collect data for the first mask. This it does by inspecting all the instructions, and only operating on those instructions that relate to the first mask. Next the design is divided into a large number of regions, each of which is allotted a serial number. A serial number is also allotted to the calculated contours. All the data for the first mask are then punched on the control tape in the sequence of the serial numbers. This is done in such a way as to reduce the number of operations required later for drawing or cutting. The designer may however give his instructions in any arbitrary sequence.

The next thing that the computer does is to inspect all the instructions once again in order to collect data for the second mask and punch them on the control tape. This process is continued until all the masks needed for making the circuit have been processed.

The control tape is checked by feeding it to a numerically controlled drawing machine, which draws the masks in different colours one above the other. Once this layout drawing has been approved, the same tape goes to the mask workshop, where the masks are automatically cut in scribing film. No further check on whether these masks fit together is required. It is sufficient to make sure that the sheets of scribing film have been properly stripped. The procedure that we have described has greatly reduced the chance of errors, while much time is saved because the masks can be cut much faster and the design no longer has to be drawn to the last detail. The designer only has to draw contours required for the coding. Thus, the circuit of fig. 3 requires about 40 metres of control tape, containing some 16 000 characters. But the computer produces this control tape from the program shown in fig. 4, which contains only about 600 characters.

Summary. The article shows how numerical control can be applied to the manufacture of integrated circuits. It starts by describing the part played by scribing-film masks in the making of the components, and goes on to illustrate how the making of these masks can be simplified by drawing and cutting them automatically. This can be done by supplying a computer with only a limited number of easily coded data. The coding of resistors and transistors, and the use of copying instructions, are dealt with at some length. The article concludes with a complete program for making a simple circuit, and outlines the way in which a computer processes the instructions.

The "BOL" nuclear research project



Photograph Nuclear Research Institute, Amsterdam

The measuring system shown here, built by the Nuclear Research Institute, Amsterdam, supported by the Foundation for Fundamental Research of Matter (FOM), is designed for detailed investigations into the structure of atomic nuclei. Hydrogen or helium nuclei, accelerated in the Institute's synchro-cyclotron, travel through the narrow tube visible on the left and strike a target of the type of atom under study, which is situated at the centre of a spherical chamber. The mass number and energy of the reaction products, and the angle at which they leave the target, are determined by 64 telescopic detector units mounted in the wall of the sphere. In the photograph one of these units is being put into position. The system can also detect coincidences, i.e. simultaneous emissions of different reaction products [1]. Semiconductor detectors in the form of "checker-board counters", with 10×10 squares, were specially designed and built for this system by Philips Research Laboratories [2]. Each telescope at present contains one such detector and a second semiconductor detector; the number may later be extended.

The installation includes an elaborate system of electronic circuits and computers for controlling the measuring process and processing the enormous amount of information to be collected. The hoses on the right of the photograph supply cooling air for the electronic equipment.

[1] The system has been briefly described by K. Mulder, J. E. J. Oberski, R. van Dantzig and L. A. Ch. Koerts: On the BOL system, Proc. Int. Symp. on nuclear electronics, 1968, Part 2, p. 120/1-120/7.

[2] The principle of these detectors has been described by W. K. Hofker in: Semiconductor detectors for ionizing radiation, Philips tech. Rev. **27**, 323-336, 1966.

The polytypism of silicon carbide

A. H. Gomes de Mesquita

Silicon carbide is a substance with many interesting applications, some of which impose some strict requirements on material control. One of the greatest problems here is the phenomenon of polytypism, that is, the existence of innumerable different ways of stacking the double layers of carbon and silicon atoms forming the crystal lattice. This gives rise to as many variations again in some of the physical properties.

Silicon carbide is found only very sporadically as a naturally occurring mineral, and then only in unexploitable quantities. Very nearly all of the material used today is therefore made in laboratories or in industrial processes. The first synthesis dates from the last century. Since that time, interest in the substance has steadily increased, because of its interesting combination of chemical and physical properties. Silicon carbide is exceptionally hard and rigid and possesses great tensile strength. It is resistant to chemical action up to high temperatures, it is an excellent heat conductor which will withstand high temperatures, and it is also a semiconductor. Such a material clearly has interesting potential applications in various fields.

Its hardness was the first property to be used: powdered silicon carbide is well known as an abrasive and polishing agent. Next came applications as an electric heating element, as a voltage-dependent resistor, and as a resistor with a negative temperature coefficient primarily for use at high temperatures. In the last few years several types of diode have been made of silicon carbide, although not as yet on a large scale. The latest development is the work now in progress on the strengthening of plastics and metals by thin "whiskers" (thread-like crystals) of SiC.

This brief summary of properties and applications, which is far from complete, explains why Philips Research Laboratories have been carrying out investigations on silicon carbide for the last twenty years or so. This work has contributed to technical development and has led to many publications [1].

The chronological sequence of the principal applications: abrasive — heating element — diode, clearly indicates that the control over the material has improved with time. A little impurity in an abrasive does not usually do much harm, but in semiconductor materials it is essential to be able to keep the degree of impurity under accurate control. This applies both to chemical impurities and to physical imperfections, the latter category including stacking faults and dislocations that disturb the crystal lattice of the substance.

In fact, there has been considerable progress in the preparation of silicon carbide. While the Acheson industrial process — which uses a mixture of sand, coke, common salt and sawdust as starting materials — is still widely used, a pure laboratory synthesis is now available which gives a product whose total content of chemical impurities has been reduced to less than one foreign atom in 10^7 silicon and carbon atoms.

Control of the physical imperfections has not yet progressed so far. This is directly related with the phenomenon of polytypism, which is found in its most pronounced form in silicon carbide. This implies that, unlike nearly all other known substances, silicon carbide is known to occur in a very large number of crystalline modifications (types) which, although closely related, differ from one another quite distinctly. Up to now about 140 have been counted, and others are still being discovered.

The crystal structures of all these modifications can be described in the hexagonal system with an a-axis of approximately 3 Å (0.3 nm); the length of the c-axis, however, varies from 5 Å to 1000 Å (0.5-100 nm). This latter distance is 100 times greater than the range of all known ordering forces in crystals — a strange feature indeed.

Polytypism, that is to say the occurrence of a large number of types of one substance with crystal structures differing only in one dimension, is observed not only

Dr. A. H. Gomes de Mesquita is with Philips Research Laboratories, Eindhoven.

in silicon carbide but also in zinc sulphide, cadmium sulphide and cadmium iodide, to quote a few examples. There is also quasi-polytypism, which is found in substances in which slight differences in chemical composition affect the stacking of the atomic planes in only one direction. The members of the large group of ferrites that includes materials like $BaFe_{18}O_{27}$ and $Ba_2Zn_2Fe_{12}O_{22}$ are probably the best-known examples. Here too the crystal structure can be described in the hexagonal system, and the *a*-axis of the unit cell is fixed, in this case at 6 Å, while the *c*-axis varies from 15 to 1500 Å.

The physical aspects of polytypism are particularly interesting. The energy difference between the electrons in the highest levels of the valence band and the lowest levels of the conduction band is found to vary with the length of the *c*-axis, though the relation is not a simple linear one. In silicon carbide this "indirect" band gap can assume values between 2.4 and 3.3 eV, a variation of as much as 30 to 40%. The mobilities and the effective masses of the electrons are intimately connected with the crystal structure. The electroluminescence generated in silicon carbide may have almost any colour of the rainbow, depending on the crystal modification. The colour of material doped with nitrogen (or other donors) is strongly dependent on the type of structure.

In spite of a great deal of research, it is not yet completely understood why there are so many structure types. The crystal structure of a substance like silver iodide seems to be determined by slight deviations from the stoichiometric composition; in silicon carbide this has not been shown to be the case. In zinc sulphide, structures with large repeat distances occur during the cooling of crystals already formed; however, there is every indication that in silicon carbide the various structure types occur during growth.

In this situation it is not as yet really possible to treat the phenomenon of polytypism in a truly general way. The facts and arguments about silicon carbide which are presented in this article are therefore not always necessarily applicable to other substances in which polytypism occurs.

Crystal structure

Since the turn of the century there has been a great deal of crystallographic research on silicon carbide. Studies of the external crystal faces brought out at an early stage the existence of a number of different modifications with trigonal, hexagonal, rhombohedral and cubic symmetry and with widely different lattice constants. With the aid of X-ray analysis methods very many more modifications have been discovered, and in a large number of them the crystal structure has also been determined on an atomic scale [2]. We shall now discuss some of these modifications in greater detail, at the same time explaining the nomenclatures used to indicate the different types of structure.

To picture the crystal structure of silicon carbide, think first of a flat layer consisting exclusively of silicon atoms which are at a distance apart of 3.08 Å and are arranged in a close-packed structure of equal spheres (*fig. 1*). The (two-dimensional) periodic pattern is a rhombic unit cell with sides of 3.08 Å. Vertically above this layer, and at a distance of 1.89 Å, we now place a congruent layer of carbon atoms. In the resultant

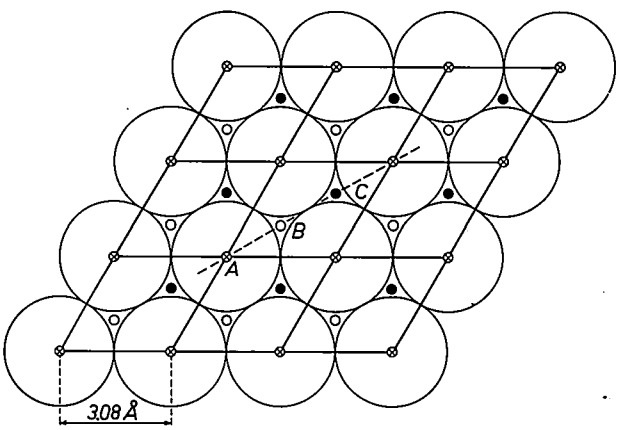


Fig. 1. Close-packed arrangement of atoms in a plane. The two-dimensional repetition pattern is a rhombic figure with atoms at the position *A* (the corner points) and interstices at *B* and *C*. In an isolated layer the last two positions are equivalent, but this is not true any more when the stack contains more than one layer.

double layer we can distinguish two sets of interstices at the sites marked *B* and *C*. Above one of these two sets we now place the next double layer, in which there are again two sets of interstices, and so on. Note that there are *two* possible ways of placing each double layer; this is the *essential characteristic of polytypism.*

With this method of stacking, all the atoms come to lie on axes of the type *A*, *B* and *C*, which are perpendicular to the plane of the drawing of fig. 1. Each crystallographic unit cell contains one axis of each of these three types. Since they lie in one plane, the whole crystal structure can be represented two-dimensionally by a cross-section in the plane *ABC* at right angles to fig. 1.

[1] See W. F. Knippenberg, Growth phenomena in silicon carbide, Philips Res. Repts. **18**, 161-274, 1963; W. F. Knippenberg and G. Verspui, Influence of impurities on the growth of silicon carbide, to be published in Mat. Res. Bull.

[2] Until recently this appeared to be a hopeless task, but now a fairly simple standard method exists that can be used in a large number of cases. On this subject see: M. Farkas-Jahnke, Acta cryst. **21**, A173, 1966; A. H. Gomes de Mesquita, Acta cryst. **B24**, 1461, 1968 (No. 11).

If successive double layers occupy the consecutive positions *ABCABC*... (*fig. 2a*) the result is a structure which closer examination shows to be cubic (sphalerite structure), which means that by the transformation of coordinates we can describe the structure in the cubic system. Stacking in the opposite direction *CBACBA*... also corresponds to the cubic structure. We shall not go any further into this subject here, but simply note that cubic β-SiC is the end product of many chemical reactions in which silicon carbide is formed.

If, during the crystal growth of β-SiC, a stacking fault is made in which the sequence of the layers is reversed, changing for example from *ABC*... to *CBA*..., the result is a twinned crystal (fig. 2*b*): *ABCABCBACBA*. The repeated occurrence of a stacking fault is referred to as multiple twinning. Where the frequency of their occurrence is great, and the distance between the faults irregular, the structure

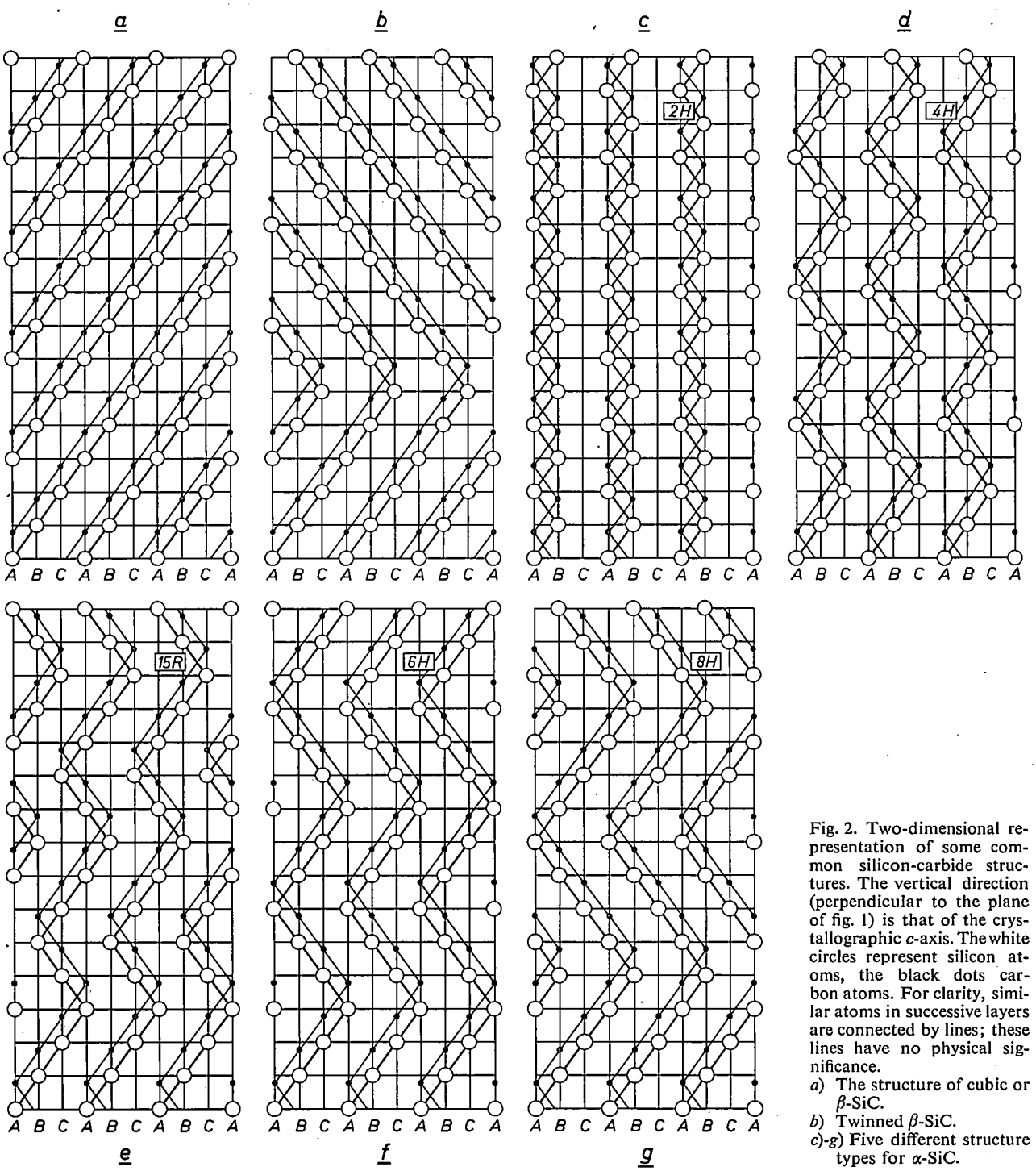


Fig. 2. Two-dimensional representation of some common silicon-carbide structures. The vertical direction (perpendicular to the plane of fig. 1) is that of the crystallographic *c*-axis. The white circles represent silicon atoms, the black dots carbon atoms. For clarity, similar atoms in successive layers are connected by lines; these lines have no physical significance.
a) The structure of cubic or β-SiC.
b) Twinned β-SiC.
c)-*g*) Five different structure types for α-SiC.

is said to be disordered, which is here one-dimensional disorder. In some structures the stacking may even be completely random. Experiments have shown that all these modes of stacking occur in silicon carbide.

It is also possible, however, for the stacking order to change direction frequently and *periodically*. The many modifications of silicon carbide in which this occurs are generally referred to as α-SiC. The individual designation indicates that α-SiC is regarded as a form of silicon carbide that is clearly distinct from the cubic phase. In particular, it should not be seen as a kind of misgrowth of β-SiC.

Some of the most frequently encountered structures of α-SiC are shown in figs. 2*c-g*. The stacking *ABAB*... (fig. 2*c*) corresponds to the well-known wurtzite structure. Along the crystallographic *c*-axis, i.e. the vertical direction in fig. 2, the repetition pattern consists of two double layers. Closer examination shows that the unit cell is hexagonal. This modification is therefore often given the designation 2H.

Somewhat more complicated is the stacking *ABCBABCBA* . . . which is also hexagonal (fig. 2*d*). The atomic stacking is repeated after every four layers, giving what is called a 4H structure. Since the environment of the layers in which the atoms are situated on the *B*-axis is the same as in the cubic structure, they are known as "c" layers, unlike the layers with atoms on *A* and *C* whose immediate environment corresponds to the 2H structure and are therefore referred to as "h" layers. In the nomenclature system based on this the 4H structure is described as hchc . . .. Another system of nomenclature, introduced in 1945 by Zhdanov [3], counts the number of layers situated between the reversal points in the atomic stacking. In the 4H structure this number is always two, and the structure can thus be described as (22). Similarly 2H may be described as (11), while β-SiC is given the symbol ∞ (for infinity).

Figs. 2*f* and 2*g* now present no further difficulties. The modifications shown there are 6H and 8H respectively, which may also be described as hcchcc and hccchccc or (33) and (44).

The structure shown in fig. 2*e* is rather more difficult. Cubic and hexagonal layers alternate in the pattern with hcchc and this atomic stacking is repeated after every five layers. Seen along the crystallographic *c*-axis, however, these five layers do not form a repeat distance. If the atoms of the first layer lie on the axis *A*, then those of the sixth layer do not lie on *A* but on *B*. Now the structure can be described in a cell containing only five layers, but this crystallographic cell is not hexagonal but rhombohedral. If, for the sake of clarity, we also wish to describe the rhombohedral structures in a hexagonal system, we take the unit cell three times larger, so that it contains — in this case — 15 layers. The structure illustrated in fig. 2*e* is therefore called 15R or $(32)_3$ [4].

The α-SiC structures we have discussed so far are among the simplest encountered; 6H, 15R and 4H, in that order, are also the most common. We shall not give a description of all the other silicon-carbide structures. To illustrate the possible degree of complexity however, we can quote the following [5].

$$174R = [33333633333333334]_3;$$

$$120R = [322222222222322233]_3;$$

$$33T = [3333353334];$$

$$168R = [232323232323232232333]_3.$$

Polymorphism and polytypism

The occurrence of more than one crystalline modification of a substance is not in itself particularly unusual. The phenomenon is known as polymorphism when it occurs in a compound, and allotropy when it occurs in an element. Standard examples are white and grey tin, graphite and diamond, and white and violet phosphorus, where the same solid has two forms of different appearance.

In the case of tin there is a phase transition at 13 °C. At atmospheric pressure the grey tin is stable below that temperature, and above it the white. The transformation takes place so very slowly, however, that objects made of pure tin do not usually decompose as a result of ambient temperature fluctuations. Nevertheless, this does sometimes happen and it is known as "tin plague".

Diamond and graphite can be reversibly transformed into one another at very high pressures. Under atmospheric conditions graphite is the stable form; the transformation of diamond into graphite, however, does not take place at a speed that can be measured.

Investigations have shown that in these cases there is a specific region of existence ("homogeneity region") for each modification, i.e. a certain range of pressures and temperatures within which it is stable. This is not the case for violet and white phosphorus; the white variety is metastable under all conditions. The direct conversion of violet to white phosphorus is therefore not possible. The reverse transformation from white to violet phosphorus does of course take place, too

[3] G. S. Zhdanov, C. R. Acad. Sci. U.R.S.S. **48**, 43, 1945.

[4] In an analogous way the structure of ß-SiC is also designated as 3C. The figure "3" again indicates the repeat distance along the crystallographic *c*-axis, and the letter "C" indicates cubic symmetry.

[5] There is only one cubic modification. All the others are rhombohedral (R), hexagonal (H) or trigonal. In the literature the trigonal modifications are usually also denoted by H, but sometimes however by T, thus here we have 33T.

slowly to be measured at room temperature, but at a measurable speed at higher temperature.

Returning now to our subject proper, we can at once see the connection with what has just been discussed. Clearly, the question of what causes polytypism leads at the same time to a question about the stability of the different structure types. Are regions of existence to be indicated for a few or for all the modifications, and if so, why is it that under certain conditions one structure is more stable than all the others?

As we saw in the above examples, transformations in the solid state are often very slow. The fact that a single furnace charge usually yields many different types of silicon-carbide crystals at the same time, and that different types of structure may even frequently be found in the same fragment of silicon carbide therefore proves very little about their stability. It chiefly illustrates that phase transitions in silicon carbide are also very slow processes, certainly at room temperature, so that we do not reach the equilibrium situation, in which a single stable structure occurs. At higher temperatures one might expect the transition effects to be accelerated, but even here, strange to say, transitions *in situ* have never been observed. This approach thus does not yield direct evidence showing whether any type of structure is stable or not.

In principle, information about the stability of certain modifications can also be obtained from the results of accurate measurements of thermodynamic quantities, such as the heat of combustion and the specific heat as a function of temperature. For silicon carbide these quantities have been determined in only two cases: for β-SiC and for α-SiC of the type 6H. The values of the heat of combustion found were identical within the accuracy of the measurements, and more exact experiments are therefore needed in order to demonstrate any possible differences. Nor is this so very surprising, since the disposition of the nearest and next-nearest neighbours of the atoms is the same in all structures. Related to this is the fact that the densities of the different structure types differ by no more than about 0.1%. In tin, carbon and phosphorus the corresponding differences are very much greater. We saw in the introduction that the conspicuous features of polytypism are the large number of modifications and the peculiarly large repeat distances in the crystal lattice. In view of what we have just said we ought to add that polytypism may also be distinguished from normal polymorphism by the subtlety of the differences that determine the stability of the crystalline phases. Because of this it has still not been found possible to give any clear demarcation of the possible existence regions, even for the most common structure types of silicon carbide.

Some theories on polytypism

In the last twenty-five years many special theories have been put forward to account for the occurrence of polytypism. This again shows that polytypism is regarded as something which is clearly distinct from normal polymorphism and requires distinct treatment. In some theories each modification is considered to be a stable phase. Most of them, however, assume that fortuitous causes such as impurities or lattice defects give rise to the growth of long-period structures. In some of the theories the two types of argument — thermodynamic and growth-kinetic — are combined.

The first attempt at an explanation was made by Lundqvist [6], who attributed polytypism to the effect of chemical impurities. He based this view on differences which he had observed in the aluminium content of industrially prepared 4H, 15R and 6H crystals (*fig. 3*).

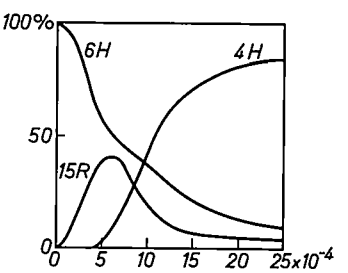


Fig. 3. The relation between the aluminium content and the relative number of crystals (in %) of the three most common types of α-SiC (after D. Lundqvist [6]).

Some initial experiments carried out by Knippenberg of these Laboratories with aluminium-doped silicon carbide crystals gave no confirmation on this point, for with crystals grown under otherwise identical conditions it was not possible to show any relation between aluminium content and crystal structure.

Recently, however, it has been shown that the growth of a few simple structure types can be stimulated by the addition of certain impurities — discovered quite by chance. With complicated structures this has never yet succeeded; for that matter, it would be difficult to see how the presence of impurities could cause the formation of these structures. This, and the fact that highly complicated structures, like 387H, have also been found in crystals of the highest purity, indicate that we must look elsewhere, at least for the origins of the more complicated structures.

A later theory attributes the growth of all except a few of the simplest structures to the effect of screw dislocations. The idea underlying this theory, originally put forward by Frank [7], can be understood by pictur-

ing a growing crystal, for example of the type 6H, in which for one reason or another a screw dislocation occurs during growth, and is propagated along a growth spiral. A very clear diagrammatic illustration of such a crystal-growth mechanism has been given by Read [8]; this is shown in *fig. 4*. When the Burgers vector of the

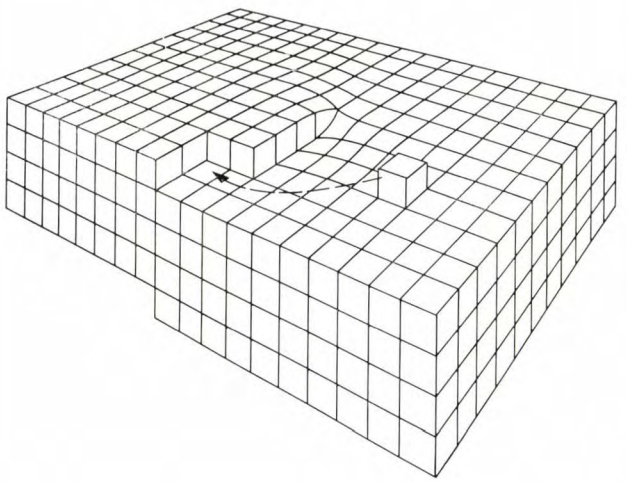


Fig. 4. Schematic representation of crystal growth by means of a screw dislocation mechanism (after W. T. Read [8]). The height of a block is equal to the Burgers vector. In Frank's theory [7], this is at the same time the repeat distance of the structure type undergoing growth.

dislocation, i.e. the height of one block in the figure, is a whole multiple of six layers, then according to this theory the 6H structure will grow without further modification. If the length of the Burgers vector is equal to five layer-spacings, however, then we find the 15R structure $(32)_3$, while seven layer-spacings gives $(331)_3 = (34)_3 = 21R$, eight layer-spacings gives $(332)_3 = (35)_3 = 24R$, eleven layer-spacings gives $(3332)_3 = 33R$, and so on. This theory can thus explain many structures that have been found, and indeed can even predict new structures, many of which have since been found. This theory is also supported by the fact that structures with large repeat distances are usually found in parallel intergrowth (syntactic coalescence) with simple types such as 4H, 15R or 6H; it therefore seems quite possible that the simple structures serve as a basis for the more complicated ones. Sometimes, too, beautiful growth spirals are found on the basal planes of silicon-carbide crystals as visible evidence of the presence of a screw dislocation inside the crystal (*fig. 5*); the pitch of the spiral has in some cases been found to be equal to the length of the crystallographic *c*-axis (repeat distance) of the type that has been grown, demonstrating the relation between the crystal structure and the length of the Burgers vector.

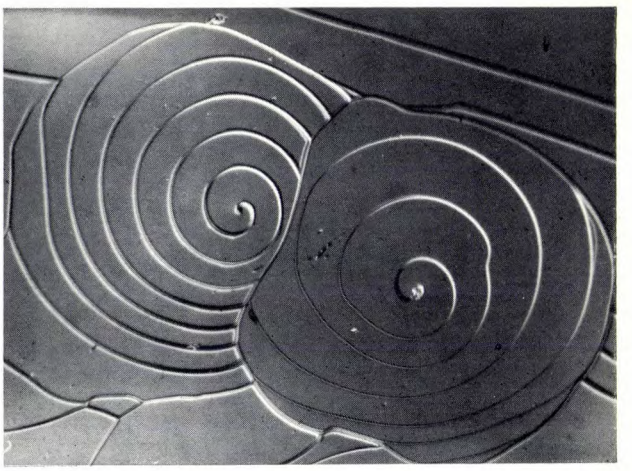


Fig. 5. Growth spirals on the surface of a silicon-carbide crystal. This photograph and figs. 8*b* and 10 were supplied by courtesy of Dr. W. F. Knippenberg.

However, there are a large number of cases in which this relation cannot be shown at all, either because of the complete absence of any growth spiral where one would be expected, or because the pitch does not correspond to the periodicity of the structure. Vand and Hanoka [9] have pointed out that this does not invalidate the theory, because there is always a reasonable chance that the crystal surface has grown differently from the interior from which, with the aid of X-ray diffraction methods, the structure has been determined.

There are however other objections to the screw-dislocation theory. The energy of such a dislocation is to a first approximation proportional to the square of the length of the Burgers vector. For a length of the order of 1000 Å this energy becomes very large, and it is therefore not clear why the crystal should be able to provide this energy and not that for a large number of edge dislocations and stacking faults, which would require much less energy. Moreover, since these faults can pass through the screw dislocation, they would destroy the periodicity of the screw and hence of the structure. Nor is it to be understood why — except in the 2H structure — two successive hexagonal double layers (corresponding to the figure "1" in the Zhdanov symbol) have never yet been found in any known type of SiC. If this were already an energetically unfavourable configuration, the energy effect would still be insignificant compared with the energy required for some screw dislocations.

[6] D. Lundqvist, Acta chem. scand. **2**, 177, 1948.
[7] F. C. Frank, Phil. Mag. **42**, 1014, 1951.
[8] W. T. Read, Jr., Dislocations in crystals, McGraw-Hill, New York 1953, p. 144.
[9] V. Vand and J. I. Hanoka, Mat. Res. Bull. **2**, 241, 1967.

Partly because of some of the above-mentioned objections, Jagodzinski [10] put forward a theory about fifteen years ago, the "vibration-entropy" theory, which is based on quite different assumptions. The basic assumption of this theory is that β-SiC is really the stable modification. The cubic crystals would have the outer form, however, of an octahedron, a closed figure for which in this case growth is slow—an assumption which can be justified. The growth process can be greatly accelerated by the introduction of stacking faults, and larger crystals are then formed. If we select these crystals for our investigations, then we are applying a negative selection. To explain why the stacking faults are built into the cubic structure *periodically*, the following reasoning is used. The differences in lattice energy between the structure types must be very slight, because the disposition of the nearest and next-nearest neighbours of the atoms is the same in all structures. If we neglect these differences, all that really remains for us is to consider the probability of the formation of each of the structures. At first sight one might be inclined to say that any disordered stacking of double layers is more probable than any ordered succession, but upon closer consideration this is not so evident. In fact, at the high temperatures at which SiC crystals are formed the vibrations of the atoms in the crystal lattice (lattice vibrations) represent a significant part of the energy content. Now these vibrations are seriously impeded by irregularities in the lattice. Hence a periodic stacking would after all be preferred. The lattice vibrations play such an important part in this case because their (three-dimensional) effect is set off against the effect of the one-dimensional sequence of the stacking of the atomic layers. This theory is therefore based entirely on probabilities. The probability of an ordered lattice is high. However, as the stacking of the double layers becomes more complicated and the difference compared with a completely disordered structure becomes smaller, the probability of disorder increases. The theory thus predicts a certain probability of stacking faults, which increases with the length of the stacking period.

These predictions were tested experimentally by Jagodzinski and at first found to be correct. Disorder is found in many crystals of simple structure types, and in the majority of crystals with structures with large repeat distances. Crystals of the type 6H and 15R are found to be either well-ordered, or the degree of disorder (the relative number of disordered double layers) fluctuates around 12%, a value which the theory suggests is likely (*fig. 6*).

There is however some doubt about the stability of β-SiC which Jagodzinski has postulated, and this tends to undermine his theory. An even greater difficulty with this theory is that no evidence has ever been found of any statistical distribution of the disorder, which would be expected with a probability theory. On the contrary, such investigations as have been made have always shown that the disorder was localized in the crystals. Moreover, various examples have meanwhile been found of structures that have large repeat distances but no demonstrable disorder, the occurrence of which would be extremely improbable according to the theory. Finally Jagodzinski's experimental results themselves give indications that it is not the structures with large repeat distances that have most disorder but the *simple* structures.

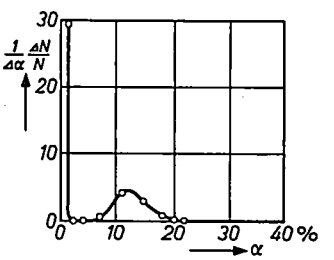


Fig. 6. The frequency distribution of the degree of disorder α in an arbitrary sample of N crystals ($N = 150$), practically all of the sample being of the types 6H and 15R; ΔN is the number of crystals with a degree of disorder between α and $\alpha + \Delta\alpha$. (After H. Jagodzinski [10].)

In spite of their differences, the three theories noted above are in agreement that the formation of the various types is to some extent a matter of chance. The thermodynamic stability of the higher structures plays no part in these theories [11]. It seems improbable, therefore, that these theories are applicable to the simple SiC structures, where there are sound reasons for accepting the existence of real differences in free energy. Knippenberg [1], for example, has found that the frequency of occurrence of types 4H, 15R, 6H and 8H is a function not only of the degree of impurity, as mentioned earlier, but also of the growth temperature (*fig. 7*). Furthermore, the energy difference between the highest level of the valence band and the lowest level of the conduction band appears to differ considerably in these and various other simple types. It therefore seems not unreasonable to assume that the contribution of the electron energy to the total lattice energy is dependent on the structure. The lattice constants and the average "thickness" of the double layers vary, though only very slightly, with the type of structure, as has been shown for types 2H, 6H and 3C. Refined X-ray-analytical determinations have shown that "h" layers are somewhat thicker than "c" layers. Presum-

ably, this again will involve small variations of the lattice energy.

At this stage it is difficult to establish which structures are to be considered as true equilibrium phases; this is still the subject of much debate. From the available experimental material one might be inclined to draw a distinction between structures with large and small repeat distances, and attribute an existence region to the latter structures only. In this connection we should mention the work of Van Vucht and Buschow [12] of these Laboratories on compounds of rare earths (R) with aluminium of the type RAl_3. These have structures, related to SiC, of the types 2H, 9R, 4H, 15R, 6H and 3C. Which type is formed depends on the atomic radius of R and on the growth temperature. For instance, the structure of $ErAl_3$ corresponds to 3C, $HoAl_3$ corresponds to 15R, while $DyAl_3$ takes the 15R or 4H structure, depending on the temperature. There can be no doubt about the thermodynamic stability of these structures. Zinc sulphide has the same structure as silicon carbide; below 1000 °C the cubic phase is stable, above 1200 °C the 2H structure. Studies of the transition region have yielded evidence in support of the stability of the 4H and 6H types of zinc sulphide grown in that region.

No such evidence exists for the complicated types, whether of SiC or of ZnS. Rather than attempting to explain their existence and origin in terms of unknown long-distance interactions between atoms, we are inclined, like Frank and Jagodzinski, to attribute them to special circumstances during growth. To determine whether this is justified or not, we shall have to look at what is known about the inter-relationships between growth conditions, crystal habit and structure types. The investigations of these subjects have nearly all been carried out since the theories we have mentioned were put forward. Conversely, these theories are largely based on data applicable to industrial-quality silicon carbide and not to material prepared under well-defined conditions. It is therefore not surprising that the later crystal-growth experiments have led to the postulation of a new and alternative explanation for the polytypism of silicon carbide. This will be dealt with at the end of this article.

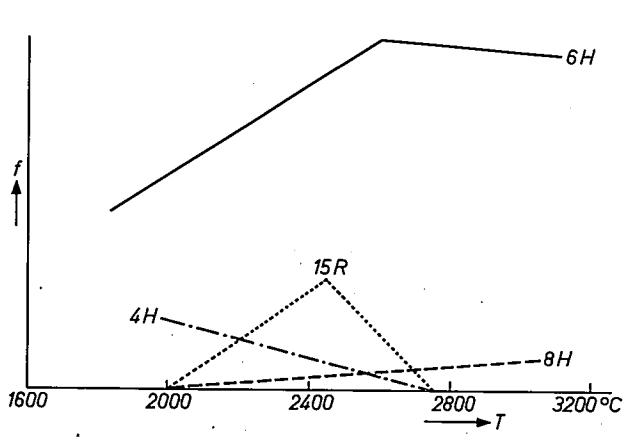


Fig. 7. Relation between the crystallization temperature and the relative quantities f of resultant crystals of a number of simple structure types. (After W. F. Knippenberg [1], as are also figs. 8*a*, 9 and 11.)

The growth of silicon-carbide crystals

As we mentioned earlier, the raw materials for the industrial preparation of silicon carbide by the Acheson process are sand, coke, common salt and sawdust. The reaction $3C + SiO_2 = SiC + 2CO$ takes place in a furnace whose centre is first heated to 1900 °C, then slowly to 2700 °C, and is finally maintained for some time at a temperature of 2000 °C. The chemical reaction in fact starts to take place at 1800 °C; the higher temperatures are needed to effect the recrystallization, which results in an end product that can be used as a grinding and polishing agent.

In laboratory syntheses this process is divided into two stages:

a) the preparation of SiC,

b) the recrystallization.

The initial materials used for the formation of extremely pure SiC are of course quite different from those mentioned above. One such material is methyltrichlorosilane, CH_3SiCl_3, which is decomposed on a pure graphite rod at 1300-1800 °C in a hydrogen atmosphere, resulting in the formation of SiC. With this process, known as the Van Arkel process, crystals of high chemical purity can be prepared. Crystallographically, however, the products are very impure; the crystals are of very irregular shape and are built up from a skeleton of α-SiC of indeterminate structure, between which large amounts of β-SiC have grown, constituting the bulk of the crystals.

The recrystallization takes place at higher temperatures, normally at 2300-2700 °C. For this next operation the SiC obtained is made into hollow cylinders (*fig. 8a*) which are heated in a graphite crucible in an argon or helium atmosphere (Lely's method [13]). On the outside surface of the cylinders there is partial decomposition of SiC, while at the same time there is transport of material to the inside surface of the cylinder, where large crystals are formed, usually in the

[10] H. Jagodzinski, Neues Jahrb. Mineral. Monatsh. **3**, 49, 1954.

[11] A theory that attempts to explain the phenomenon of polytypism on thermodynamic grounds, but which has found little or no experimental support, has been put forward in: C. J. Schneer, Acta cryst. **8**, 279, 1955.

[12] J. H. N. van Vucht and K. H. J. Buschow, J. less-common Met. **10**, 98, 1966.

[13] J. A. Lely, Ber. Dtsch. Keram. Ges. **32**, 229, 1955.

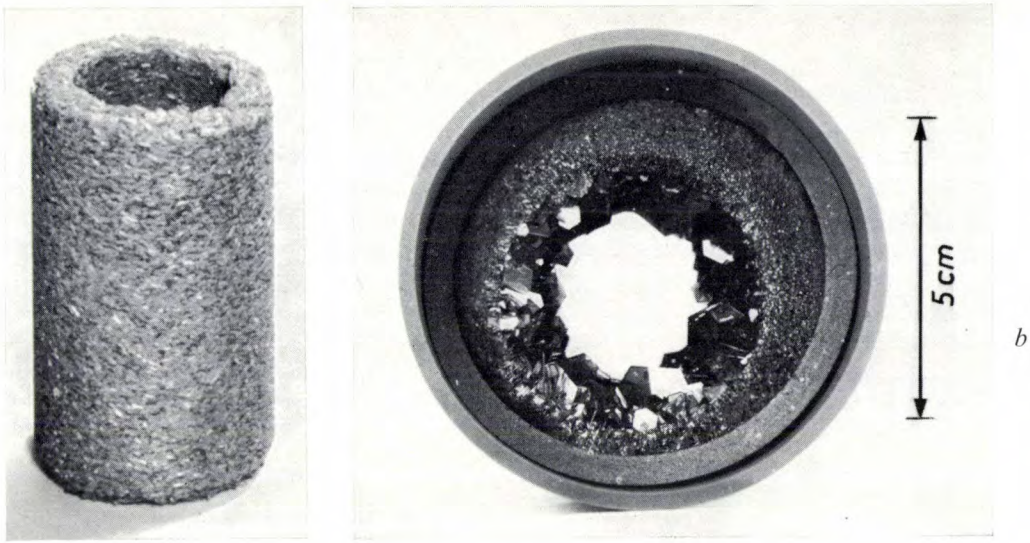


Fig. 8. Hollow cylinders of silicon carbide, *a*) before and *b*) after recrystallization in a graphite crucible.

shape of platelets (fig. 8*b*). Very occasionally polyhedral crystals are found among them (*fig. 9*). Among the crystal platelets the types 6H, 15R and 4H are by far the most strongly represented; the relative quantities depend to some extent on the recrystallization temperature. If the cylinder is rapidly cooled, β-SiC is deposited from the supersaturated vapour upon the hexagonal platelets (*fig. 10*). The higher structure types are found only occasionally, as a rule intergrown with one of the simpler structures mentioned above. Most crystals, moreover, exhibit non-periodic stacking faults.

The crystal structure of these platelets closely resembles that of industrially prepared SiC. The difference is that their chemical purity is very much higher.

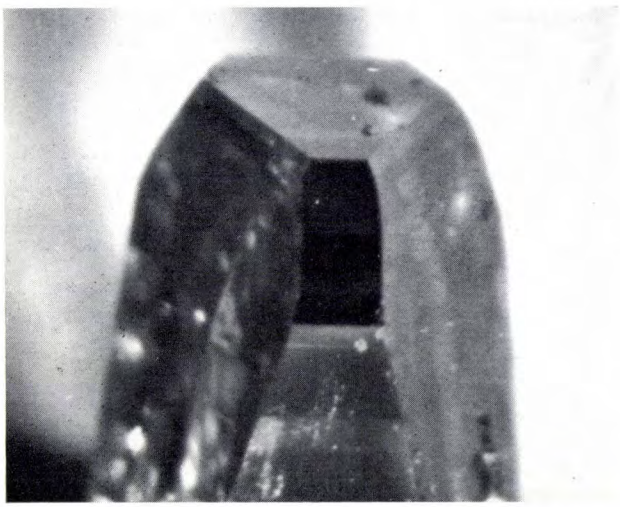


Fig. 9. Columnar crystal of SiC grown among platelet crystals at 2550 °C in an argon atmosphere.

The polyhedra, which we will presently deal with at greater length, usually consist of pure 6H, in which stacking faults or other types of structure do not occur.

Apart from the above, there are many other methods of preparing silicon-carbide crystals. The reaction between carbon and SiO_2, the first stage of the Acheson process, has been investigated in detail by Knippenberg [1]. When the reaction takes place between 1000 °C and 1800 °C by means of gas transport or surface diffusion, whiskers of various shapes are formed, as described earlier in this journal [14]. These whiskers are usually of cubic structure. The 2H structure is only formed as a result of the presence of certain impurities. If monocrystalline α-SiC is used as a substrate, the

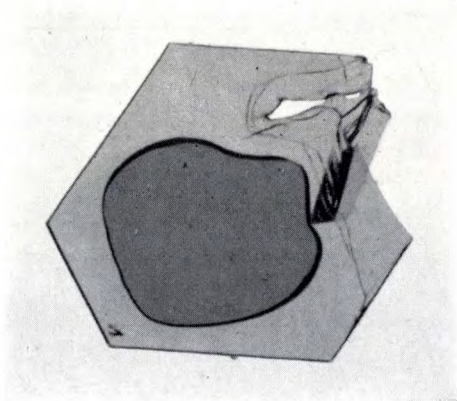


Fig. 10. Hexagonal crystal platelet on which a circular piece of cubic silicon carbide has formed during rapid cooling of the furnace.

whiskers growing at right angles to the *c*-axis of the substrate assume the same structure. However, the whiskers growing along the *c*-axis are usually of cubic structure. In other words, the hexagonal stacking in the substrate continues in a parallel direction in the whiskers. If, however, new layers are formed in the whiskers, then in this situation cubic stacking takes precedence. Above 1900 °C reactions in the gas phase give platelet crystals of the types 6H and 15R.

Silicon-carbide crystals can also be prepared from *melts* of different composition. Microcrystalline β-SiC has been prepared from silicon melts and crystals of the type 6H have been prepared from various metal melts. A related but more refined method is the travelling-solvent method, known as TSM [15], in which a chromium zone is introduced between a single-crystal seed and a polycrystalline rod. Using a kind of zone-refining process, where the chromium is drawn through the rod by means of a heat treatment, the seed can be made to grow at the expense of the rod. With this method, under identical conditions, it has been found possible to grow the types 4H, 15R and 6H with no change of structure type, in the direction of the *c*-axis. It is therefore evident that crystals of this structure growing under these conditions need no specific outside influence to enable growth to continue; they seem to have a kind of internal memory which regulates the stacking of the atomic layers. Unfortunately it is not yet known whether this is also the case for the more complicated structures, since growth experiments of this kind have not yet been performed with such structures.

Attempting now to survey the whole field of crystal-growth experiments, we arrive at the following broad summary.

The simple types of structure can be formed under widely different conditions, the 6H structure being by far the most frequently encountered.

The complicated structures, on the other hand, are only obtained (as far as is known at present) at higher temperatures in crystals grown from the gas phase or by surface diffusion during reactions or recrystallization (Lely's method). The great majority of these crystals are in the shape of platelets.

The growth mechanism of these crystals, prepared by the Lely method, has been investigated by Kroko [16] and by Knippenberg [1] using an ingenious method.

These investigators found that the crystals, which in themselves are virtually colourless, can be coloured by the addition of a small quantity of nitrogen during growth. This makes 6H crystals green, 15R yellow, 4H brownish, and so on. If nitrogen is then suddenly admitted into the furnace during recrystallization, crystals are obtained with sharp colour transitions, so that two stages of growth can be distinguished one from the other. *Fig. 11* gives an example which shows that the crystals in the form of platelets, or sometimes wedges, grow in a lamellar form on the polycrystalline cylinder wall. A fact entirely in accordance with this mechanism is that when different structure types are present in one crystal, they also tend to occur in the form of

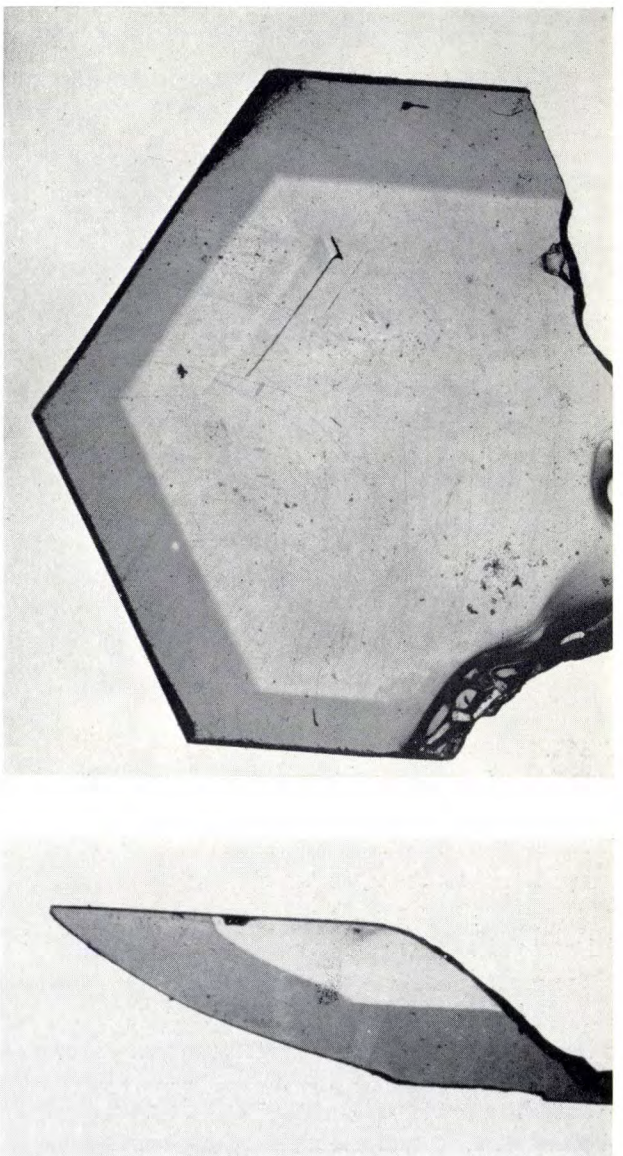


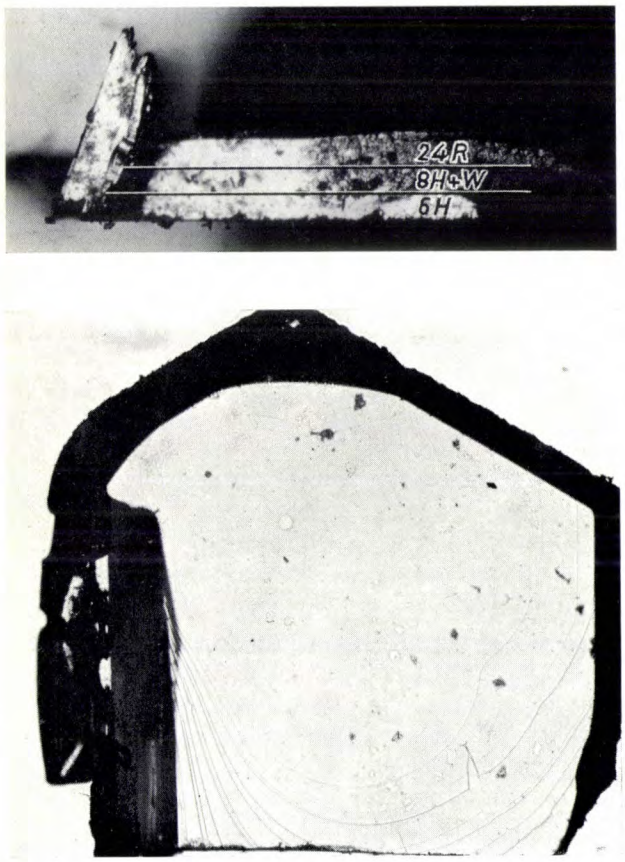
Fig. 11. Wedge-shaped crystals of α-SiC to which nitrogen has been added at an advanced stage of growth. This has made the stages in the growth clearly visible.
a) Plan view of the basal plane. *b*) Cross section parallel to the hexagonal axis, i.e. perpendicular to the plane of (*a*).

[14] W. F. Knippenberg, H. B. Haanstra and J. R. M. Dekkers, Philips tech. Rev. **24**, 181, 1962/63; H. B. Haanstra and W. F. Knippenberg, Philips tech. Rev. **26**, 187, 1965; W. F. Knippenberg and G. Verspui, Philips tech. Rev. **29**, 252, 1968 (No. 8/9).

[15] W. F. Knippenberg and G. Verspui, Philips Res. Repts. **21**, 113, 1966.

[16] L. J. Kroko, J. Electrochem. Soc. **113**, 801, 1966.

stacked lamellae. This can be demonstrated by grinding away layers from the crystal and making an X-ray exposure at each stage. It can then be seen that the structure types disappear successively from the X-ray diagram. We have carried out an experiment of this kind on a crystal which contained, in addition to a zone with random stacking faults, zones with 6H, 8H and 24R (*fig. 12*). An analogous case, reported by Kuo Chang-lin [17], relates to the structure types 6H, 15R and 69R.



a

b

Fig. 12. Two crystal platelets of α-SiC intergrown. *a*) Side view; *b*) plan view of the larger crystal. The structures indicated in (*a*) were determined by taking X-ray diffraction pictures after repeated grinding away of layers of the crystal (*W* represents the disorder).

Crystal intergrowths as a possible cause of polytypism

It appears that the problem of polytypism in silicon carbide is of a dual nature, and that a distinction must be made between the simple structures and the more complicated ones. The reasons have already been mentioned. The simpler structures grow under widely different conditions, their growth is sometimes stimulated by impurities, there is evidence of differences in thermodynamic stability, stable analogues are found among the intermetallic compounds, and in certain circumstances they grow in single-crystal form. The more complicated structures, on the other hand, are formed exclusively at higher temperatures from the gas phase and by surface diffusion, their occurrence is fortuitous and no means of stimulating their growth are known; structurally they resemble the simpler types (174R is like 6H, 168R like 15R, and 120R like 4H), they have no stable analogues and almost without exception they occur intergrown with one or more of the simple types.

The explanation for the growth of the simple structures does not yet seem to be within reach; there are rather more promising ideas for the structures with large repeat distances, but none of the theories mentioned seem really satisfactory. Now all of these theories assume that the crystal under investigation has its own *internal* memory controlling the growth, in the form of a screw dislocation, say, or lattice vibrations, so that the stacking of the atomic layers is continued in the way it has been started. Zhdanov and Minervina, on the other hand, suggested more than twenty years ago [18] that the growing structure might well have an *external* memory instead. Their actual suggestion was that the structure of a crystal is determined by the way in which it is intergrown with other crystals.

We have recently taken up this suggestion, and we have noted [19] that the higher structure types occur in platelet crystals which are always found to grow in the Lely furnace with the "memory direction" (the *c*-axis) roughly parallel with the polycrystalline cylinder wall. An external memory-influence, which may even differ considerably from one place to another, is thus quite conceivable; it could perhaps reside in the relative orientation of wall and growing single crystal (*fig. 13*).

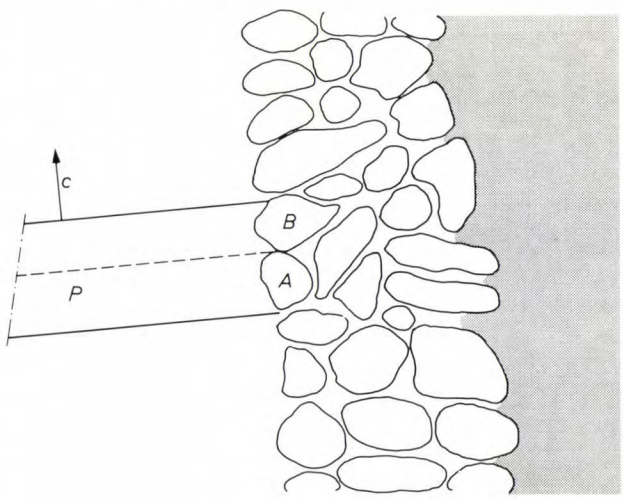


Fig. 13. Illustrating the growth of crystal platelets on a polycrystalline cylinder wall. The SiC double layers are imagined to be perpendicular to the plane of the drawing, the direction of stacking (*c*-axis) in the plane of the drawing. There are reasons for assuming that the crystal platelet *P* consists of lamellae whose structure is partly determined by intergrowth with the grains (*A*, *B*, . . .) of the wall (see text).

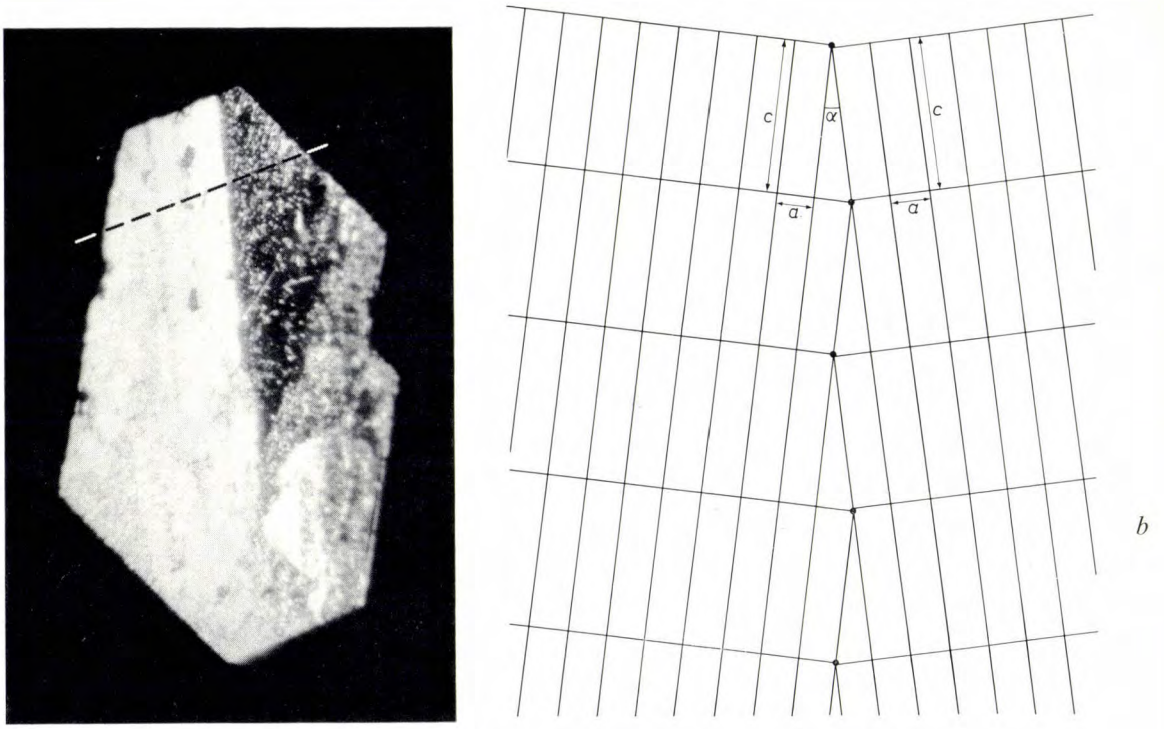


Fig. 14. *a*) Top face of a crystal of the type 33R placed in a beam of light. The crystal clearly consists of two parts at a small angle to one another. One part is exactly in the right position to reflect the incident light, and thus gives a very strong reflection; the other part is not. *b*) Diagram of a cross-section of the type 33R crystal. This cross-section is taken along the dotted line of (*a*) and at right angles to the crystal face shown there. The hexagonal axes of the two halves of the crystal enclose a small angle α, which probably determines the type of structure formed ($\alpha = a/c$; see text).

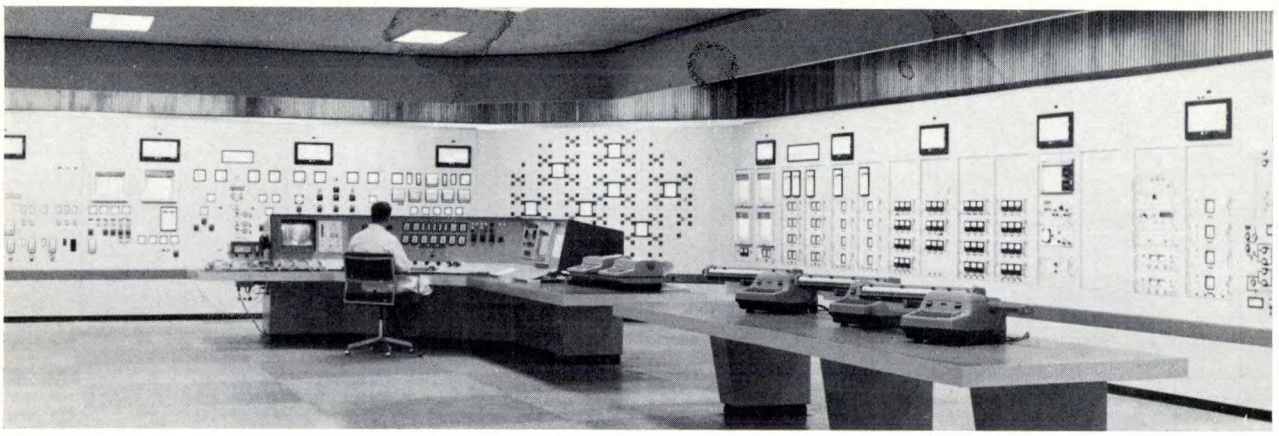
This hypothesis is supported by a number of facts. For example, SiC polyhedra, grown in the same furnace but with their *c*-axis practically perpendicular to the wall, contain no higher types. This orientation of the crystals apparently excludes the memory effect. Only 6H is formed and perhaps also 15R. Complicated structures are again found however in polyhedra that are intergrown with one another, as can be seen from the occurrence of re-entrant angles. Here one intergrown part apparently forms the memory of the other one and *vice versa*. In one case we have established a quantitative relation [20] between the structure type formed (33R) and the angle α at which two halves of a polyhedron were intergrown (*fig. 14a*). The ratio of the crystallographic *a*- and *c*-axes was exactly equal to α (expressed in radians), strongly suggesting the existence of a relation between the manner of intergrowth and the type of structure (fig. 14*b*).

Evidence for the existence of such a relation has also been found in investigations with the related CdI_2, which can be prepared in different ways: the method yielding most crystal intergrowths also leads to the formation of the largest number of different structures [21].

It seems to us, therefore, that this is where the key to the solution of the problem of polytypism is hidden, at any rate for the more complicated structures. However, the whole story will not be known until we have also understood the mechanism underlying their formation on an atomic scale.

Summary. Several theories have been put forward to explain the phenomenon of polytypism as it occurs in silicon carbide. Some of the principal theories are discussed. A view common to all of these is that chance factors (lattice defects) present in the crystal itself are responsible for the growth of the often very complicated type of structure. None of these theories are entirely satisfactory. Closer consideration shows that a small number of simple structures can grow under very different conditions. This makes the mechanism of their growth more difficult to explain than that of the more complicated structures which, although fortuitous in origin, are nevertheless grown under better defined conditions. Experiments have provided clear evidence in support of the hypothesis that the growth of these more complicated structures can be caused by external influences, in particular by the way in which they are intergrown with other individual crystals.

[17] Kuo Chang-lin, Scientia sinica **15**, 604, 1966.
[18] G. S. Zhdanov and Z. V. Minervina, J. Phys. USSR **9**, 244, 1945.
[19] W. F. Knippenberg and A. H. Gomes de Mesquita, Z. Kristallogr. **121**, 67, 1965.
[20] A. H. Gomes de Mesquita, J. Crystal Growth **3,4**, 747, 1968.
[21] W. Kleber and P. Fricke, Z. phys. Chemie (Leipzig) **224**, 353, 1963.



Monitoring the control-rod operation of the nuclear reactor at Dodewaard (Netherlands)

The first nuclear power station in the Netherlands, with an output of 54 MW, has recently been put into service at Dodewaard. Uranium dioxide, packed in tubular cans, is used as the fuel for the reactor of this station, and water is used as a moderator and heat-transfer agent. The reactor works at a temperature of 286 °C and a pressure of 70 bars. The steam generated in the reactor is used to drive a turbine.

The operation of the station requires many electronic aids: there are numerous measuring devices, some with recording systems, a communication system which includes television cameras set up at strategic points, radiation meters to ensure the safety of the staff, and various computer systems for processing the measurement data coming in from the many sub-units of the station. Much of this equipment is set up in the control room, part of which is shown in the photograph of *fig. 1*.

The maintenance of a controlled chain reaction requires that the reactivity k must be held exactly to a value of unity. What this means is that a neutron liberated in a nuclear fission liberates on average exactly one other neutron by the fission of another uranium nucleus. In this as in many other reactors this is achieved by the use of *control rods* of neutron-absorbing material at certain locations. These control rods (there are 37 of them) can be set to a greater or lesser insertion between the tubular fuel elements. Since the density of the boiling water (and hence its reactivity) is greatest at the bottom of the reactor, the best control is achieved when the control rods are inserted from underneath.

One of the computer systems mentioned above has the function of monitoring the control-rod operation and preventing incorrect adjustments by the reactor operator. This article gives a few details of this system, whose control gear with the actual computer unit was made by Philips.

An essential for economical running of the reactor is that the control rods should be adjusted, not all together as a unit, but individually. This should be done in such a way that the heat generated is distributed as uniformly as possible over the reactor and also so that each fuel element is expended as far as possible when the charge has to be renewed. Knowing the power required and the situation in the reactor, the operator makes these adjustments by changing the positions of the control rods one by one in accordance with a particular programme until one of a number of previously calculated and tested patterns of settings is reached [1]. The procedure is chiefly of interest in the starting-up period, in which the reaction is made critical ($k = 1$) step by step by pulling out control rods. Once the reactor is running at full power a critical state of equilibrium is reached which requires very little adjustment of the control rods.

The control-rod computer checks the upward and downward movement of each rod separately, thus ensuring that the rods keep to the desired pattern of settings, which is stored in the computer. An important additional condition here is that the technician is never allowed to depart so far from the prescribed procedure that the number of neutrons absorbed by one of the

control rods exceeds a certain value, the permissible "control-rod worth". This precaution prevents the possibility of the reactor going out of control if a fault causes one of the rods to fall out. With this in mind, there is also a "scram" system which can slam all the control rods back into the reactor simultaneously, thus stopping the chain reaction.

In fig. 1 the control panel of the control-rod computer can be seen to the reactor operator's left. The wall behind the desk is the "core mimic display" with instruments which give information about the control rods and the reactivity at various locations in the reactor core: the location of the indications and instruments on the core display corresponds to the actual locations of the 156 groups of fuel elements in the reactor (each group contains 36 fuel elements). The control rods are cross-shaped in section and each rod is surrounded by four groups of fuel elements; see *fig. 2a*. Fig. 2*b* gives a more detailed photograph of the core display; the display at the centre of the picture gives information about the corresponding control rod. The readings of this display and the indicator lights show the position and movement of the control rod. This information is provided by 23 position-sensing contacts mounted in the guide channel of each rod. The operation of the control-rod motors can be observed at any time by means of a television camera.

The computer is of the type PR 8000, with a memory store of 8192 words of 24 bits and the appropriate peripheral equipment. The program takes up 4700 words, and another 2700 words are required for the measurement data.

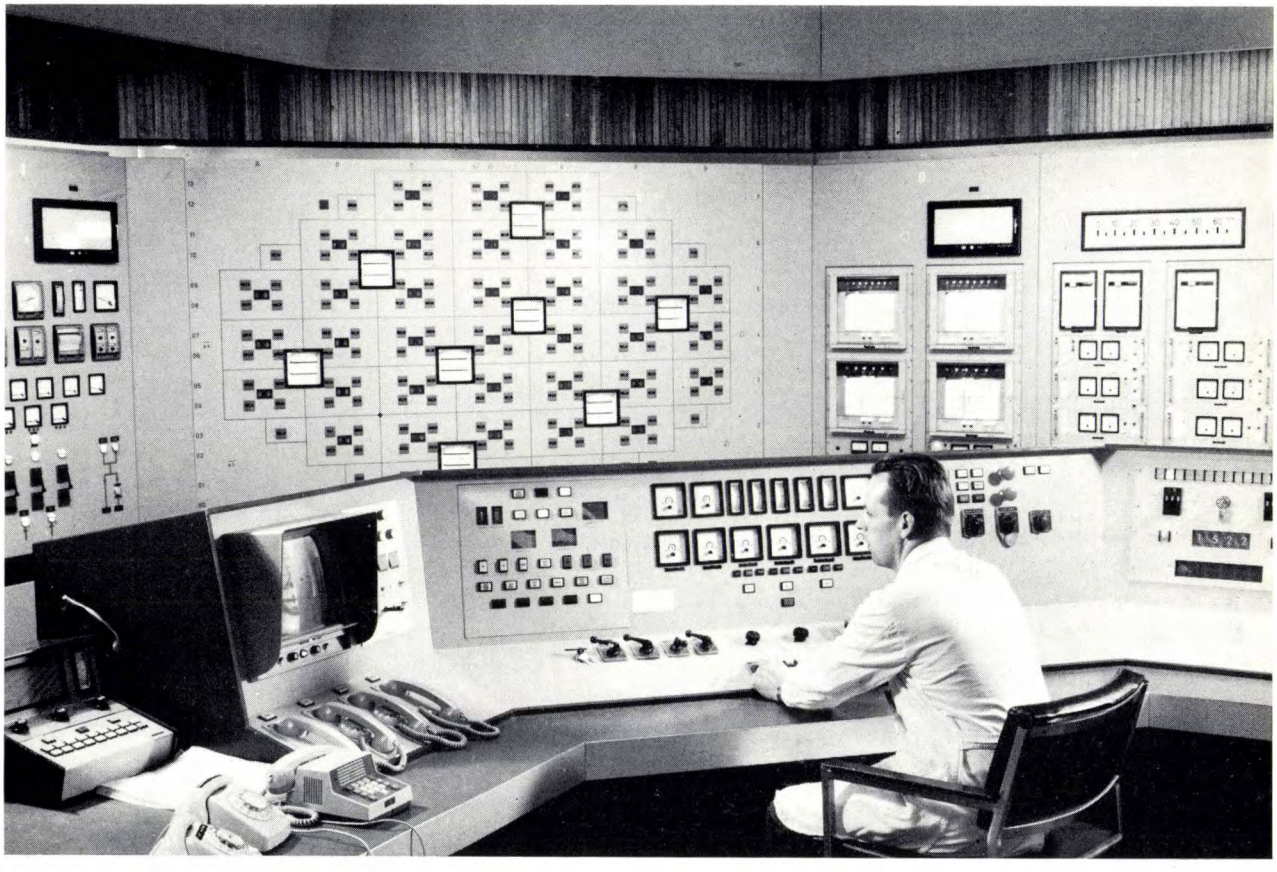


Fig. 1. Part of the control room of the nuclear power station. The "core display", located behind the operating-desk, indicates the position and movement of each control rod by means of 37 counters and signal lamps, which are arranged in the same pattern as the fuel elements and control rods in the reactor.

The program entered in the store is divided into a number of subsidiary programs:

1) *Starting program* for the computer procedure that limits the reactivity worth (initial setting-up).
2) *Scan programs* for collecting information about the positions of the rods, the rod to be moved by the operator, the direction of movement of the rods, the rod drift, etc. The computer has an input system

[1] J. C. Bruggink, Ingenieur **80**, O 128, 1968 (No. 37).

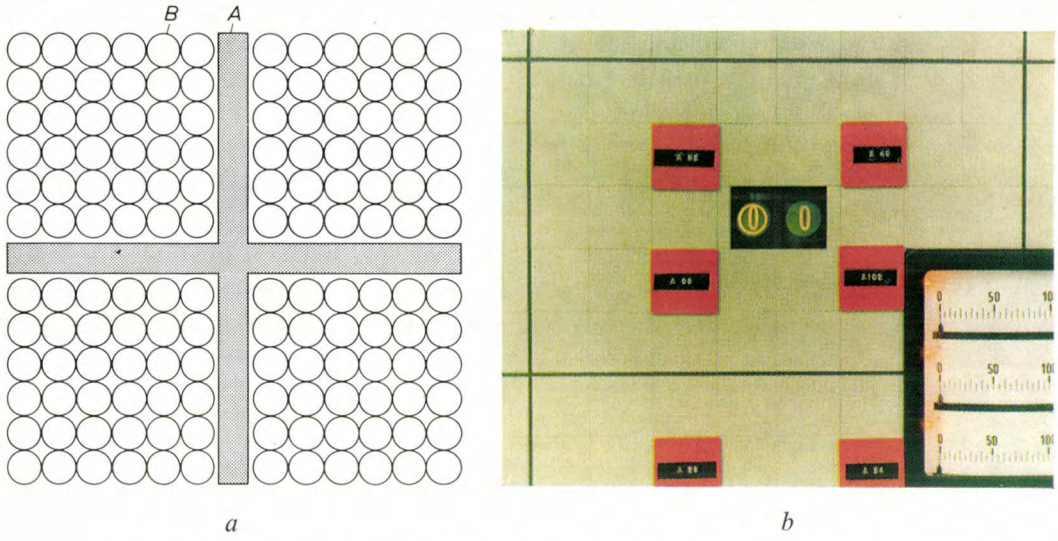


Fig. 2. *a*) Diagram showing how a control rod (*A*) is located centrally between four groups of fuel elements. The control rod, which has a section shaped like a cross, consists of stainless-steel tubes filled with boron carbide; the complete assembly is enclosed by plates of stainless steel. *B* is one of the fuel elements (dia. 13 mm); each element consists of a tube filled with uranium dioxide.
b) Enlarged view of part of the core display. The small central window in each panel of the core display shows information about the corresponding control rod. The information is given as a two-digit decimal number registering the distance by which the rod is withdrawn. When one of the control rods has been selected for adjustment the corresponding panel lights up, and the final settings are also indicated by coloured indicator lights. The four square red indicators relate to the fuel elements grouped around each control rod, and the meters show the local reactivity at corresponding points in the reactor core.

which scans 800 binary signal indicators in each 125 ms. The output unit supplies digital information to the operating panel and also to interlock circuits in the drive system for the control rods: the interlock circuits prevent incorrect operation. Where necessary the reliability is increased by redundancy in the signal information.

3) *Information transfer* between the operating panel and the drive systems of the control rods.
4) *Typewriter print-out* of information about the control-rod limiting system, reactor power, sequence of operations in adjusting the rods, and of a "topographical" survey of the positions of the rods (corresponding to the indications on the core display), plus records of faults, with a time scale, and rod drift.
5) *Indications* on the display and operating panel of rod chosen, sequence of operations in use, operational faults, system faults and corrections.
6) *Computer checking* by means of a test program which examines parity (check digits), encoding, processing of instructions, access to the store, and mains supply voltage. If the supply voltage fails an emergency program comes into operation; this picks out the information being transferred and puts it into the store, which is then protected from the effects of a failure in the units which control it.

H. van der Kooi

Ir. H. van der Kooi is with the Philips Industrial Equipment Division (PIT), Eindhoven.

A two-stage compressor with rolling diaphragm seals

H. J. Verbeek

The current development of small piston-driven compressors has been strongly influenced by the need in certain applications for oil-free compression — used for some years in large industrial compressors. In the two-stage compressor described here the gas and oil are kept separate by means of rolling diaphragm seals.

In various applications, such as the liquefaction of gases, it is important that oil from the compressor should contaminate the gas as little as possible. On this account, much of the recent development work on piston-driven compressors has been concentrated on reducing, or better completely stopping, the leakage of oil along the piston to the gas to be compressed. Oil leakage can be greatly reduced by the use of labyrinth-seal pistons or pistons with rings of special plastic, both of which can work without lubrication. In compressors equipped with such pistons the crankcase is not directly connected to the cylinder. The partition wall does however have an opening for the connecting-rod, provided with a packing or with scraper rings, and this of course cannot be perfectly leakproof.

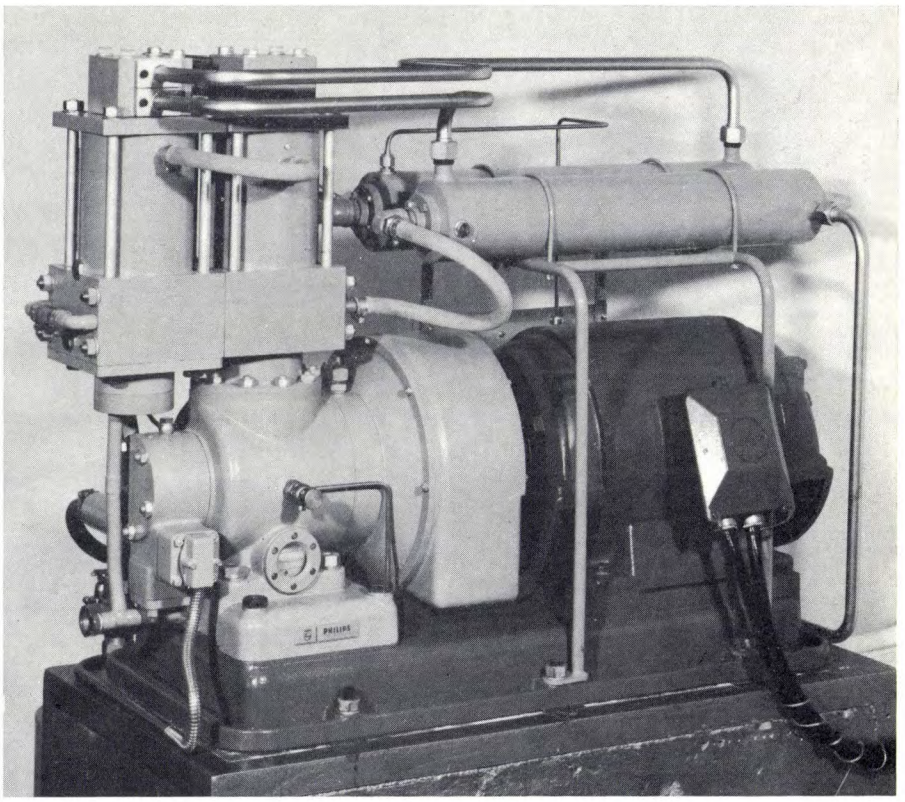


Fig. 1. Two-stage compressor with rolling diaphragm seals. Intake pressure about 2½ bars, output pressure about 20 bars. Gas flow when compressing helium is about 2½ g/s. Speed 960 r.p.m. Maximum shaft power 8 kW. Cooling water consumption about 500 l/h.

A complete separation between the space containing the gas and the other spaces in the compressor can be obtained by using a diaphragm; this also prevents the escape of gas, which is another requirement frequently imposed in the applications referred to. A disadvantage of the flat diaphragms of metal or plastic used in some compressors, however, is their relatively short life. A very much longer life can be obtained from the rolling diaphragm seals [1] developed at Philips Research Laboratories a few years ago. These dia-

Ir. H. J. Verbeek is with the Philips Industrial Equipment Division (PIT), Eindhoven.

[1] J. A. Rietdijk, H. C. J. van Beukering, H. H. M. van der Aa and R. J. Meijer, A positive rod or piston seal for large pressure differences, Philips tech. Rev. **26**, 287-296, 1965. An experimental compressor using rolling diaphragm seals is discussed in this article.

phragms can be rolled up and down more than two thousand million times. The compressor discussed below (*fig. 1*), which was developed for the Philips helium liquefier [2], is fitted with diaphragms of this kind. The compressor compresses the helium in the liquefier in two stages from about $2\frac{1}{2}$ to about 20 bars, the gas flow rate being about $2\frac{1}{2}$ g/s. It can also be used for the compression of other inert gases and of hydrogen.

Fig. 2 shows a schematic cross-section of the compressor. The piston P_1 in the cylinder on the right is driven by a motor (not included in the drawing) via a

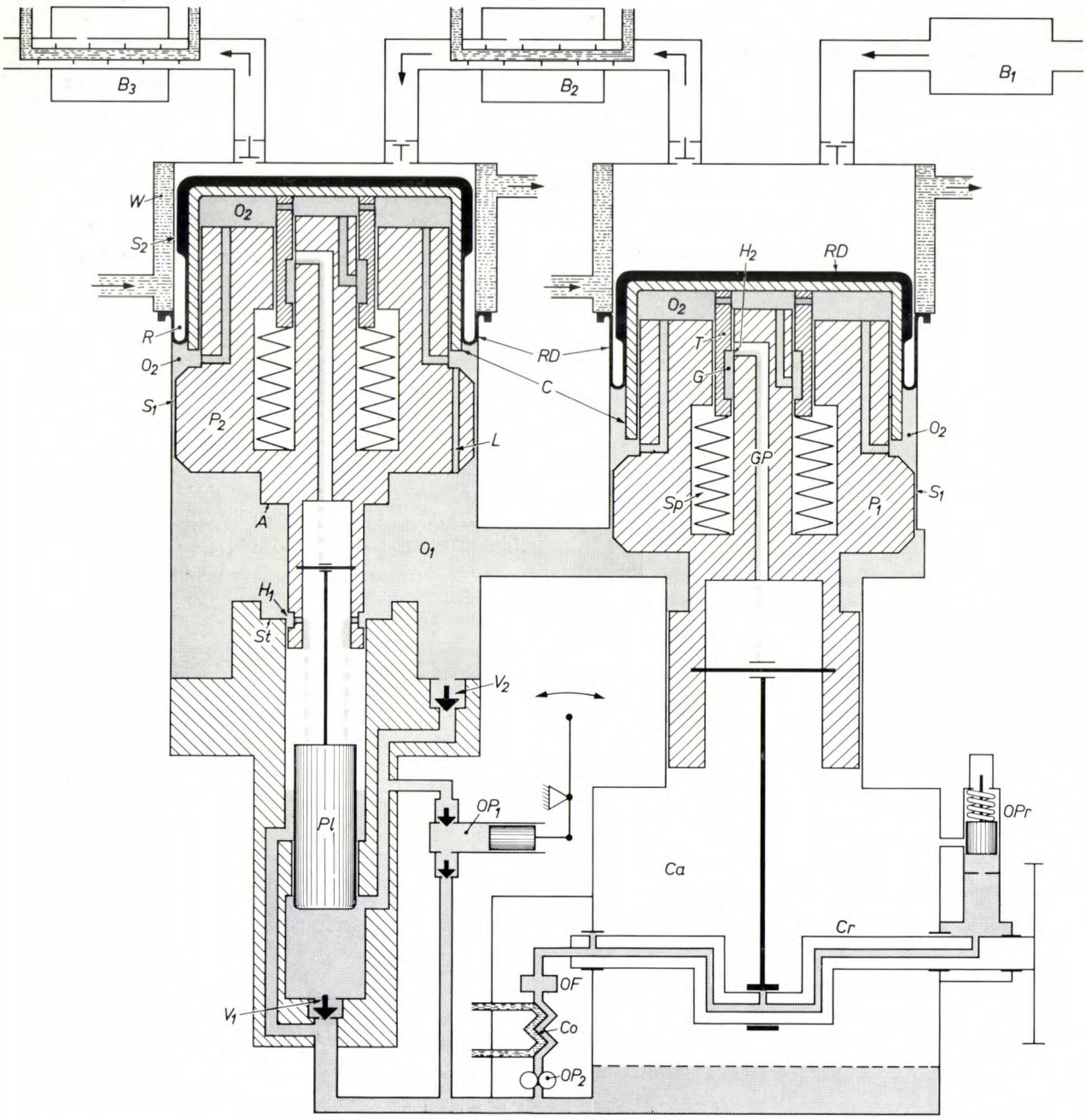


Fig. 2. Simplified schematic cross-section of the two-stage compressor with rolling diaphragm seals. *Ca* crankcase. *Cr* crankshaft. P_1 and P_2 pistons (diameter 110 mm, stroke 60 and 25 mm respectively). O_1 and O_2 oil-filled spaces. *Pl* pump plunger. V_1 suction valve. V_2 compression valve. H_1 overflow port which determines the highest position of P_2. OP_1 priming pump. *RD* rolling diaphragm. *C* piston cap. *T* thick-walled tube carrying *C*; the tube *T* slides around the guide pin *GP* and rests on a stack of copper spring washers *Sp*. S_1 gap between piston and cylinder wall. S_2 gap between rolling diaphragm and cylinder wall. *W* water-cooling jacket. *R* diaphragm cavity. *A* lower end of P_2 which comes to rest on the stop *St* when the compressor is stationary and sufficient oil has leaked away from O_1. B_1 buffer vessel in the suction line. B_2 and B_3 buffer vessels in the intermediate and compression lines, each with water cooling (above the motor in fig. 1). OP_2 oil pump (gear type), *Co* cooler, *OF* oil filter and *OPr* oil pressure regulator, which together provide the lubrication for the crankshaft.

crank and connecting-rod mechanism. The piston P_1 sets P_2 in motion by hydraulic action. Since the upper part of P_1 has a larger diameter than the lower part, at every downward stroke this piston displaces some of the oil in space O_1 (shaded), so that P_2 is pushed upwards, and *vice versa* [3]. Each piston is hermetically sealed with a rolling diaphragm *RD*. The rolling diaphragm seals used here are not ring-shaped but form a closed surface which completely covers the head of the piston. The cap of the seal is much thicker than the skirt, which performs the rolling action; there are several advantages in this kind of design: it is less likely to spring a leak after slight damage, it has good heat insulation and is easier to make.

Under the second cylinder there is an oil pump, formed by the pump plunger *Pl* and the valves V_1 and V_2. When the compressor is in operation, this pumps oil continuously from the crankcase *Ca* to O_1, pumping slightly more than leaks away from O_1. In every upward stroke of P_2 the outlet port H_1 opens at a certain instant, allowing the surplus pumped oil to flow back to *Ca*; the position of H_1 determines the highest point that P_2 can reach.

When the compressor is stationary, piston P_2 will slowly sink, because of the leakage of oil from O_1, until its surface *A* comes to rest on the stop *St*. Before the compressor can be started up, enough oil must therefore be pumped into O_1 to bring P_2 to its highest point — P_1 being at its lowest position when the compressor is stationary. This is done by means of the priming pump OP_1.

As fig. 1 shows, the rolling diaphragms do not rest on the piston but on a cap *C* which slides around the piston and rests on a thick-walled tube *T*. This tube is carried by a stack of cupped spring washers *Sp* and slides around the guide pin *GP* located in the cavity of the piston. The spaces O_2 under *RD* and *C* are filled with oil. The whole arrangement acts as a regulating system which ensures that the oil pressure in the spaces O_2 is always lower than the gas pressure in the cylinder above the piston by a constant amount ($1\frac{1}{2}$ bars), which is necessary for rolling diaphragm seals of this type [1].

This regulating system is identical for both pistons and works as follows. The oil leaking along the cylinder from space O_1 to space O_2 cannot push the cap any higher than the position at which the groove *G* reaches the opening H_2 of the duct running through *GP*, i.e. the duct through which oil can flow back to the crankcase from O_2. The difference between the oil pressure in O_2 and the gas pressure in the cylinder therefore always corresponds to the force which the spring exerts on the cap at the position at which *G* reaches H_2.

The regulating systems of the two rolling diaphragm seals do not of course operate under identical conditions. The oil pressure in O_1 only differs slightly from the gas pressure in the second cylinder, but differs considerably from the gas pressure in the first cylinder. Consequently the difference in oil pressure between O_1 and O_2 is also very much smaller for the second cylinder than for the first, so much so that the oil leak through S_1 is not sufficient for the proper operation of the regulating system in the second cylinder. Piston P_2 is therefore also provided with a leakage duct *L*.

The rapid to-and-fro motion of the pistons when the compressor is running naturally causes fairly large inertial forces on the piston caps. To prevent these forces from causing fluctuations in the pressure difference across the diaphragms, steps have been taken to compensate for them.

In the present design this was done by choosing a suitable value for the spring constant of the spring washers. When the piston rises, the curved part of the diaphragm does not go up as far (*fig. 3*). As a result, part of the oil filling O_2 is displaced and causes the piston cap to rise slightly with respect to the piston. In this higher position the spring *Sp* exerts a smaller upward force on the cap. The effect of the inertial force on the pressure difference across the diaphragm is

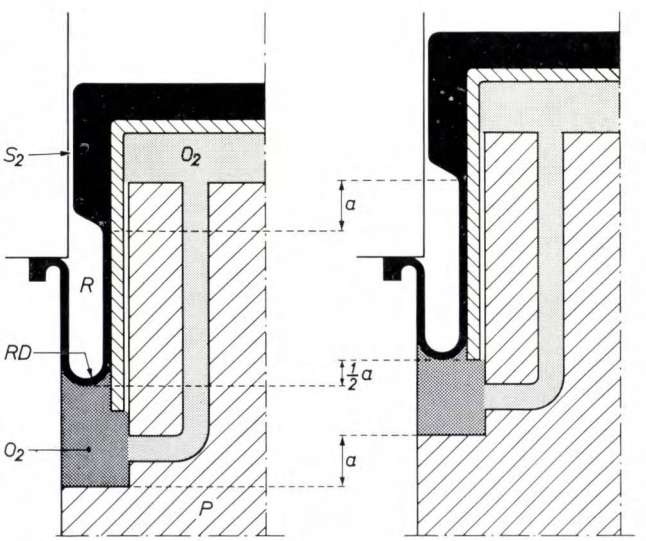


Fig. 3. Schematic cross-section of part of the piston *P* and of the rolling diaphragm *RD*, which separates the oil-filled space O_2 from the gas-filled space *R*, the rolling-diaphragm cavity. When the piston makes an upward stroke *a*, the curved part of *RD* — the base of *R* — is only displaced by $\frac{1}{2}a$, so that the volume of *R* increases and that of the heavily shaded part of O_2 decreases.

[2] G. J. Haarhuis, The Philips helium liquefier, Philips tech. Rev. **29**, 197-204, 1968 (No. 7).

[3] This type of drive was first used in the Philips type C gas refrigerating machine. See A. A. Dros, An industrial gas refrigerating machine with hydraulic piston drive, Philips tech. Rev. **26**, 297-308, 1965.

cancelled out when this reduction of the spring force at the lowest position is equal to the inertial force.

The temperature of the gas in the cylinders fluctuates considerably about a mean value of about 100 °C. The thin, rolling part of the diaphragms cannot however come into contact with gas at this high temperature. To reach the rolling part of the diaphragm the gas must first pass through the narrow gap S_2, where it is cooled down by the local water cooling in the cylinder wall. Moreover, in the space R, the diaphragm cavity, a small quantity of cold is generated. It can be shown that this effect arises because the variation of the volume of R with time is exactly out of phase with the variation of the volume of the space above the piston, since the upper wall of R is formed by the thick part of the diaphragm, which describes the same stroke as the piston (fig. 3).

The diaphragms are made of polyurethane rubber. The use of special processes for purifying the rubber and for pressing the diaphragms ensures that the diaphragms meet the very strict requirements for purity and dimensional accuracy. They have a working life of at least 2000 hours, but they will generally operate for even longer than 4000 hours. Changing a diaphragm only takes about 30 minutes.

The crankcase Ca is hermetically sealed. The pressure in it, under normal conditions, is about 1 bar. If one of the rolling diaphragms should start to leak, the gas to be compressed would flow into Ca, causing the pressure in it to rise. In such an event a pressure-sensitive switch stops the compressor. The crankcase is designed to be able to withstand the high pressure which would appear inside it if the diaphragm failed.

Above each of the cylinders there is a double-acting valve unit which can quickly be replaced if necessary. The two valves in each unit are in the form of concentric rings. The valve units are identical for both cylinders.

The principles described above can also be applied to the design of compressors with more than two stages, giving a higher output pressure, or with a different intake pressure, e.g. 1 bar or lower.

Summary. The compressor was developed for the Philips helium liquefier, in which it is used to compress flowing helium gas (2½ g/s) from about 2½ to 20 bars in two stages. The first piston is motor-driven via a crankshaft, and the second is driven hydraulically (with oil as the medium) by the first. The stroke of the second piston is set by an overflow device, the oil that leaks away being continuously pumped back. The compression spaces are sealed with rolling diaphragms of polyurethane rubber, made by a special process, which completely cover the heads of the pistons. The diaphragms are not however stretched over the pistons themselves but rest on a piston cap which has oil beneath it; the position of the cap with respect to the piston is determined by an overflow device. The difference between the oil pressure and the gas pressure in the cylinder is constant and is determined by the force which a spring exerts on the cap at the highest position of the cap. The pressure difference is not affected by the inertial forces acting on the piston cap. The gas coming into contact with the rolling part of the diaphragms is much colder than the gas in the other parts of the cylinder. The same principles can be used in designing compressors with more than two cylinders and a different intake pressure.

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

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.

F. Berz: A generalized integral formula for two-dimensional fields.
J. appl. Phys. **39**, 4853-4854, 1968 (No. 10). *M*

G. Blasse: Energy transfer in oxidic phosphors.
Physics Letters **28A**, 444-445, 1968 (No. 6). *E*

B. Bölger & J. C. Diels: Photon echoes in Cs vapour.
Physics Letters **28A**, 401-402, 1968 (No. 6). *E*

E. A. Boonzajer Flaes (Philips Philite and Metalware Manufacturing Division, Eindhoven): Untersuchung über die Verwendbarkeit organischer Farbstoffe zum Einfärben von anodisiertem Aluminium.
Galvanotechnik **59**, 955-962, 1968 (No. 12).

G. A. Bootsma & F. Meyer: Ellipsometric investigation of physisorption at low temperatures.
Surface Sci. **13**, 110-118, 1969 (No. 1). *E*

A. Broese van Groenou, P. F. Bongers & A. L. Stuyts: Magnetism, microstructure and crystal chemistry of spinel ferrites.
Mat. Sci. Engng. **3**, 317-392, 1968/69 (No. 6). *E*

T. D. Clark: Experiments on coupled Josephson junctions.
Physics Letters **27A**, 585-586, 1968 (No. 9). *M*

T. D. Clark & D. R. Tilley (Physics Dept., University of Essex, Colchester): Resistive transition tailing in an array of Josephson junctions.
Physics Letters **28A**, 62-63, 1968 (No. 1). *M*

U. Dibbern: Zur Ausbreitung des Gleitfunkenkanals.
Z. angew. Physik **25**, 336-341, 1968 (No. 6). *H*

G. Diemer & J. H. Emck (Technical University, Eindhoven): The education of physicists for work in industry.
Proc. int. Seminar, Eindhoven 1968, publ. Centrex, Eindhoven, 55 pp. *E*

F. C. M. Driessens: Thermodynamics and defect chemistry of some oxide solid solutions: I. Nearest-neighbour interactions and the effect of substitutional disorder, II. Pair interactions.
Berichte Bunsenges. phys. Chemie **72**, 754-764, 764-773, 1968 (No. 7). *E*

G. Engelsma: Photoinduction of phenylalanine deaminase in gherkin seedlings, III. Effects of excision and irradiation on enzyme development in hypocotyl segments.
Planta **82**, 355-368, 1968 (No. 4). *E*

B. Frank & R. Groth: Herstellung und Eigenschaften von Schichten aus Y_2O_3 und Oxiden der Seltenen Erden auf Glas.
Thin Solid Films **3**, 41-50, 1969 (No. 1). *A*

Z. van Gelder: New high-intensity spectral source with a narrow line profile.
Appl. Spectroscopy **22**, 581-582, 1968 (No. 5, Part 1). *E*

C. A. A. J. Greebe: Use of minority carriers for ultrasonic amplification in piezoelectric semiconductors.
Physics Letters **28A**, 455-456, 1968 (No. 6). *E*

G. Groh: Vacuum deposition of thin films by means of a CO_2 laser.
J. appl. Phys. **39**, 5804-5805, 1968 (No. 12). *H*

R. Groth & J. Liebertz: Dichroismus von Cr^{3+}-dotierten $LiNbO_3$-Einkristallen.
Naturwiss. **55**, 540-541, 1968 (No. 11). *A*

E. E. Himmelbauer (Philips Electronic Components and Materials Division, Eindhoven): Design considerations for a wide-band oscilloscope tube with scan magnification.
Thesis, Groningen 1968.

K. Hoselitz: Educating physicists for work in industry.
Nature **220**, 1280-1282, 1968 (No. 5174). *M*

J. Kaashoek: A study of magnetic-deflection errors.
Thesis, Eindhoven 1968. *E*

Y. Kamp & J. Neirynck: Cascade synthesis of a multivariable transfer matrix.
Electronics Letters **4**, 369-370, 1968 (No. 17). *B*

D. Le Coq, C. Piaget, G. Piétri & G.-A. Boutry: Obtention de cibles pour tubes analyseurs d'images à partir de lames minces monocristallines.
C.R. Acad. Sci. Paris **267B**, 635-638, 1968 (No. 14). *L*

D. Le Coq, G. Piétri & G.-A. Boutry: Réalisation par alliage de mosaïques de jonctions sur silicium pour l'utilisation dans des tubes analyseurs d'images.
C.R. Acad. Sci. Paris **267B**, 601-604, 1968 (No. 13). *L*

W. Lems, P. J. Rijnierse, P. F. Bongers & U. Enz: Photomagnetic effect in a chalcogenide spinel.
Phys. Rev. Letters **21**, 1643-1645, 1968 (No. 24). *E*

G. Marie: Nouveau tube relais optique pour projection d'images de télévision.
Onde électr. **48**, 925-929, 1968 (No. 499). *L*

A. van Oostrom: Dependence of the critical field strength for vacuum breakdown of tungsten on the field emitting area.
Les décharges et l'isolement électrique dans le vide, Comptes-rendus IIIe Symp. Int., Paris 1968, p. 174-180. *E*

L. J. van der Pauw: Fourpole properties of electroacoustic delay lines with conventional and unconventional transducers.
Thesis, Delft 1968. *E*

P. Reijnen: The formation of ferrites from the metal oxides.
Science of Ceramics **3**, 245-261, 1967. *E*

J. F. Schouten (Institute for Perception Research, Eindhoven): Weergave en presentatie in rekenen en regelen, 1. Wat verlangt de mens van weergave, presentatie en hanteerbaarheid?
Ingenieur **81**, O 1-3, 1969 (No. 1).

Ch. G. Sluijter: An analog circuit simulating and explaining the triggering process of an electronic flash-tube.
High-Speed Photography, Proc. 8th int. Congress, Stockholm 1968, p. 211-214. *E*

M. J. Sparnaay: The donor and acceptor character of surface states on germanium as affected by the adsorption of carbon monoxide.
Surface Sci. **13**, 99-109, 1969 (No. 1). *E*

M. J. Sparnaay: The free energy of double layer systems, with an emphasis on semiconductor surfaces and on a membrane system.
Surface Sci. **13**, 222-237, 1969 (No. 1). *E*

F. L. Stumpers: Radiostoringen, I. De internationale samenwerking in het radiostoringsonderzoek en de radiostoringsbestrijding.
Ingenieur **80**, ET 147-149, 1968 (No. 42). *E*

C. G. Venis: Eenvoudige methoden om zelf kleine aantallen artikelen van rubber te maken uit gevulcaniseerde latex.
Polytechn. T. Werktuigbouw **24**, 17-27, 1969 (No. 1). *E*

Contents of Electronic Applications **28**, No. 4, 1968:

Contents of Mullard Technical Communications **10**, No. 97, 1969:

Contents of Valvo Berichte **14**, No. 5, 1968:

PHILIPS TECHNICAL REVIEW

VOLUME 30, 1969, No. 3

A step-and-repeat camera for making photomasks for integrated circuits

F. T. Klostermann

In the current technique for making monolithic integrated circuits a large number of identical electronic circuits, each consisting of a number of circuit elements, are formed upon a silicon wafer perhaps 50 mm in diameter. This technique has only been made possible by the most extensive refinements in mechanical and chemical processing. A particularly important part is played by the photographic step-and-repeat process which has to meet some exceptionally difficult accuracy requirements. The cameras which Philips have developed for this process can produce photomasks with an accuracy of 0.2 μm or better.

Introduction: technology of monolithic integrated circuits

Monolithic integrated circuits consist of very small but complicated structures, with even smaller detail and tolerances, which are formed in a silicon crystal. The desired configurations are obtained by using a refined photographic technique. The basic steps in the process are shown in *fig. 1* [1]. A coating of photosensitive lacquer (photoresist) is deposited on a wafer of silicon oxidized at its surface. A microphotographic negative, the photomask, is placed on the wafer and the coating of photoresist is exposed to the light of a mercury lamp. The exposed areas of photoresist are dissolved in a liquid developer (with another type of photoresist it is the unexposed parts which are dissolved). At the places no longer covered by photoresist the layer of SiO_2 can be dissolved in a selective etching bath, and the remaining photoresist is then removed. This photo-etching process is followed by a diffusion process which gives the desired doping to the areas of silicon from which the oxide has been removed by etching; a new layer of SiO_2 is then formed. These processes, from the application of the photoresist to the formation of a new layer of SiO_2, have to be repeated a number of times with different masks and different types of diffusion, as

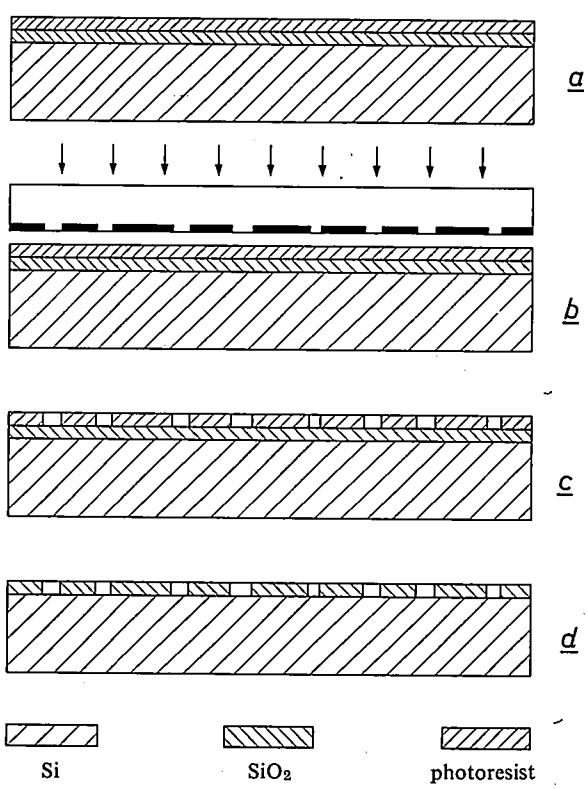


Fig. 1. Photographic process in the manufacture of integrated circuits.
a) Silicon substrate with SiO_2 layer and coating of photoresist.
b) Exposure in contact with a photomask.
c) The exposed parts of the photoresist are dissolved.
d) The SiO_2 layer is etched away at the places where the photoresist has been dissolved, then the remaining photoresist is dissolved.

Dr. F. T. Klostermann is with Philips Research Laboratories, Eindhoven.

[1] For a more detailed description see A. Schmitz, Solid circuits, Philips tech. Rev. 27, 192-199, 1966.

required for the circuit to be produced. Finally, a "wiring" of deposited aluminium is formed by using similar processes with what is called the aluminium or wiring mask.

In this article we shall not go into the electrical characteristics required of the circuits or the silicon-doping processes involved, but we shall look instead at the mechanics of the photographic process, with particular attention to the manufacture of the photomasks used in this process.

details of this size on the photoresist on the silico wafer (fig. 1*b*).

b) In the succession of processes that are applied i making an integrated circuit, *windows* have to b made at each step in newly formed layers of SiO These windows should match each other very clos ly, often with tolerances of not more than sa $\pm \frac{1}{2}$ µm. *Fig. 2* shows a circuit together with th set of photomask patterns (in this case five) whic were used successively in making it.

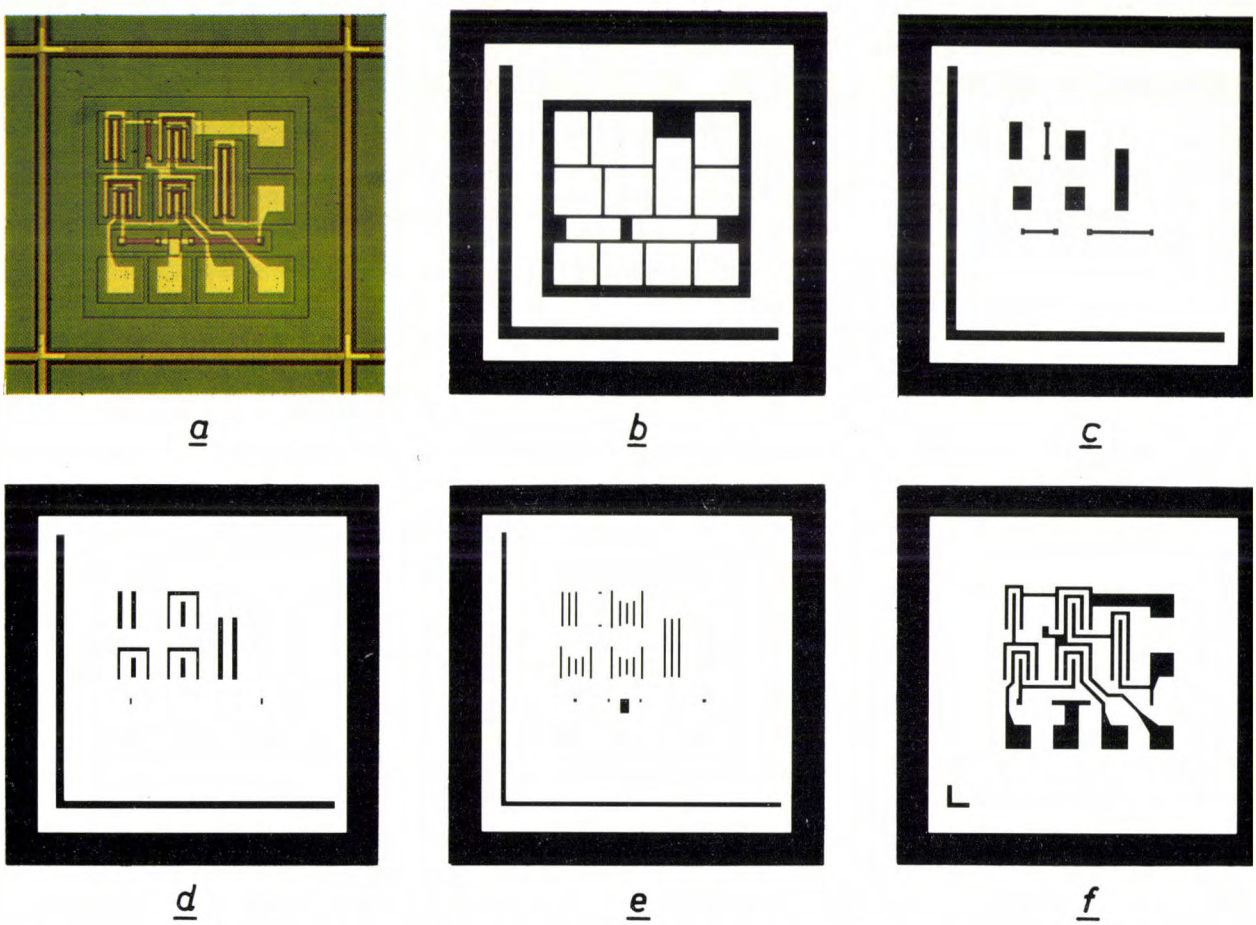


Fig. 2. *a*) An integrated circuit enlarged about 40×. *b-f*) Patterns for the photomasks used to produce circuit (*a*).

Accuracy requirements for photomasks

The requirements which the photomasks have to meet are determined by three special features of the technique employed:

a) The dimensions of elements in an integrated circuit or in a transistor for high frequencies have to be extremely small (down to 1 µm in extreme cases). The patterns on the photomask have to be sharp enough to give accurate and reliable copying of

c) A large number of circuits are formed simultan ously on a single silicon wafer. Each of the phot masks therefore has to be made up from a lar number of regularly spaced identical patterns; the are 100 to 10 000 patterns on a mask. *Fig. 3* sho a multiple photomask of this type for one of t mask patterns shown in fig. 2 (the wiring mask).

Each of the multiple photomasks in a set has to copied in contact with the coating of photoresist on t

silicon wafer in such a way that *all* the successive patterns register with each other over the *entire* surface within the $\pm \frac{1}{2}$ μm tolerance. This is only possible if the mask patterns are very accurately located in the same way in each of the photomasks. In placing the successive photomasks on the silicon wafer, errors in positioning are unavoidable and also contribute to the final error in registration. The error in the position of any one pattern in the multiple mask must therefore be only part of the total tolerance, say $\pm \frac{1}{4}$ μm. This should be compared with the size of the complete mask, which is 50 000 μm.

Limitations imposed by image quality

The degree of sharpness of the pattern which can be achieved in the microphotographic negative is limited by aberrations and diffraction in the optical image, by light scattering during exposure, and by chemical diffusion during the development of the photographic plate. If visible light and a very fine-grained silver-

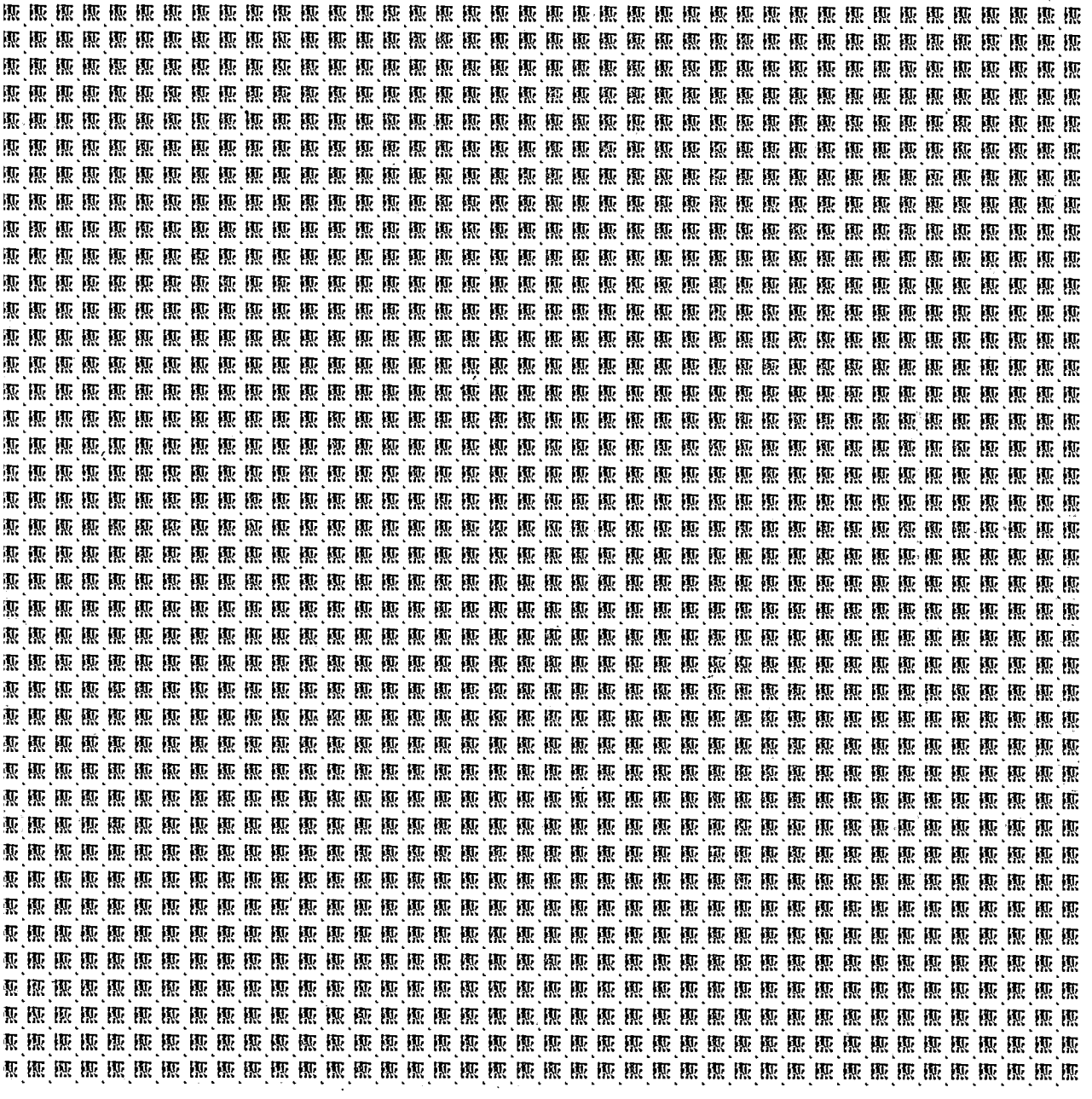


Fig. 3. Multiple photomask, enlarged 3×, of the pattern shown in fig. 2*f* ("aluminium" mask used to apply the wiring).

halide plate (Lippmann plate) are used, lines 1 μm wide can be obtained with just sufficient sharpness.

Ultra-violet light has the advantage that it suffers less diffraction than visible light. The Rayleigh scattering of light in Lippmann plates, however, increases with decreasing wavelength and with blue or ultra-violet light less sharp photographic negatives are obtained than with green light [2]. If the pattern is obtained with ultra-violet light directed at a non-scatter or reduced-scatter photographic material (such as a photoresist coating [3] or PD material [4]), lines of width less than 1 μm can be obtained which are sharp enough for practical application. In this case the limit is probably somewhere near 0.5 μm.

Much sharper images can be obtained with the aid of electron optics. Experiments using electron beam techniques have yielded particularly good results [5]. For example, conductor patterns have been produced whose line width was less than 0.1 μm. A limitation with electron beam techniques is that details in a negative which are too fine to be obtained with light optics cannot — for this very reason — be reproduced satisfactorily in the light-optical copying process mentioned above (fig. 1*b*). This makes it necessary to work without a negative and to expose each silicon wafer using the electron-optical equipment. The method is consequently very expensive, especially since the image field in which very good image quality is obtained is no larger than a few square millimetres, so that each silicon wafer has to be exposed in a great many places in turn. However, this method has other advantages besides the improved sharpness of image, so that further developments along these lines seem likely.

Making masks photographically by the step-and-repeat technique

One of the fundamental problems of the technology we have described is that of making the multiple photomasks to the required accuracy. In one ingenious method a pinhole array camera with a large number of uniformly spaced holes is used, giving a reduced image of the same mask pattern behind each hole [6]. Unfortunately, the image quality which can be obtained is not good enough. In the "fly's-eye" camera proposed by IBM each hole is replaced by a simple moulded plastic lens. The image quality achieved by this means is acceptable for certain applications [7].

However, the most important and almost universally employed method of producing photomasks is the step-and-repeat technique. With this technique the multiple photomask is obtained by optically recording the mask pattern not simultaneously but *successively*, on a reduced scale, at a number of positions on a photographic plate. The chief advantage of this method is that the "step-and-repeat camera" used for this process can be provided with a very good optical system whose image field has been matched to the size of the mask patterns. Achieving the very high register accuracy which is required for the multiple photomasks of a set now depends on the accuracy of the displacement system of the step-and-repeat camera.

Before discussing this matter in further detail let us briefly mention some important features of the step-and-repeat technique [8].

To begin with, mask patterns are usually prepared at a scale of say 200 : 1, with the aid of a numerically controlled drawing-machine [9]. These patterns are then reduced in two stages: the first reduction, which could typically be 10×, provides the object plates for the step-and-repeat camera, and the second reduction, of 20×, is carried out by the step-and-repeat camera itself [10]. In this second reduction, the plates used are of the silver-halide type used in microphotography, like the Kodak High Resolution plate. The plate can be exposed "in motion" in the step-and-repeat camera by using a xenon flash lamp; the photographic plate moves uniformly in its own plane and the flash lamp is fired whenever a position is reached at which an image of the pattern must appear. This position is determined by means of a measuring system. The duration of the flash from the lamp can be so short (e.g. 10 μs) that the movement causes no loss of sharpness at quite practical rates of advance of the photographic plate (e.g. 1 mm/s). The operation of a step-and-repeat camera with "in motion" exposure is shown schematically in *fig. 4*. The carriage *S* carrying the photographic plate moves uniformly on a slideway and this permits a row of very accurately located patterns to be obtained relatively easily and quickly. The starting point of each successive row has to be set each time (manually or by servo) by displacing a second carriage which carries the projection system and moves at right angles to *S*.

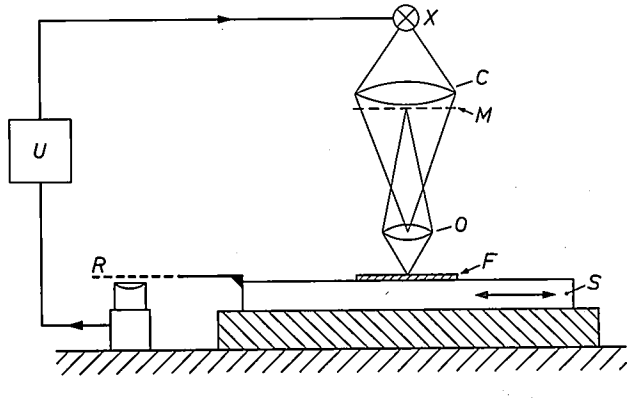


Fig. 4. Step-and-repeat camera with "in motion" exposure for generating a multiple photomask. *M* object plate with a mask pattern, positioned in a projection system with xenon flash lamp *X*, condenserlens *C* and objective *O*. The photographic plate *F* is on a carriage *S* which moves at uniform speed. The measuring system using a grating *R* determines the distance covered and gives this information to the control and supply circuit *U* which fires the flash lamp at the correct instants. The projection system is placed on a second carriage, similar to *S* and moving at right angles to *S*, whose displacement is determined with a second measuring system.

To increase the production capacity a *multiple* step-and-repeat camera (a "multi-barrel camera") is generally employed. The principle shown diagrammatically in fig. 4 remains unchanged and no drastic modifications need be made to the slide systems, the measuring system or the control circuit. A number of projection systems — six, for example — rigidly fastened to each other, and with their flash lamps fired by the same pulse, are placed side by side on one carriage and six photographic plates are fitted to the other, which moves at right angles to the first (*fig. 5*). Six different object

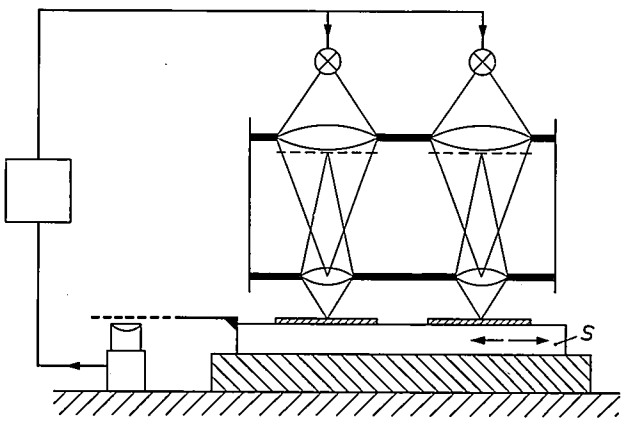


Fig. 5. Multiple step-and-repeat camera with "in motion" exposure. Two rows of three photographic plates are mounted on carriage *S*, each plate being exposed by its own projection system. The two rows of three projection systems are themselves mounted on a second carriage which moves at right angles to *S*.

plates, each showing one of the mask patterns required for a set of photomasks, are inserted in the projection systems. In this way all the photomasks required for an integrated circuit (usually not more than six to ten) can be prepared in one or two runs.

With the multi-barrel camera, it is essential that both simultaneously and successively made photomasks should match each other exactly. The most critical factors here are:

a) The accuracy of the slide systems.
b) The accuracy of the measuring system.
c) The positioning of the object plates.
d) The maintenance of focus and the correct reduction factors.
e) Temperature control.

An equipment that satisfies current (and even more stringent) requirements in all these respects has been developed in our Laboratories: under the least favourable conditions of use it gives a total error from all these sources which is no greater than $\pm$ 0.2 μm. The various critical factors will be considered in the discussion of the equipment which follows.

The Philips step-and-repeat camera

The slide systems

The ultimate accuracy of the photomasks depends to a very large extent on the slide systems of the step-and-repeat camera. The two most important requirements which the slide systems must satisfy are that the random lateral deviations (perpendicular to the direction of travel) must be extremely small and that rotation about the optical axes of the camera must be kept to an absolute minimum. Lateral deviations give rise to errors in photomasks which are made successively. Rotation of the carriages about the optical axes causes errors in photomasks that are made simultaneously by the different projection systems. If the distance between the optical axes is d and there is a rotation of $\Delta\alpha$, relative errors of $d\Delta\alpha$ are introduced.

A slide system with hydrostatic bearings which is ideally suited for the step-and-repeat technique has been developed (for other purposes) in these Laboratories by De Gast [11]. The design of a slide system of this

[2] G. W. W. Stevens, Microphotography; photography and photofabrication at extreme resolution, Chapman and Hall, London 1968, p. 27.
[3] H. J. Schuetze and K. E. Hennings, Large-area masking with patterns of micron and submicron element size, Semicond. Prod. Solid State Technol. **9**, No. 7, 31-35, 1966.
[4] H. Jonker, C. J. Dippel, H. J. Houtman, C. J. G. F. Janssen and L. K. H. van Beek, Physical development recording systems, I. General survey and photochemical principles, Phot. Sci. Engng. **13**, 1-8, 1969 (No. 1); parts II, III, IV, V to appear shortly.
[5] T. H. P. Chang and W. C. Nixon, Electron beam formation of 800 Å wide aluminium lines, J. sci. Instr. **44**, 231-234, 1967.
[6] J. J. Murray and R. E. Maurer, Arrays of microphotographs for microelectronic components, Semicond. Prod. **5**, No. 2, 30-32, 1962.
P. A. Newman and V. E. Rible, Pinhole array camera for integrated circuits, Appl. Optics **5**, 1225-1228, 1966.
[7] W. E. Rudge, W. E. Harding and W. E. Mutter, Fly's-eye lens technique for generating semiconductor device fabrication masks, IBM J. Res. Devel. **7**, 146-150, 1963. — An interesting new method that promises high image quality is the holographic method of simultaneous image multiplication recently proposed by G. Groh of the Philips laboratory at Hamburg (Appl. Optics **7**, 1643-1644, 1968), now being further investigated there.
[8] A. J. O'Malley, The ABC's of photomasking, Semicond. Prod. Solid State Technol. **7**, No. 11, 27-32, 1964.
[9] R. Ch. van Ommering and G. C. M. Schoenaker, The "COBRA", a small digital computer for numerical control of machine tools, Philips tech. Rev. **27**, 285-297, 1966. A recent further development of this work is described by C. Niessen and H. E. J. Wulms, Automatic drawing of masks for integrated circuits, Philips tech. Rev. **30**, 29-34, 1969.
[10] More direct methods for the first step have been developed and further development is expected: in these methods the object plates for the step-and-repeat camera are obtained directly by numerically controlled exposure of the photographic plate. See for example: H. Freitag, Generating IC masks automatically, Electronics **40**, No. 18, 88-92, 1967.
[11] 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; published in Advances in machine tool design and research 1966 (Proc. 7th Int. M.T.D.R. Conf., Univ. Birmingham), Pergamon Press, Oxford 1967, pp. 273-298.
See also an article on a precision lathe, to appear shortly in this journal.

type is shown in *fig. 6*. The hydrostatic bearings are situated at H_1, H_2 and H_3. Oil delivered under high pressure (20 to 40 bars) by an external pump system passes first through hydraulic restrictors, with resistances R_a and R_b and then through the restrictors R_a' and R_b' formed by the narrow bearing gap (about 30 μm). This gives pressures between R_a and R_a', and also between R_b and R_b', which depend on the ratio of these resistances. A downward force on the carriage causes the gap height and hence also the intermediate

granite (diabase) slab with retaining bolts (fig. 6). The granite slab is flat to within 3 μm over the whole of its upper surface (400×600 mm). The design shown has been chosen because it enables the tendency of the carriage to rotate about the optical axes (the z-axis in fig. 6) to be effectively suppressed. In fact, any rotation about the z-axis would be caused by the guide surfaces of the bearings H_1, i.e. the opposing surfaces of bars B_2 and B_3. The average deviation of these surfaces from the straight is measured while the bars are being mounted.

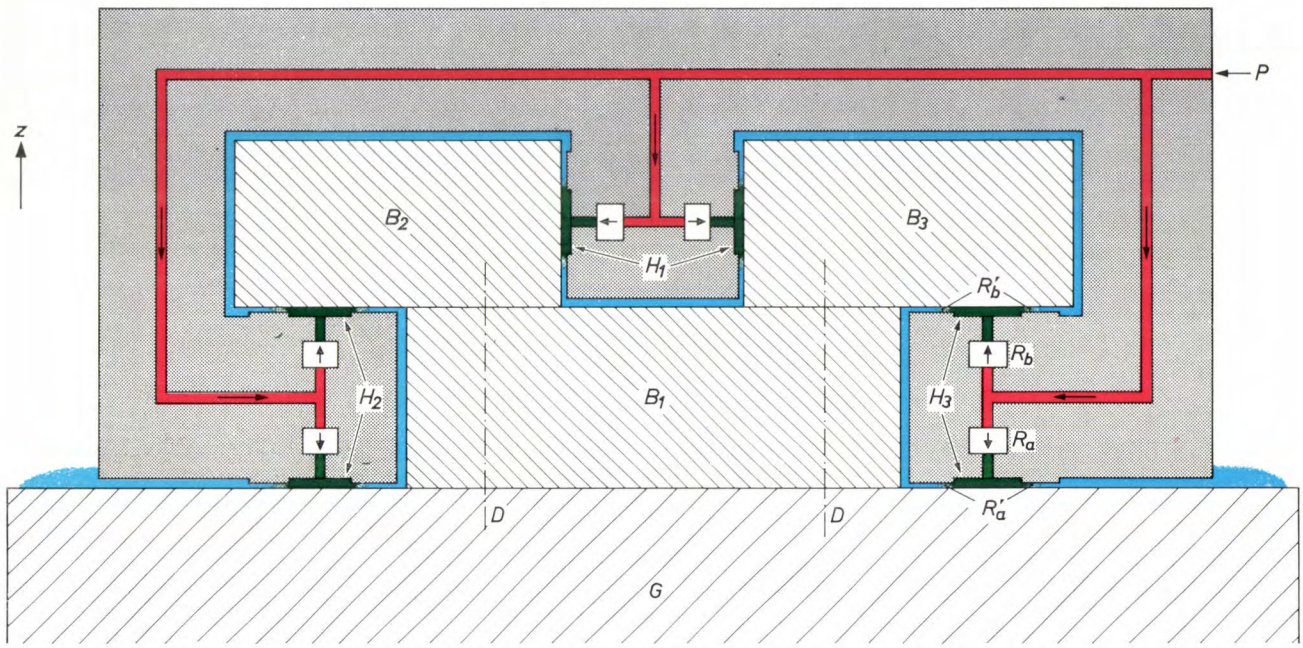


Fig. 6. Schematic diagram of the slide system. The carriage is shown in grey; it moves at right angles to the plane of the drawing. G granite baseplate; B_1, B_2 and B_3 steel bars. D retaining bolts; H_1, H_2 and H_3 hydrostatic bearings. Oil is pumped in at P; oil at high pressure is shown in red, oil at low pressure in blue (intermediate pressure in green). R_a and R_b restrictors; R_a' and R_b' bearing gaps.

pressures to change, so that a new equilibrium can be found. The slide systems designed by De Gast differ from other systems with hydrostatic bearings in that the high hydrostatic forces (approximately 10^4 newtons) are very evenly balanced. When a carriage moves, changes occur in the forces acting on the slideway. Small unavoidable asymmetries in the elastic deformation of the slideway then adversely affect the accuracy of travel. In our case however such forces are balanced, and because of this there is hardly any elastic deformation even with a carriage and slideway of relatively light construction.

The slideway for each of the two slide systems consists of three steel bars fastened to a 160 mm thick

Any deviations measured can then be corrected by making small elastic deformations in the bars B_2 and B_3 and the corrected arrangement is made permanent with the aid of the retaining bolts.

The reproducibility of the paths traced out by these carriages was found by measuring the capacitance between a small metal disc mounted on the moving carriage and a fixed metal bar that had been very accurately lapped. A typical result is shown in *fig. 7*. The lateral deviations of the carriage do not exceed ± 0.03 μm.

The rotation of a carriage about the z-axis while it moved horizontally was measured with the aid of a photoelectric autocollimator telescope (measuring

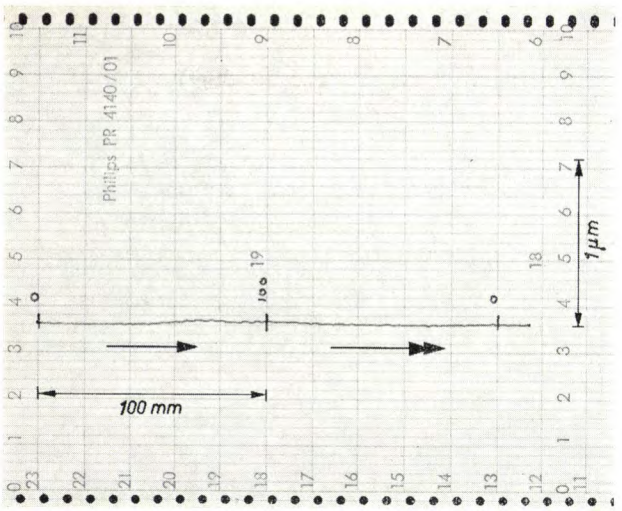


Fig. 7. Trace showing lateral deviations from the correct path of a carriage with hydrostatic bearings. The carriage has been moved 100 mm and back. The deviations are smaller than $\pm$ 0.03 mm.

accuracy 0.05 of a second of arc). A typical result is shown in *fig. 8*. The rotation about the z-axis for a travel of 50 mm does not exceed $\Delta\alpha = \pm 0.1'' = \pm 5\times10^{-7}$ rad. This has to be substituted in the formula $d\Delta\alpha$ for the resultant error. The distance required between adjacent optical axes of a multiple step-and-repeat camera depends chiefly on the size of the photomasks required. We have chosen 60 mm for this distance in one case and 80 mm in others. The largest distance, d, in our system is therefore 160 mm.

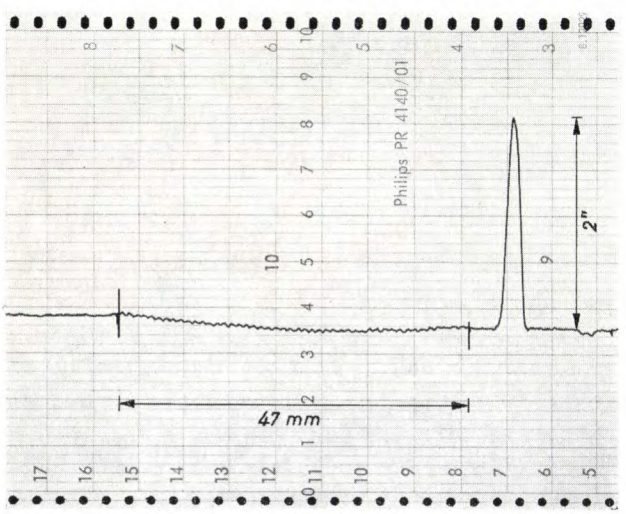


Fig. 8. Trace showing the rotation about a vertical axis of a carriage with hydrostatic bearings. Carriage movement 47 mm. Rotations are smaller than $\pm$ 0.1 second of arc. This trace was made after the slide system had been used nearly every day for 18 months.

The magnitude of the worst-case error introduced by the rotation of *both* carriages is therefore:

$$\Delta x = \pm 2\times16\times10^{4}\times5\times10^{-7} = \pm 0.16\ \mu\text{m}.$$

Fig. 9 shows the slide system for the six-barrel step-and-repeat camera before mounting. The carriage on which the six photographic plates are mounted can be seen at the front; the carriage at the back carries the holder for the six projection systems and moves at right angles to the first carriage. The two slide systems are mounted independently of each other on the granite slab. The dimensions of the clamping surface of each

Fig. 9. The slide systems of a six-barrel step-and-repeat camera. The carriage on which the photographic plates are mounted can be seen at the front, and the carriage at the back carries the projection systems on an arm, here already fitted.

slideway are 200 $\times$ 180 mm. The photograph in *fig. 10* shows the complete equipment, i.e. the actual step-and-repeat camera, and the ancillary electronic equipment which controls the movements of the carriages, the firing of the flash lamps, etc.

The measuring system

Indirect methods of measuring carriage travel, such as the combination of an angle-measuring system and a lead-screw, are much used in engineering workshops. However, the indirect methods are not very suitable in our case, since such methods are barely capable of the high linear measuring accuracy required. For the accuracy that we require, a direct measurement of dis-

placement, by optical means, is by far the most suitable. Two such systems have been developed in these Laboratories by De Lang and others; both systems can be considered for the step-and-repeat camera. One is a grating measurement system, and the other a laser-interferometer system [12]. Of the two, the former depends on the barometric pressure, and it introduces a measuring unit that is basically foreign to the metric system. The small systematic errors that can occur with a measuring grating (perhaps $\pm$ 0.2 μm) are no disadvantage in our case where we are chiefly concerned with the *relative* registration of the photomasks.

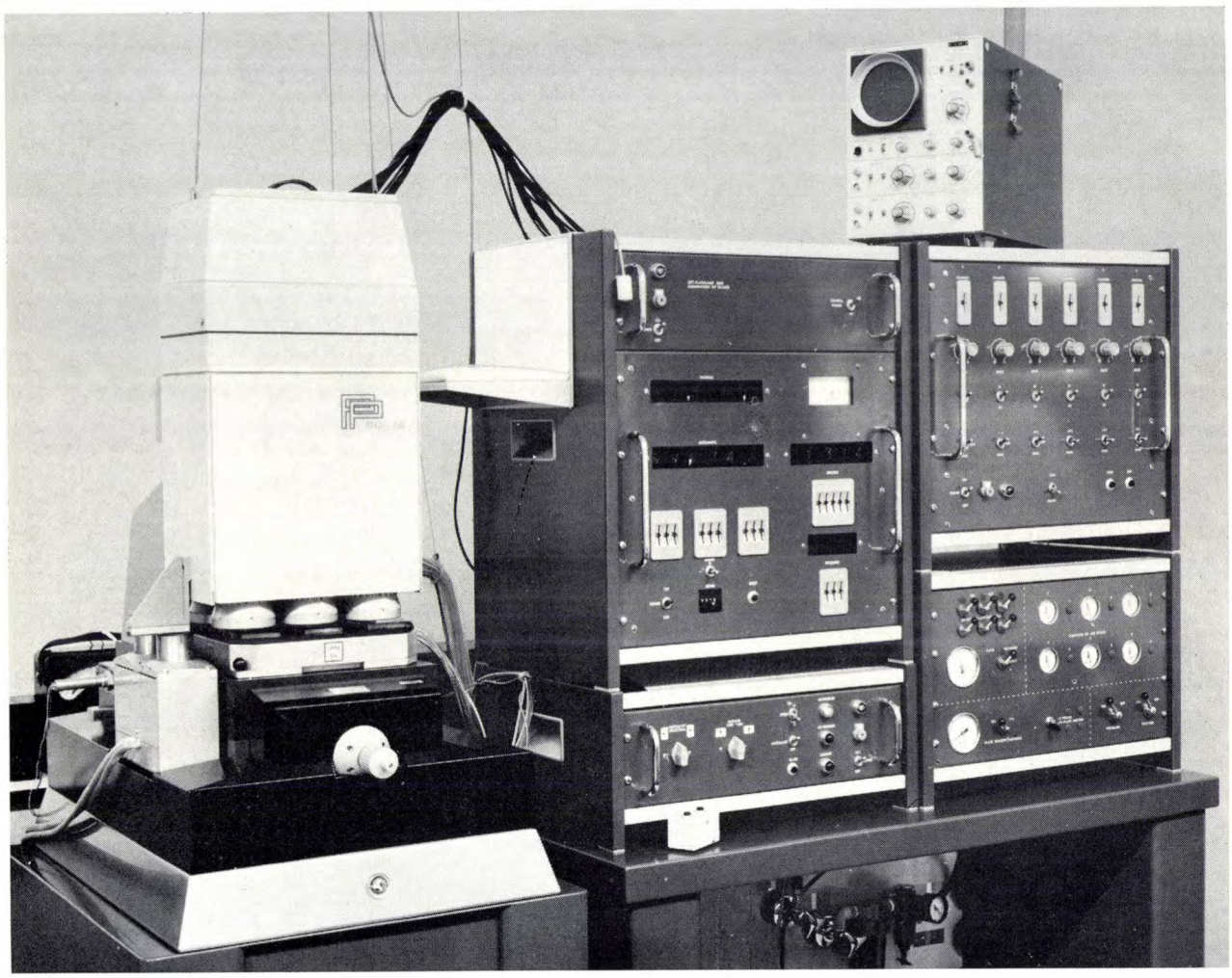


Fig. 10. The complete six-barrel step-and-repeat camera. The panels on the right contain the electronic control and supply circuits, the reducing valves for pressurizing the air bearings, and other accessories.

appears more suitable for the step-and-repeat technique because the thermal expansion coefficient of the measuring grating can be matched to that of the glass photographic plate by appropriate choice of the material of the grating. The actual temperature level of the step-and-repeat camera is then far less critical, which is quite the reverse of the situation with a laser interferometer, whose thermal expansion coefficient is some 10 times smaller than that of most types of glass. Moreover, measurement with a laser interferometer

For a detailed description of the grating measurement system, which is being used with very good results in our step-and-repeat camera, the reader is referred to the publication already noted under [12]. A detail which should be mentioned here is that measurement with this system makes use of a mirror vibrating at a frequency f. When the system is used dynamically, as when flash lamps are fired during the uniform movement of the carriage, the frequency f and the velocity v of the carriage together impose a limit on the repetition

accuracy that can be obtained by the measurement. This is because the uncertainty Δt of the instant of firing is $\pm 1/2f$, and the uncertainty in the location of the images is therefore $\Delta x = \Delta t \times v = \pm v/2f$. With $v = 1$ mm/s and $f = 4$ kHz, we obtain $\Delta x = \pm 1/8$ μm.

Fig. 11 is a close-up view of the measuring head and measuring grating assemblies in the six-barrel step-and-repeat camera. The measuring heads, which are firmly fixed to the granite base plate, read off the gratings from below.

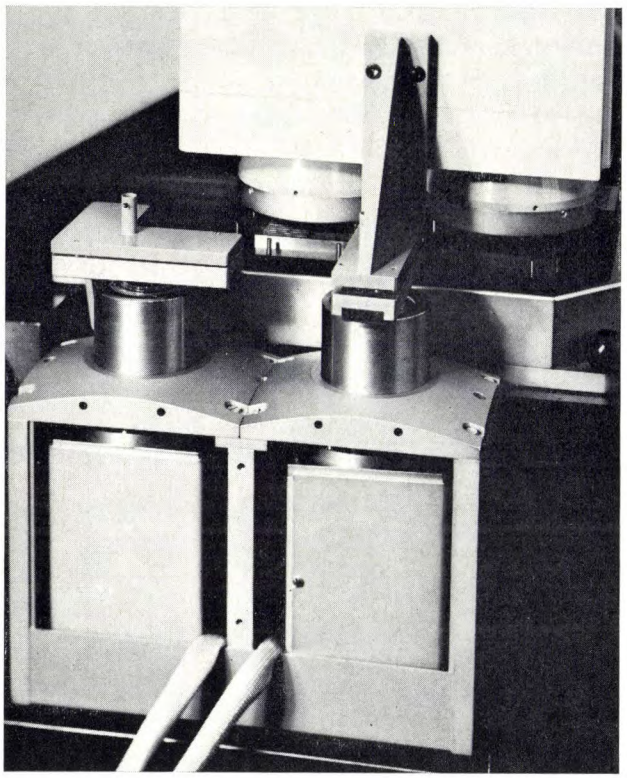


Fig. 11. The measuring system. The measuring head and the grating on the left are for the carriage on which the photographic plates are mounted; the measuring head and grating on the right are for the carriage carrying the projection systems.

Positioning of the object plates

The patterns $a_1, b_1, c_1 \ldots$ of a set of multiple photomasks can only yield accurately registering prints if they all make the same angle with the directions of travel in the step-and-repeat process. This condition can be satisfied by positioning the object plates for patterns $a_1, b_1, c_1 \ldots$ in the projection system very accurately to the same angle. Sometimes a pattern $a_1, b_1, c_1 \ldots$ has to be left out at certain positions on the photomask, and another pattern $a_2, b_2, c_2 \ldots$ has to be put in at these positions (for example for putting in special test components). If this is the case a second condition is added: the object plates for patterns $a_1, b_1, c_1 \ldots$ and $a_2, b_2, c_2 \ldots$ must also be positioned very accurately to the same coordinates.

This positioning adjustment can be performed by the following simple and reliable method. Each object plate has two optical fiducial marks and the top plate of the projection column (the housing of the projection systems) on which the object plates have to be successively placed, has "enclosing" lines which fit round the marks (*fig. 12*). The two marks are 25 mm apart. A double

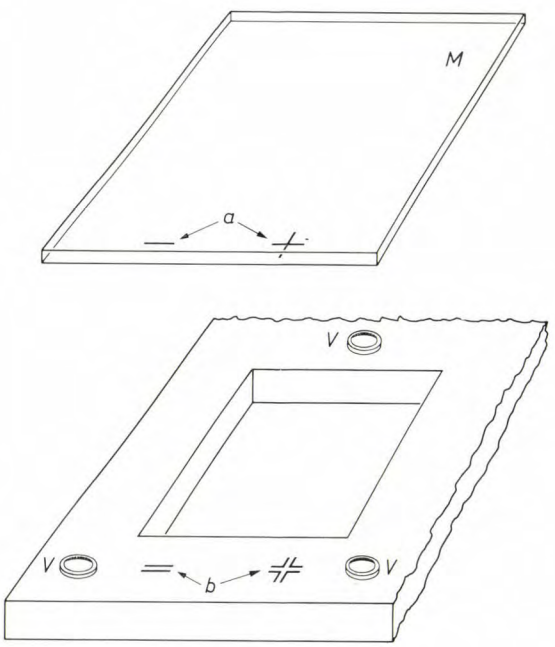


Fig. 12. Each object plate M is adjusted with the aid of two fiducial marks a on the plate. These have to be positioned accurately between the enclosing lines b on the top plate of the projection column. V supports for the object plate which is clamped to them by vacuum suction after adjustment.

microscope with a magnification of 30× enables both marks to be observed simultaneously (*fig. 13*). Firmly attached to the microscope is an x-y-α micromanipulator which can be used to move the object plate so that the marks fit between the enclosing lines. The combined adjustment microscope and micromanipulator can be positioned over each of the projection systems by means of a subsidiary slide system and held there temporarily by pistons powered with compressed air.

The positioning of a black line centrally between two black enclosing lines is in fact checked visually by comparing the light intensity of the strips left between

[12] Both systems will shortly be described in this journal.

the lines; the eye can detect an extremely small asymmetry between the two intensities. With a suitable configuration of the enclosing lines [13] it is possible in this way to detect a deviation as small as $\Delta x = \Delta y = \pm 0.3$ μm when the magnification is only 30×. A setting accuracy of ± 1 μm is easily and quickly ensured with the micromanipulator.

One of the reasons why this method is so reliable is that the accuracy of the subsidiary slide system does not affect the result.

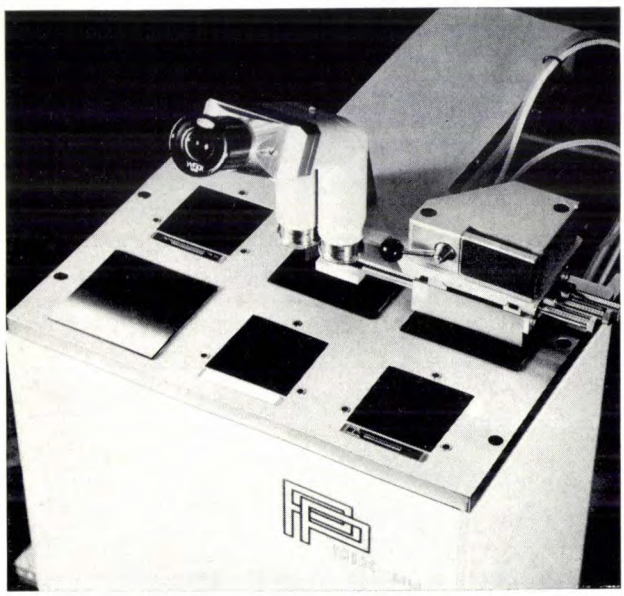


Fig. 13. The adjustment microscope with the x-y-α micromanipulator to the right of it.

Focusing and reduction factor

Each of the six optical systems of a step-and-repeat camera has to project an image of its object plate on each of the successive desired positions of the appropriate photographic plate. In this operation the optical system must remain accurately focused on the surface of that plate. Any defocusing affects not only the sharpness but also the size of the image.

Contrary to what might be expected, the effect on image size is in some cases more serious than the loss of sharpness. This should be clear from the following estimate of the tolerances of the two effects. To obtain good image quality in microphotography a fairly large numerical aperture value N is necessary, e.g. $N = 0.2$ to 0.4. The depth of focus, given by $\Delta b = \pm 0.6\,\lambda/N^2$, and hence the amount of defocusing allowed if adequate image sharpness is to be retained, is consequently very small. Using the above values for N and a wavelength of $\lambda = 0.5$ μm we find $(\Delta b)_{max} = \pm 8$ to 2 μm. On the other hand, a variation Δb in the image distance b causes an image point at a distance r from the optical axis to be shifted by Δr, and this variation of the image size means that projected images will be out of register. As a first approximation $\Delta r/r = \Delta b/b$. (More correctly, b should be replaced by the distance between image plane and exit pupil.) In the optical systems used in practice b/r_{max} often has a value between 10 and 20. If we assume Δr, the tolerance on r, to be $\frac{1}{8}$ or $\frac{1}{4}$ μm, we see that in some (in fact important) cases the defocusing Δb allowed giving the maximum allowable change in image size will only be about ± 2 μm. If this requirement is met, the depth-of-focus requirement given above is automatically satisfied.

The most important cause of defocusing is out-of-flatness of the photographic plate. Kodak High Resolution plates are generally used, in the versions known as "selected flat" and "ultra flat" (the "micro flat" version is not very suitable because of the thick glass). Over a length of 25 mm the guaranteed tolerance on flatness is ± 25 μm for "selected flat" plates and ± 10 μm for "ultra flat" plates. The length we are concerned with, however, is considerably larger than 25 mm, since the size of the field with mask patterns is about 50 mm in multiple photomasks for integrated circuits. Since other factors besides flatness, e.g. placing the photographic plate against stops, contribute to the total defocusing Δb, the tolerance calculated above is far exceeded even with "ultra flat" plates.

This difficulty has been overcome by continuously correcting the image distance b while the photographic plate is moving, using the local surface of the plate itself as reference level [14]. We have achieved this in our camera by mounting each objective in diaphragm springs and allowing it to rest on an air bearing over the part of the photographic plate which is to be exposed (*fig. 14*). With a supply pressure of 2 bars and a nominal gap height of 20 μm in the air bearing, the gap height varies by 1 μm per 0.1 bar change in the supply pressure. The gap height in this case can easily be kept constant within ± 1 μm. This air-bearing focusing arrangement can be seen in *fig. 15*. The photographic plates have to be brought up to the air bearings from underneath. To load the camera, the plate holder is swung out, loaded with plates, swung under the air bearings in the lowest position and automatically adjusted to the correct height by three compressed-air pistons.

Temperature control

The principle of the step-and-repeat camera is that the plate (or group of six plates) is displaced by increments — longitudinal for the continuous movement

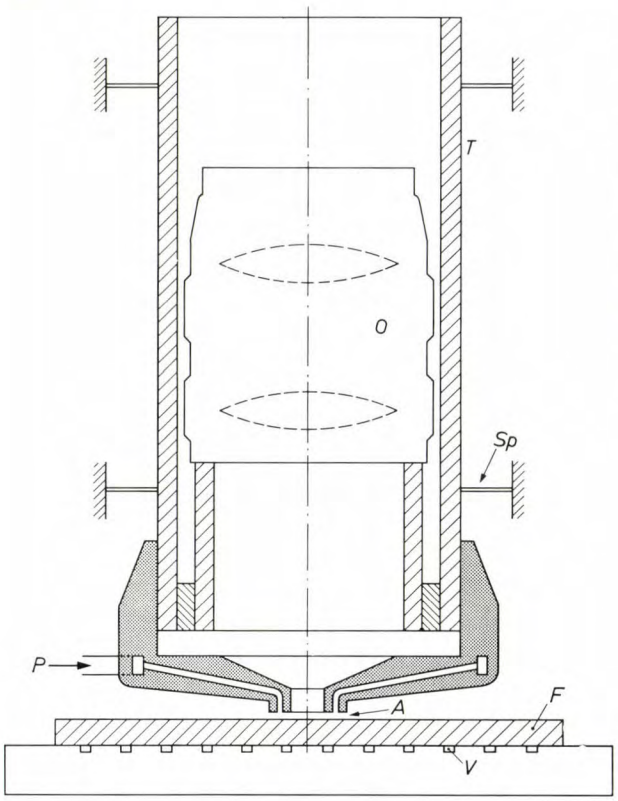


Fig. 14. To prevent defocusing due to out-of-flatness of the photographic plate, the objective mounted in tube T is mounted in diaphragm springs Sp and held by an air bearing A at an accurately constant height from the part of the photographic plate F to be exposed. The compressed air for the bearing is supplied at P. V vacuum suction plate.

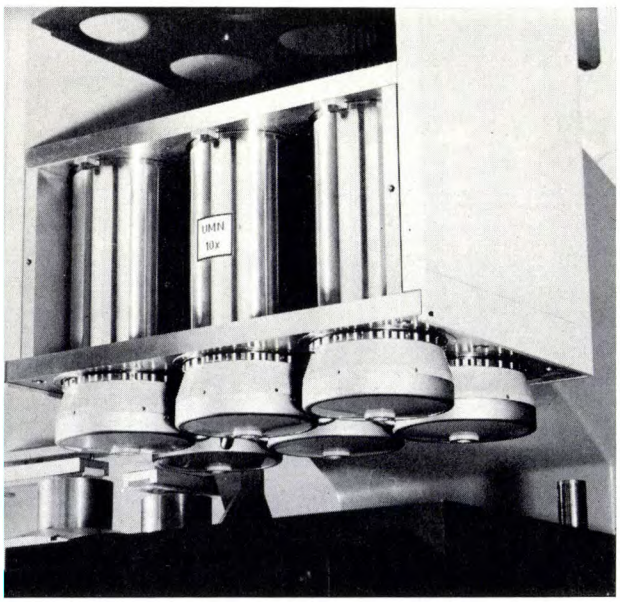


Fig. 15. Arrangement for focusing by means of air bearing. The objective tubes are mounted in diaphragm springs. The circular air bearing can be seen at the lower end of each tube.

and lateral at the start of a row of new patterns — which are accurately established with respect to the axis of the projection system. With the measurement system just described, however, the displacement produced inevitably has to be measured at a certain distance from the optical axes. Due allowance must be made for the possibility that the accuracy of location of subsequent rows of patterns is adversely affected by changes in the distance between the measuring point and the optical axis, or between the measuring point and the photographic plate; see *fig. 16*. Such changes may occur

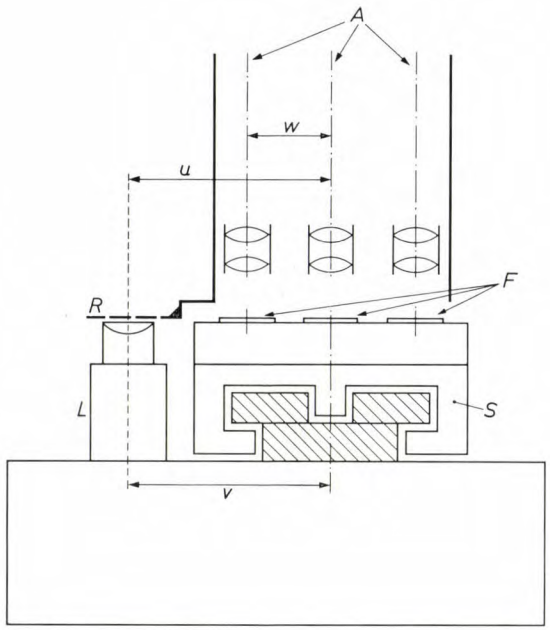


Fig. 16. Displacement measurement, using grating R and measuring head L, is performed at the distances u and v from the optical axis A and photographic plate F respectively.

during a step-and-repeat process, mainly because of thermal expansion. The underframe of the camera, which determines the distance v between the measuring head and the photographic plates, is largely made of granite; the upper part of the structure, however, which determines the distance u between the measuring grating and the optical axes, is mainly made of steel and is much more lightly constructed. The thermal expansion of the two parts thus differs greatly and

[13] P. Kissam, Optical tooling for precise manufacture and alignment, McGraw-Hill, New York 1962, pp. 116-118.

[14] A forerunner of this method is the one proposed by Bovey, for finding the distance by measuring an airflow resistance. See E. Bovey, Pneumatic gauging applied to the focusing of a microscope objective, British Scientific Instrument Research Association, Research Report No. 51, July 1954.

temperature changes during the step-and-repeat process therefore introduce errors.

Because of this situation the temperature of the oil supplied to the hydrostatic bearings is kept constant to within ± 0.05 °C and the flow of oil to the bearings is also maintained during temporary halts in the process. The temperature of the air supplied to the air bearings is controlled to within ± 0.1 °C. To bring the oil and air to the same final temperature the two controlled temperatures have to be set to different values, since the frictional heat developed in the oil restrictions makes the oil about 1 °C warmer, while the Joule-Kelvin effect cools the air in the air-bearing gap by about 0.5 °C.

One factor that helps to give good temperature control is that the camera is in much better thermal contact with the bearing oil than with the air in the room where it is installed — the temperature in this room varies by considerably more than 0.05 °C. Camera warm-up by radiant heat from the operator can be prevented by a simple radiation screen.

To eliminate as far as possible thermal expansion at the measuring head — where errors would have a particularly undesirable effect — the lamps in the measuring heads have been chosen to have a dissipation as small as possible: it is less than 0.1 W per head.

We have measured the temperatures at a number of critical points in one of our step-and-repeat cameras in normal use in a room whose temperature remained constant within ± 0.5 °C. The temperature variation ΔT at these points remained within ± 0.1 °C during a period of 1 hour. In the overall repetition error the contribution that can be expected from temperature changes is proportional to ΔT; it is also proportional to the distance w (see fig. 16) for photomasks made simultaneously and to the distance u for photomasks made successively. If $w = 80$ mm and $u = 150$ mm, the error contributions are found to be ± 0.08 μm and ± 0.15 μm respectively. This estimate only makes allowance for the contraction or expansion of structural elements. Other types of distortion, such as bending, can also occur but are not particularly significant in our design of camera.

Overall repetition accuracy

From the information given above we can draw up the following list of contributions Δx to the total repetition error.

a) Slide system.
Lateral reproducibility errors: $\Delta x_1 = \pm 0.03$ μm.
Rotation ± 0.1 second of arc; therefore, for a distance of 80 mm between adjacent optical axes: $\Delta x_2 = \pm 0.16$ μm.

b) Measuring system (with vibrating-mirror frequency of 4 kHz).
For a carriage speed of 1 mm/s: $\Delta x_3 = \pm 0.12$ μm.

c) Thermal expansion.
For photomasks made simultaneously: $\Delta x_4 = \pm 0.08$ μm.
For photomasks made successively: $\Delta x_5 = \pm 0.15$ μm.

Using these data we can make an estimate of the total expected repetition error from the sum:

$$\Delta x_{tot} = \sqrt{\Sigma(\Delta x_n)^2}.$$

For photomasks made simultaneously we thus find:

$$\Delta x_{tot} = \sqrt{\Delta x_2^2 + \Delta x_4^2} = \pm 0.1 \text{ μm},$$

and for photomasks made successively:

$$\Delta x_{tot} = \sqrt{\Delta x_1^2 + \Delta x_2^2 + \Delta x_3^2 + \Delta x_5^2} = \pm 0.25 \text{ μm}.$$

How do these predictions compare with practical experience? The relative repetition error between two photomasks, whether produced simultaneously or successively, can be measured with a comparator arrangement; see *fig. 17*. The microscope M_1 with a fixed cross-wire is aligned on a particular mark in photomask F_1 by moving carriage S. The micrometer eyepiece E of the second microscope M_2 is then adjusted until the cross-wire is aligned on the corresponding mark in photomask F_2. Carriage S is now moved a relatively large distance, e.g. 30 mm, and M_1 is again aligned on a mark and the adjustable eyepiece of M_2 is adjusted on the corresponding mark. The difference between the two readings on the micro-

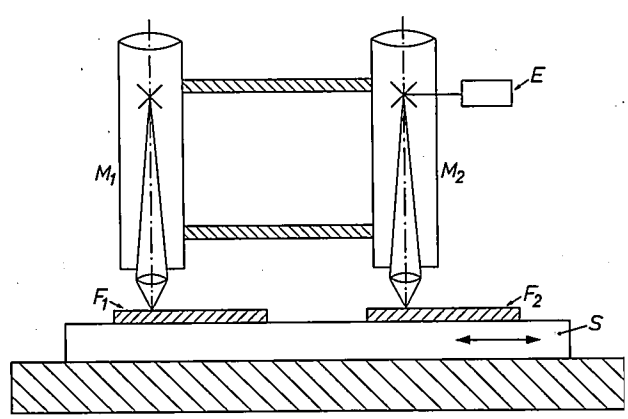


Fig. 17. Comparator arrangement for measuring the repetition error between two photomasks F_1 and F_2. These rest on a carriage S. The carriage is moved until a fixed cross-wire in microscope M_1 is aligned on a mark on F_1, and microscope M_2 is set to the corresponding mark on F_2 with the adjustable micrometer eyepiece E. The carriage is then moved through a large distance and this process is repeated.

meter eyepiece gives the relative repetition error for the two photomasks.

The main difficulty in such measurements using a comparator is that of dry friction. To overcome this difficulty, we have designed a special comparator system with hydrostatic bearings. With optimum setting marks [13] and microscopes giving a magnification of 150×, settings can be made with an accuracy of a few hundredths of a micron. We have checked this with ten successive settings of the microscope M_1 and the adjustable micrometer eyepiece E; the standard deviation calculated from the readings of E was 0.02 μm.

Using one of our six-barrel cameras we then prepared several sets of six photomasks which were given setting marks, suitable for comparator measurements. We then used the comparator to compare the distances *1-2*, *3-4*, *1-3*, *2-4*, *1-4* and *2-3*, shown in *fig. 18*, for the various photomasks. It can be seen that the greatest longitudinal or lateral displacement was 30 mm. The

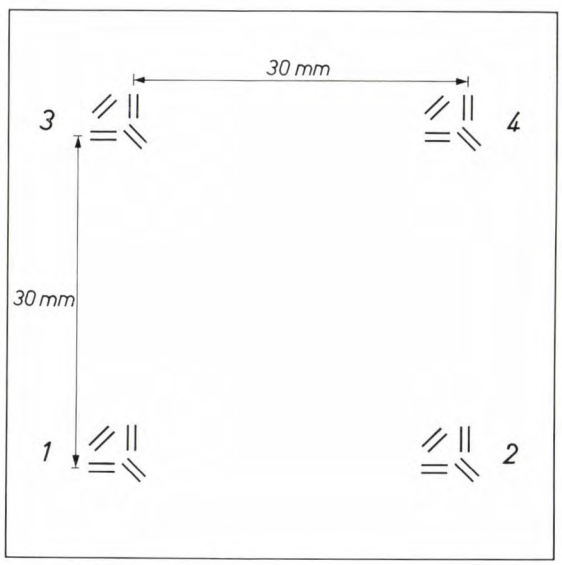


Fig. 18. Measuring points for measurements using the comparator.

results of the measurement can be summarized as follows. Comparison of two photomasks made simultaneously reveals repetition errors smaller than ± 0.05 μm, while comparison of two made successively (with the same projection system) shows repetition errors smaller than ± 0.15 μm.

It is expected that the design principles adopted in our step-and-repeat camera will ensure that the inaccuracies quoted above will not be exceeded in practice, even after a long period of use. One of the six-barrel step-and-repeat cameras has now been in service for eighteen months. The measurements reported above were made after a year of operation.

Application in integrated circuit technology

Two versions of the step-and-repeat camera have been made. The first has an optical system giving a reduction of 20× and is intended for photomasks for high-frequency transistors and for integrated circuits of relatively small dimensions. Patterns containing details with a width of 1 μm can be obtained in a field of 1.6 mm diameter with this camera. *Fig. 19* shows an example of a MOS transistor with a gate 2 μm wide, made with the aid of this camera. A 1 μm wide line is used for this gate in the photomask. By "underetching" the SiO_2 film, which is 0.6 μm thick, a line 2 μm wide is obtained.

The other version has an interchangeable optical system giving reduction of 10× or 4× and is used to

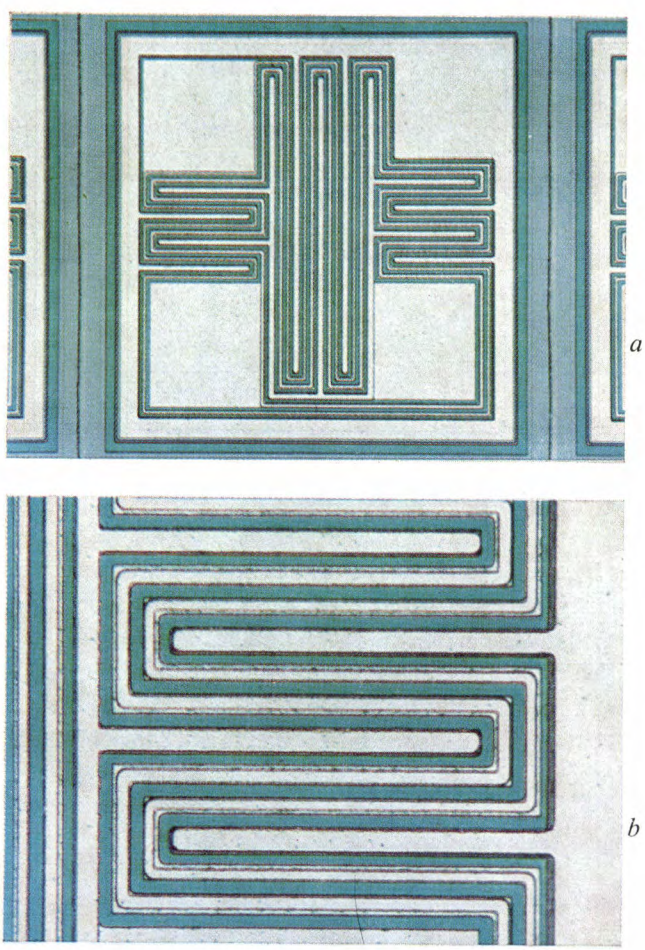


Fig. 19. One of the step-and-repeat cameras described was used in the production of this MOS transistor, which has a 2 μm wide gate. *a*) The complete transistor, 0.5×0.5 mm in size. *b*) Detail. The grey surfaces are the aluminium tracks and terminal areas.

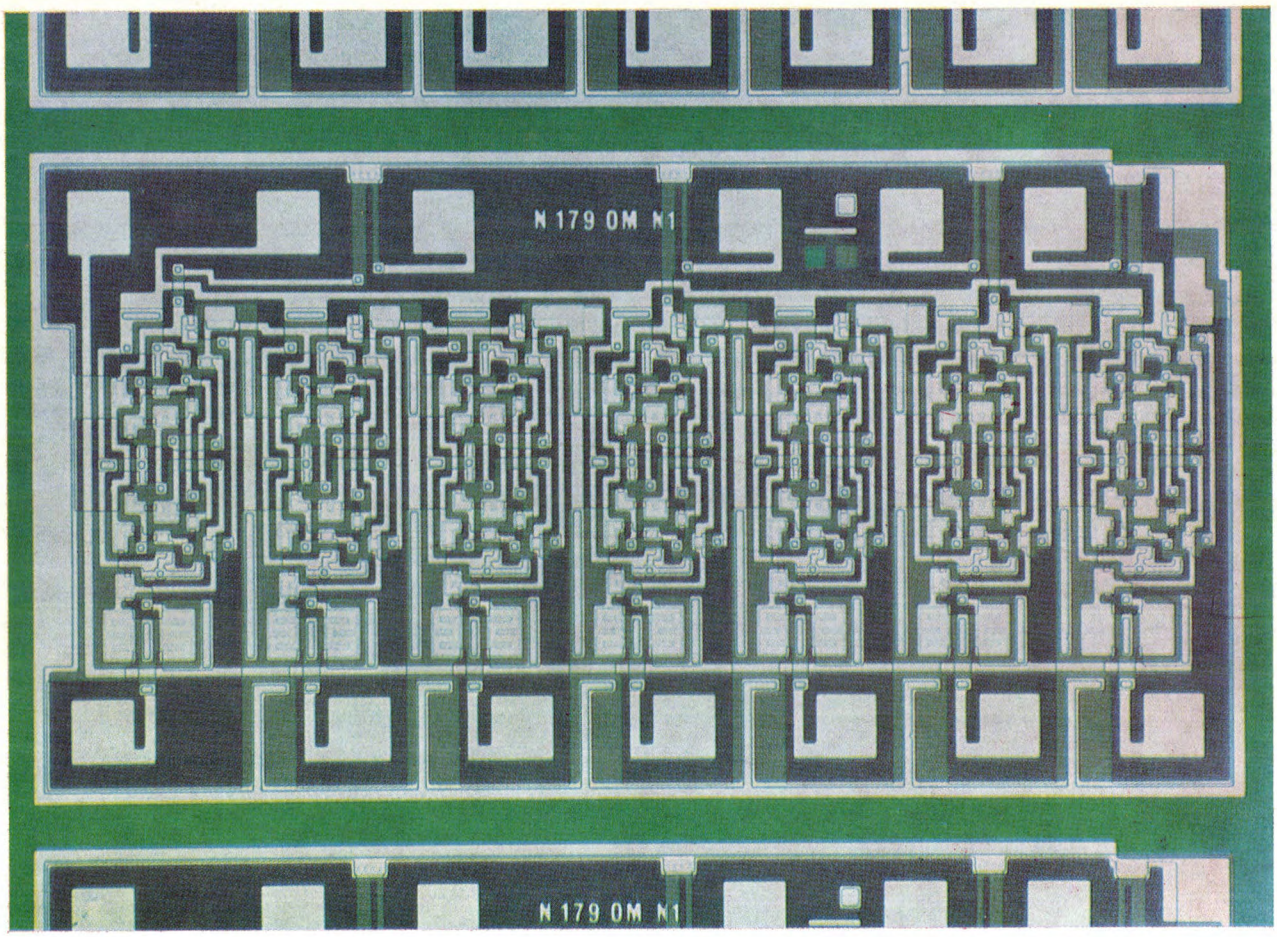


Fig. 20. A complete integrated circuit (a binary divider) made with the aid of one of the step-and-repeat cameras described. Dimensions 2.0×1.08 mm.

make photomasks for larger integrated circuits. Lines with widths of 2 and 3 μm in fields with diameters of 4.3 and 10 mm respectively can be obtained. *Fig. 20* shows an example of a larger integrated circuit produced with the aid of this camera.

The development of the step-and-repeat camera described here was made possible by the cooperation of many members of these Laboratories and of the Electronic Components and Materials Division (Elcoma). In particular the contribution made by A. G. Bouwer of the Precision Engineering Group of these Laboratories should be mentioned.

Summary. The photographic process used in the production of integrated circuits requires a set of photomasks that register very accurately with each other. These are made by photographic reduction of mask patterns; in this process 100 to 10 000 single images of the pattern for a circuit, each typically 1×1 mm, are formed on one mask for the purposes of quantity production. One method now widely used for making the masks employs a step-and-repeat camera, in a multi-barrel arrangement, for producing several masks simultaneously. The article describes a six-barrel step-and-repeat camera with carriages moving on hydrostatic bearings and a new grating measuring system for controlling their movements. As a result of numerous refinements in the bearings, in the measurement of displacement and in the adjustment of the object plates with the mask patterns and the focusing of the objectives, together with close control of the temperature, the total "repetition error", i.e. the relative error between two masks when one is placed on top of the other (whether produced simultaneously or successively), has been kept down to $\pm$ 0.2 micron. Various examples are shown.

The use of digital circuits in data transmission

P. J. van Gerwen

Data transmission is the name given to the transmission of digital signals in such a form that they can be applied directly to a computer. Now that computers are being more and more widely used, it is becoming much more important to be able to transmit these signals reliably and at high speed. In this article it is shown that where data signals have to be transmitted over ordinary telephone lines, digital circuits have several advantages over conventional circuits containing inductors and capacitors. A survey is given of digital circuits designed for this purpose. The use of such circuits has made it possible to design a data transmitter in the form of an integrated circuit on a chip a few millimetres square.

A significant development of the last few years is the growth of a new branch of computer technology: the combination of computer and telecommunications. To make the most economic use of the computer, it is located at a central point and connected to the more or less distant places where the users are situated. The information to be processed then has to be supplied to the computer, often over a telephone line, and the resultant information is communicated to the user in the same way. The same situation is found where a computer memory store is used as an information processing centre, e.g. for the book-keeping of bank transactions. Connection with the store allows any branch of the bank to have details of the state of customers' accounts without human intervention, and thus permits demands for payment to be accepted without delay.

In this kind of procedure the information (the data) has to be transmitted in such a way that it can be directly processed by the computer; this is known as *data transmission*. Existing means of telecommunication would seem to be the most convenient means for this purpose, in particular telegraph and telephone lines. Can these be used directly, without modification for transmitting data signals, i.e. signals consisting of series of *binary pulses*? Although telegraph systems are capable of handling binary signals, the bandwidth available per telegraph channel is so small that the speed at which the characters can be transmitted is far too low for modern data transmission. In telephone circuits the bandwidth is greater, giving a higher rate of transmission, but in order to use a telephone system for binary pulse trains an extra operation is required. Telephone lines are, after all, designed for speech: only signals at frequencies required for intelligibility are transmitted (ranging from 300 to 3400 Hz). Data signals, however, have a spectrum that also possesses a d.c. component and a number of a.c. components outside this band, including components at very low frequencies. Another difficulty is caused by the fact that *carrier telephony systems* with many channels are used for long-distance communications. In this case there is usually a small frequency difference between the carrier that is modulated in a channel at the transmitting end and the carrier that is used for demodulation at the receiving end. Because of this there is a small frequency shift in the spectrum of the transmitted signals. This does not matter for speech, where only the spectral distribution of the signals is of importance; for data transmission, however, the *shape* of the signals is important, since it is determined by the time function. The shift in frequency upsets the harmonic relation between the components of the spectrum, and this may lead to more than the admissible level of distortion.

This distortion could also be aggravated by the fact that the phase characteristic of telephone circuits is strongly curved, particularly at the edges of the transmitted spectrum. If one were to attempt to transmit signals below 300 Hz in a telephone channel, the signals would be severely distorted.

However, it should be possible to avoid these difficulties by applying the extra operation mentioned above. In this operation the signals to be transmitted *modulate*

P. J. van Gerwen is with Philips Research Laboratories, Eindhoven.

a separate carrier in such a way that the principal components of the modulated signal lie within the frequency band available for telephony. The original signal is then obtained by demodulation with this "data carrier" in the receiver. Now if there is a frequency shift of the spectrum because the signal has arrived via a carrier telephony system, there will also be a shift in the data carrier. If the data carrier used for demodulation in the receiver is given the same frequency shift, which can be done by conventional methods, the demodulated data signal again has the required shape.

Amplitude, frequency or phase modulation may be used. Each of these methods has its advantages and disadvantages for data transmission, but we shall not go into this here.

The essentials of a data system

Fig. 1 shows a block diagram of a conventional data transmitter. The binary elements of the data signal, the "bits", are usually supplied to the transmitter in time intervals controlled by a *clock signal*, which consists of a series of equally spaced pulses. The frequency of this signal, the *clock frequency*, is thus equal to the number of bits per second, the *bit frequency*. Transmission using a clock signal is known as *synchronous data transmission*. The clock signal can be transmitted in one form or another to the receiver, where it is used for reconstitution of the data signal already demodulated with the appropriate carrier frequency.

The data signal modulates the carrier at the transmitter in a *modulator*. The carrier signal can be produced by an oscillator. If this is tied to the clock signal, so that a fixed relation exists between the clock frequency and the carrier frequency, the data system is said to be *fully synchronized*. Instead of using a separate oscillator, the carrier signal can also be derived from the clock signal.

Since the output signal from the transmitter has to remain within the frequency band laid down for telephony, it is necessary to use *filters* in the transmitter. As indicated in fig. 1, the data signal is usually put through a low-pass filter before being applied to the modulator. After the filtering process, the higher-frequency components in the data signal are no longer present; the signal is then no longer binary. The output signal from the modulator is passed through a band-pass filter to suppress unwanted modulation products. If both sidebands of the modulated signal are transmitted, the design of this filter presents no great problems. In vestigial-sideband modulation however (see below) this band-pass filter also has to nearly suppress one of the sidebands. The design of the filter is then much more difficult.

Fig. 2 shows the block diagram of a conventional *receiver* for data signals. A band-pass filter is used here at the input to suppress interference outside the required frequency band. After demodulation the signal again goes through a low-pass filter, and the original form of the data signals is then reconstituted with the aid of the clock signal.

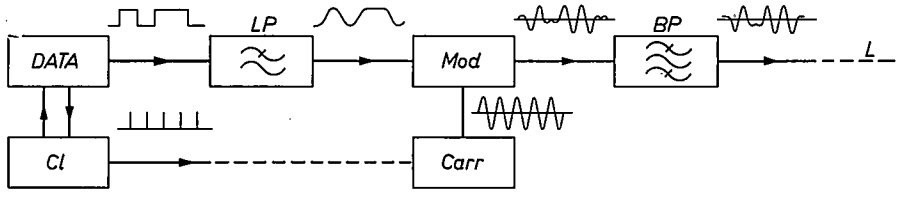


Fig. 1. Block diagram of a conventional transmitter used for transmitting data via a telephone line. *DATA* data-signal source, *Cl* clock-signal source, *Mod* modulator, *Carr* carrier source, *LP* low-pass filter, *BP* band-pass filter, *L* telephone line. The waveforms of the signals are shown beside the appropriate connections. The modulation system used here is phase modulation.

Digital circuits

The conventional filters widely used in telephony, made up from inductors and capacitors, can in principle also be used for data transmission. Their use, however, involves particular problems. Besides the *attenuation characteristic* (attenuation of the signal as a function of frequency) the *phase characteristic* (phase shift as a function of frequency) now has to be taken into account as well. To transmit the bits correctly it is desirable that the phase shift should be proportional to the frequency, in other words that the phase characteristic should be linear. (In telephony the phase characteristic is not usually important.) This makes it difficult to build conventional filters for data transmission, especially if a sharp cut-off is required at the edge of the frequency band; i.e. if the attenuation characteristic has to have steep sides. There is then no

alternative but to apply a phase correction.

A new principle of filter design has now been developed which presents new and attractive possibilities. As will be shown, signals in certain frequency bands can be suppressed by means of *delay elements*. This makes it possible to obtain an attenuation characteristic which has steep sides yet still has a linear phase characteristic. Such filters therefore require no phase correction. These filters are particularly attractive for digital signals, since they can be designed entirely with digital techniques, using bistable circuits (flip-flops) as delay elements. The operation and application of such *digital filters* [1] is one of the subjects of this article.

In conventional telecommunication systems inductors or transformers are usually used in the modulators. In a data transmission system there are some advantages if these are also made up from digital circuits. Some circuits for digital modulators will be dealt with in more detail shortly.

Another feature of digital circuits that makes their use for data transmission attractive is their great *versatility*. We mean by this that an existing system can often be combined with other units to allow operation at other transmission rates, and that other features such as the attenuation characteristic of a filter can easily be modified. In conventional techniques this usually requires an entirely new circuit design.

Transmission rate and bandwidth

The binary elements of a data signal can be transmitted at a rate which increases with the bandwidth of the transmission path. At a bandwidth of B hertz it is theoretically possible to transmit $2B$ binary elements per second. Many of the data systems in use today are still a long way from this theoretical limit, since the amplitude and phase characteristics of the transmission path also have an effect on the transmission rate. For example, a non-linear phase characteristic can cause

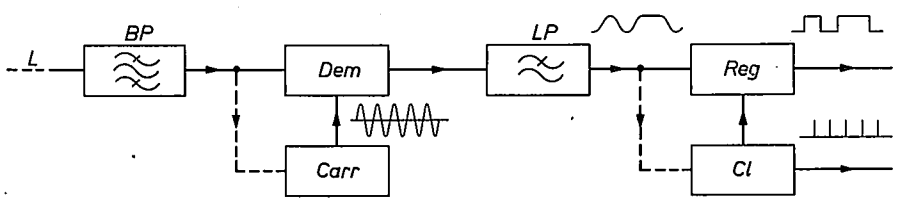


Fig. 2. Block diagram of a conventional data receiver. *L* telephone line, *BP* band-pass filter, *Dem* demodulator, *Carr* carrier, *LP* low-pass filter, *Reg* circuit for reconstituting the data signals, *Cl* clock-signal source.

The situation is different for the *data receiver*. The modulated signal entering the receiver should be considered as an analogue signal rather than a digital signal. The advantages of digital circuits seem less obvious here. In fact, however, there are many advantages in including digital circuits in data receivers, and it seems likely that this will become standard practice in future. This point will be dealt with at the end of the article.

The rapid increase in the use of digital circuits is partly due to the advances made in the *microminiaturization* of electronic circuits. Digital circuits require a fairly large number of elements, and it has now proved possible to design these circuits so as to use only transistors and resistors — components that lend themselves well to monolithic construction in an integrated circuit. This greatly reduces the disadvantage of using a large number of elements, and later on in this article we shall discuss a data transmitter incorporated in a crystal chip a few millimetres square.

distortion which increases with the transmission rate. Moreover, the permissible rate depends on the form in which the modulated signal is transmitted. The theoretical limit mentioned above of $2B$ elements per second is based on the assumption that only one of the sidebands of the modulated signal is transmitted. Often, however, both sidebands are transmitted, and this reduces the permissible rate by a factor of 2. It is difficult to suppress one of the sidebands completely. Since the data signals contain components at very low frequencies, signal components arise in the modulation process which have frequencies very close to the carrier frequency. It is therefore no easy matter to make a filter that can clearly separate the two-sidebands. A middle way can be taken by using vestigial-sideband modula-

[1] See P. Leuthold, Filternetzwerke mit digitalen Schieberegistern, Philips Res. Repts. Suppl. 1967, No. 5, also published in Mitteilungen aus dem Institut für Hochfrequenztechnik an der Eidgenössischen Technischen Hochschule in Zürich published by Prof. Dr. F. E. Borgnis, 1967, No. 1.

tion, in which one of the sidebands is only partly suppressed [2].

In the foregoing we have been concerned with a carrier which is modulated by a binary signal, that is to say the carrier can be in two distinct "states". (This may relate to the amplitude, the frequency or the phase, depending on the method of modulation used.) It has been found, however, that an even greater rate can be reached by modulating in such a way that more than two states of the carrier are used. With a "quaternary" system, for instance, using four different carrier states, the transmission rate can be doubled. A method of modulating the carrier in such a way consists in starting from a data signal which is digital, but not binary, and thus contains elements with different amplitudes. (Quaternary modulation methods do also exist, however, which start from a binary data signal.)

A disadvantage of this method of increasing the transmission rate is that it makes the system more sensitive to noise and other interference. Moreover, it imposes even stricter requirements on the amplitude and phase characteristics of the transmission path.

Filters based on delay elements

We have mentioned above that signal components at certain frequencies can be suppressed by using a delay network. This can be seen as follows.

If an input signal u_1 of frequency ω is given a delay τ and this delayed signal is added to the original one (*fig. 3*), we may write the expression for the output signal as:

$$u_2 = u_1 \{1 + \exp(-j\omega\tau)\} = 2u_1 \cos \tfrac{1}{2}\omega\tau \exp(-\tfrac{1}{2}j\omega\tau). \quad \ldots \quad (1)$$

The amplitude of u_2 thus varies with frequency as a cosine function. At the frequencies where $\cos \frac{1}{2}\omega\tau = 0$ we have $u_2 = 0$. There is also a phase shift between u_2 and u_1, which is equal to $\frac{1}{2}\omega\tau$, i.e. proportional to the frequency.

A filter may be obtained by using a large number of these delay elements. *Fig. 4* shows a series arrangement of $2n$ elements, each of which gives a delay τ. The output signals of the elements are added in a particular series of ratios, which are expressed by the introduction of weighting factors C_k $(-n \leqq k \leqq n)$, which operate on the signals before addition. If the input signal is again u_1, then the output signal obtained in this way is:

$$u_2 = u_1 \sum_{k=-n}^{+n} C_k \exp\{-j\omega(n+k)\tau\}. \quad \ldots \quad (2)$$

If the delay network is symmetrical, in other words if $C_k = C_{-k}$, then we can write (2) in the form:

$$u_2 = u_1 \left[C_0 + 2 \sum_{k=1}^{n} C_k \cos k\omega\tau\right] \exp(-j\omega n\tau). \quad (3)$$

The phase shift between u_1 and u_2 is $\omega n\tau$, which is thus again proportional to the frequency. *The phase characteristic is therefore linear.* The amplitude characteristic is:

$$A(\omega) = C_0 + 2 \sum_{k=1}^{n} C_k \cos k\omega\tau. \quad \ldots \quad (4)$$

The form of this characteristic can thus be influenced by the choice of the weighting factors. The best approximation to a specified shape for the amplitude characteristic $A(\omega)$ is obtained when the weighting factors satisfy the equation:

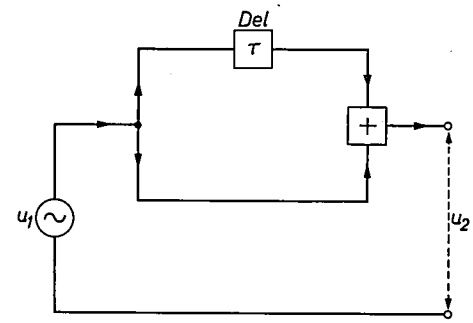


Fig. 3. When a sinusoidal signal u_1 is fed to a delay element *Del*, and the delayed signal is added to u_1, the result is an output signal u_2 which is zero at certain frequencies.

$$C_k = \frac{\tau}{2\pi} \int_{-\pi/\tau}^{\pi/\tau} A(\omega) \cos k\omega\tau \, d\omega. \quad \ldots \quad (5)$$

With the weighting factors chosen in this way the shape of the function $A(\omega)$ is usually only matched exactly for $n = \infty$, i.e. for an infinitely large number

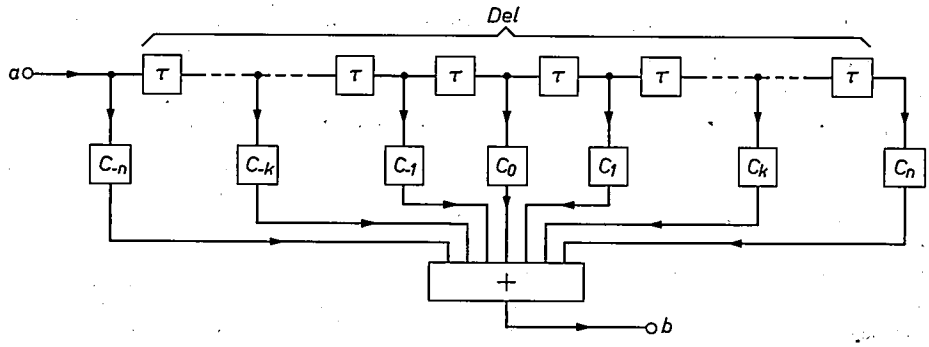


Fig. 4. Series arrangement of $2n$ delay elements *Del*, each with a delay time τ. The output signals of the elements are added in certain ratios given by the weighting factors $C_{-n} \ldots C_n$. *a* input, *b* output.

of delay elements. In practice sufficient elements are chosen to give the required degree of approximation.

From equations (3) and (4) we see that the delay networks under consideration, with our symmetry condition $C_k = C_{-k}$, are able to provide all the amplitude characteristics that can be represented by a Fourier series containing cosine terms only. This implies that A is an even periodic function of ω. An amplitude characteristic of this nature is shown in *fig. 5*. Its period f_r is equal to $1/\tau$. The value of τ is usually taken at the smallest that will give an adequate approximation to $A(\omega)$ with the chosen number of elements. The frequency f_r then becomes so high that it is easy to limit the passband to one of the bands indicated in

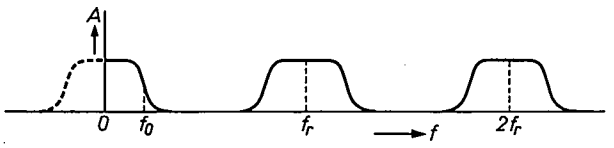


Fig. 5. Amplitude characteristic of a filter with delay elements. This characteristic can be represented by a Fourier series containing cosine terms only. The repetition period f_r of the passbands is equal to $1/\tau$ (where τ is the delay time of the elements); the cut-off frequency of the first passband is denoted by f_0. If the first passband only is required, the others can be suppressed by means of a simple filter.

$$A(\omega) = 2 \sum_{k=1}^{\infty} C_k \sin k\omega\tau. \quad \ldots\ldots \quad (6)$$

This is obtained exactly with weighting factors C_k given by:

$$C_k = \frac{\tau}{2\pi} \int_{-\pi/\tau}^{\pi/\tau} A(\omega) \sin k\omega\tau \, d\omega. \quad \ldots\ldots \quad (7)$$

In this case it also follows from equation (2) that besides a component proportional to the frequency the phase shift between input and output contains a constant component equal to 90°. This can be utilized for producing signals with single-sideband modulation [3].

Filters made on this principle are known as "transversal filters" [4]. They have not been widely applied because it is difficult to make low-loss delay elements for analogue signals from inductors and capacitors. Another difficulty is that such filters are rather large since they require a large number of delay elements.

Digital filters

Both of the above difficulties disappear if filters exclusively for digital signals are to be designed. The delay elements can then be bistable circuits (flip-flops). These are connected in series and in effect form a *shift register* (see *fig. 6*). The bistable circuits are switched from one

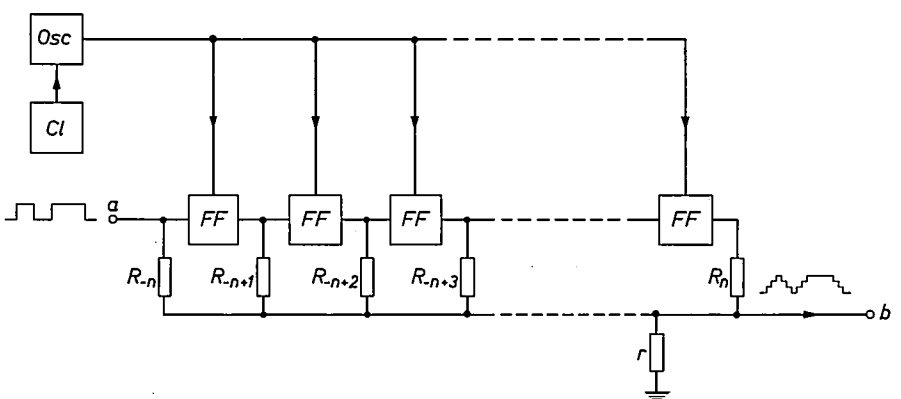


Fig. 6. Digital filter made up from a shift register consisting of a series arrangement of bistable circuits (flip-flops) *FF*. $R_{-n} \ldots R_n$ and r form a resistance network, which ensures that the output voltages of the bistable circuits are added in a particular series of ratios. *Osc* oscillator that produces the control pulses for the shift register, *a* input, *b* output.

the figure. For example, the signals at frequencies close to f_r, $2f_r$, etc., can be suppressed with a simple *RC* network, retaining only the lowest passband (cut-off frequency f_0 in fig. 5). When this is done the amplitude and phase of the signals in this passband are not significantly affected, provided f_r is large. The result is thus a low-pass filter with a linear phase characteristic. A band-pass filter can be made in much the same way.

A linear phase characteristic can also be obtained by choosing the weighting factors such that $C_k = -C_{-k}$. It follows from (2) that in this case the amplitude characteristic consists of a series with sine terms only. Assuming for simplicity that the number of delay elements is infinitely large, the characteristic is given by:

state to the other by control pulses, causing the signal to travel through the shift register. The frequency of these pulses should be higher than the clock frequency of the data signal, and is preferably an integral multiple of the clock frequency. The clock signals can for

[2] One of the sidebands can be completely suppressed, to double the transmission rate, if a special code is applied to the data signal, thus suppressing the components in the modulated signal that are close to the carrier. This is described in: P. J. van Gerwen, On the generation and application of pseudo-ternary codes in pulse transmission, Philips Res. Repts. **20**, 469-484, 1965.

[3] P. Leuthold and F. Tisi, Ein Einseitenbandsystem für Datenübertragung, Archiv elektr. Übertr. **21**, 354-362, 1967.

[4] See H. E. Kallmann, Transversal filters, Proc. I.R.E. **28**, 302-310, 1940.

example be obtained by frequency division from the oscillator supplying the control pulses. Conversely, the control pulses can be derived by frequency multiplication from the clock signal, or the control-pulse oscillator can be synchronized with a higher harmonic of the clock signal.

The weighting factors for the output signals of the bistable circuits are obtained with a resistance network consisting of resistors $R_{-n} \ldots R_n$ and r. The voltage across r is then proportional to the sum of these signals. Since a bistable circuit can deliver two output signals in opposite phase, the weighting factor may be positive or negative as desired. The output signal does of course contain components whose frequencies are close to those of the control pulse and their higher harmonics. These can be suppressed by connecting a capacitor across r.

Fig. 7 shows the attenuation characteristic of a filter of this kind which has 24 bistable circuits. The frequency f_r has been taken as $10f_0$. The dashed line represents the characteristic that would be obtained with an infinite number of elements.

It is important to note that the weighting factors, and hence the shape of the amplitude characteristic of the filter, are determined by the *relative* values of the resistors. This kind of filter (except for the capacitor) therefore lends itself very readily to integrated circuit techniques, since the relative values of the resistors in an integrated circuit can be fixed much more accurately than their absolute values.

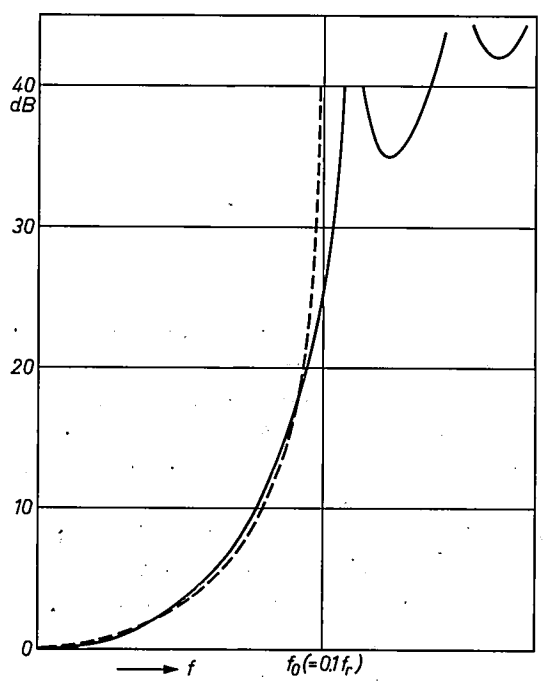


Fig. 7. Attenuation characteristic of a digital filter containing 24 bistable circuits, where $f_r = 10f_0$. With an infinite number of bistable circuits the characteristic would follow the dashed line.

Since a digital filter with a shift register is far easier to make than a filter containing a large number of delay elements for analogue signals, digital filters have also been used for analogue signals in some cases. In this case the analogue signal is first converted into a digital signal in an analogue-to-digital converter. (This can be done by using pulse-code modulation or delta modulation.) The filter is then followed by a decoder. Filters built in this way are known as *analogue code filters*.

Digital modulators

The design of a digital modulator is based on the use of a carrier in binary form, i.e. a "square wave"; its repetition frequency should be an integral multiple of half the clock frequency of the data signal. The data signal also has to be supplied to the modulator in binary form. (It is therefore not possible to use a low-pass filter as in fig. 1.) With a carrier of this form, it is fairly easy to design digital modulators for amplitude modulation and also for frequency and phase modulation, as we shall now show.

Amplitude modulation can be obtained by means of an AND gate with two inputs. If we apply to these inputs two different binary signals A and B, whose instantaneous values can be referred to as 1 and 0 respectively, we obtain at the output a signal U which has the value 1 if both A and B have this value. If $A = 0$ or $B = 0$ (or $A = B = 0$), then $U = 0$. In Boolean algebra this function is expressed as $A \cdot B$. *Fig. 8* shows the form of a binary data signal A with a "binary carrier" B whose frequency (the fundamental frequency, in analogue terms) is twice the clock frequency of A. The carrier again occurs in the output signal $U = A \cdot B$, but now modulated in amplitude by the data signal. Since the carrier is not sinusoidal, the modulation products of its higher harmonics are also obtained. As a rule these lie far outside the available frequency band and can therefore be suppressed with a simple filter. The output signal U also contains the data signal itself, and since the waveforms of A and B never cross the zero line, U also contains a d.c. term. The data signal lies in the frequency band of the modulated signal to be transmitted, and is therefore in effect a distortion of the modulated signal. It can easily be eliminated, however, by subtracting the data signal at half-amplitude from the output signal of the AND gate (*fig. 9*). This also has the result of removing the d.c. term from the output signal U. As can be seen from fig. 9, the signal thus obtained is no longer binary but ternary (with three levels). Such a signal can no longer be directly processed in a digital filter with bistable circuits, since these can only handle binary signals.

To obtain *frequency modulation* with a digital circuit

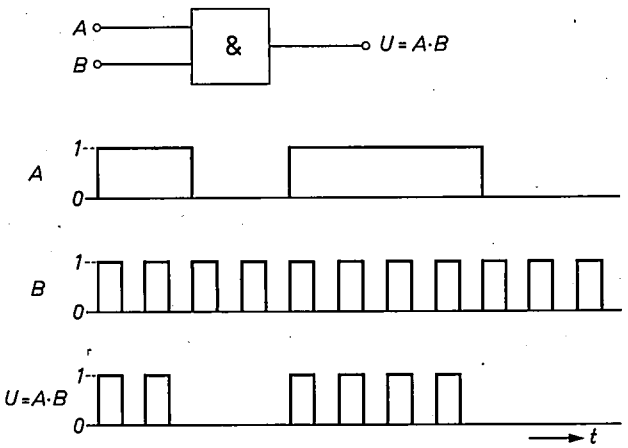


Fig. 8. AND gate with two inputs. When two binary signals A and B are applied to these inputs, the resultant output signal is $U = A \cdot B$. (Some of the symbols in this figure and figs. 9, 10 and 11 follow recent international recommendations; these are indicated below fig. 10.)

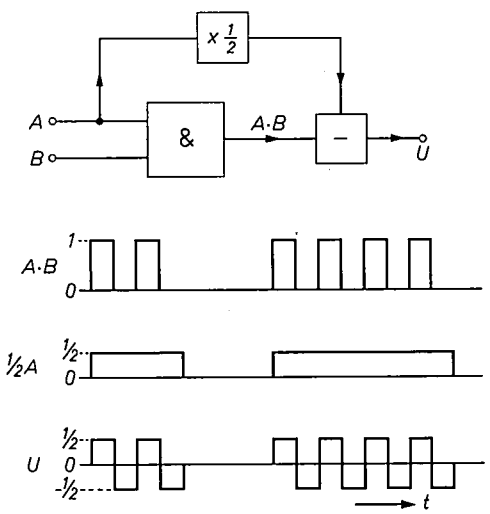


Fig. 9. When the data signal A at half-amplitude is subtracted from the output signal of an AND gate a ternary output signal U is obtained. This does not contain the data signal or a d.c. term, but only the carrier, modulated by the data signal.

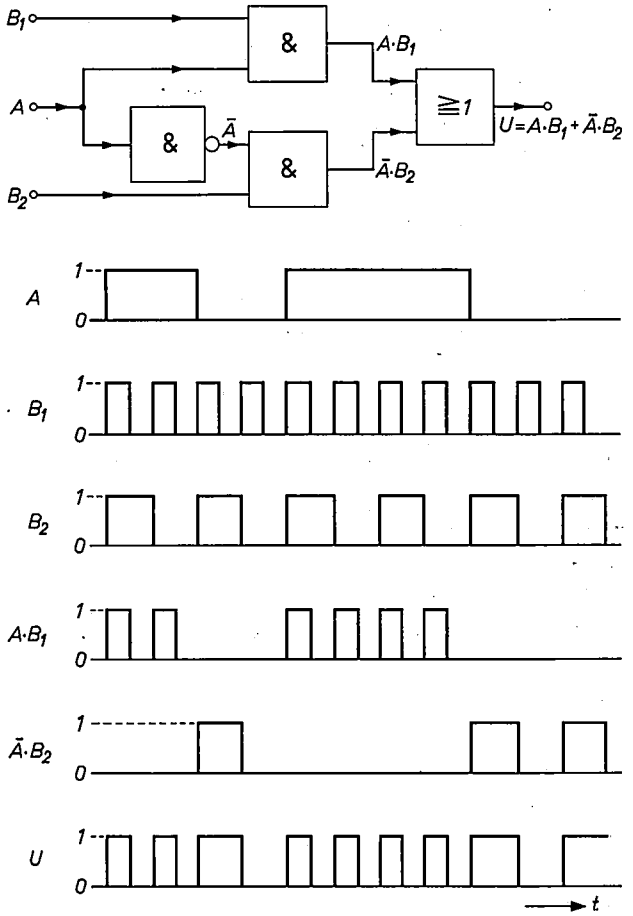
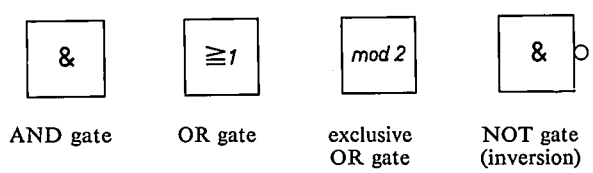


Fig. 10. Digital frequency modulator made up from two AND gates, one OR gate and one NOT gate, which supplies the inverted signal $\bar{A}$. A data signal, B_1 and B_2 binary carriers at different frequencies. The output signal U is now a binary signal with frequency modulation.



it is necessary to have two digital carriers at different frequencies. (These can be obtained by frequency multiplication or by frequency division using a single oscillator.) The circuit must now work in such a way that either the one or the other carrier is transmitted, depending on whether the data signal has the value 1 or 0. An example of such a circuit is shown in *fig. 10*. It is built up from two AND gates, an OR gate and a NOT gate. The data signal is applied to the input A and the two carriers to the inputs B_1 and B_2. The frequency of B_1 here is twice that of B_2. The NOT gate delivers the inverted data signal $\bar{A}$. The output signal is now written as $A \cdot B_1 + \bar{A} \cdot B_2$ in Boolean algebra. This signal, together with its various component parts, is also shown in fig. 10. It is a digital signal whose frequency changes whenever the data signal switches from the value 0 to 1 or *vice versa*.

A digital *phase modulator* can be obtained by using a circuit known as a "*modulo-2 adder*". This circuit, also referred to as an *exclusive OR gate*, delivers an output signal at the value 1 when the input signals are unequal. The output signal is 0 when the input signals are equal (0 or 1). This relation between the input signals A and B and the output signal U is represented by the equation:

$$U = A \cdot \bar{B} + \bar{A} \cdot B.$$

In *fig. 11* this signal is shown together with its component parts. As long as the data signal A is zero, U is equal to the carrier; if A has the value 1, then U is equal

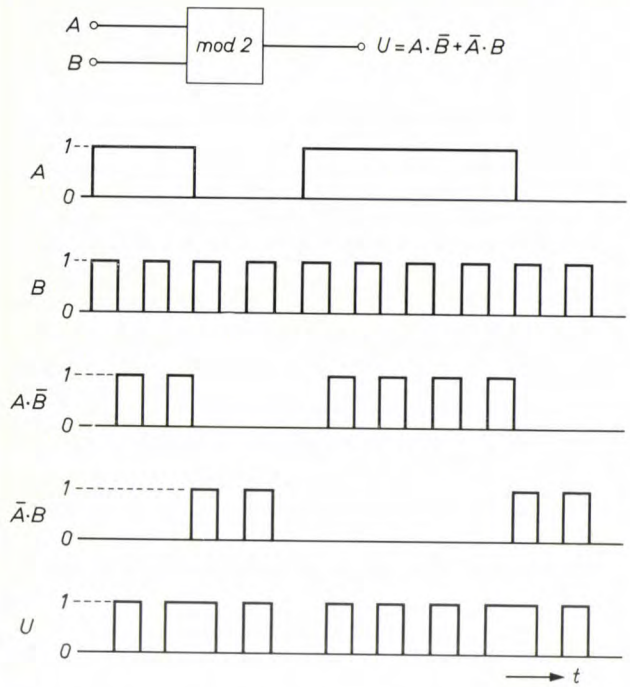


Fig. 11. Exclusive OR gate. *A* data signal, *B* carrier. In the output signal *U* the carrier is modulated in phase with the data signal.

to the inverted carrier. This is in fact phase modulation.

With binary signals a simple relation exists between phase modulation and amplitude modulation. If the inverted carrier is added to a phase-modulated digital signal the result is an amplitude-modulated signal. *Fig. 12* shows the signal *U* from fig. 11, together with the inverted carrier $\bar{B}$ and their sum *U'*. We see that the latter (ternary) signal is in fact amplitude-modulated by the data signal *A* of fig. 11.

The modulators discussed above differ from the ones normally used in radio engineering and in other branches of electronics in another respect besides the digital techniques. In ordinary modulators the carrier frequency is generally much higher than the frequency

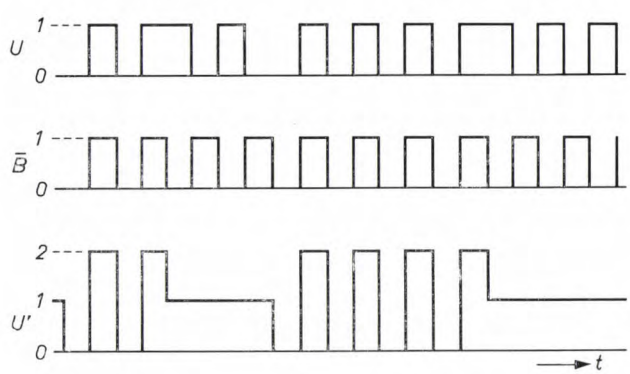


Fig. 12. When the inverted carrier $\bar{B}$ is added to the phase-modulated signal *U* (see fig. 11) an amplitude-modulated ternary signal *U'* is obtained.

of the modulating signal. The result is a signal whose spectrum is symmetrical with respect to the carrier wave. In data transmission, however, a carrier of say 1800 Hz is modulated with a data signal with a clock frequency of for example 1200 Hz. (The clock frequency and the carrier frequency may even be identical.) Since the data signal is digital, it contains higher harmonics, and in this case their frequencies are higher than the carrier frequency. If the carrier frequency is an integral multiple of half the bit frequency, the spectrum of the modulated signal will be asymmetrical. This is illustrated in *fig. 13*, which shows the spectrum of a

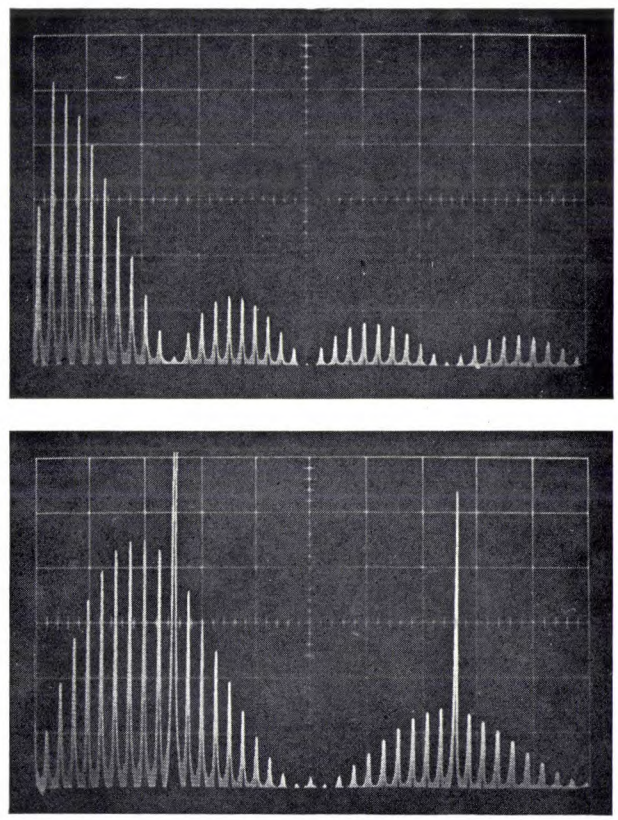


Fig. 13. *a*) Spectrum of a data signal consisting of periodic pulses. *b*) Spectrum of a phase-modulated signal, where the carrier frequency is equal to the clock frequency of the data signal.

data signal, consisting of periodic pulses (fig. 13*a*), and the spectrum that results when phase modulation is applied with a carrier frequency equal to the clock frequency of the data signal (fig. 13*b*). This latter spectrum consists of groups of components near to the carrier and near to its odd harmonics. The asymmetry is particularly evident near the carrier.

A data system is generally designed in such a way that only the spectral region surrounding the carrier is transmitted, while the other components are suppressed by a filter. The original objective that we mentioned at the beginning of the article, bringing the

spectrum inside the telephony band, has thus been achieved.

The asymmetry in the spectrum of the transmitted signal can cause interference effects. It may give rise to crosstalk between the bits of the data signal, resulting in a smaller permissible interference margin during the reconstitution of the signal in the receiver. We shall just note here that when the carrier frequency is an integral multiple of half the clock frequency the asymmetry in the spectrum can be removed by means of a band-pass filter that passes the modulated signal (see fig. 1). If this filter is given an asymmetrical attenuation characteristic, a symmetrical spectrum can again be obtained. The shape of the attenuation characteristic of a digital filter can be influenced, as we have seen, by the choice of the resistors that determine the weighting factors (fig. 6).

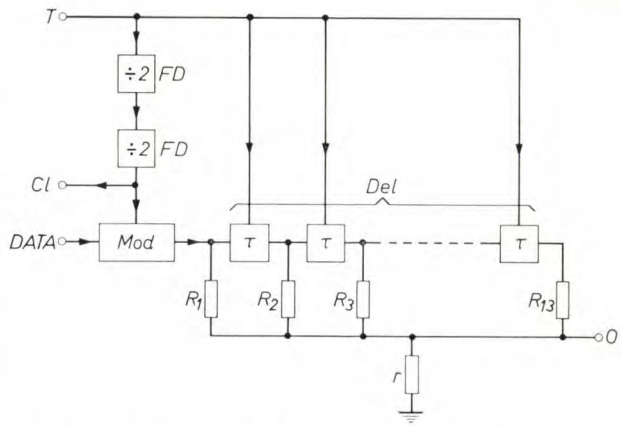


Fig. 14. Block diagram of a data transmitter using phase modulation. *Mod* modulator, *FD* frequency dividers, *Del* shift register consisting of 12 bistable circuits, *T* terminal for the control pulses for the shift register, *Cl* clock signal, *O* output.

A data transmitter as an integrated circuit

Since digital circuits, which contain no inductors and only a single capacitor, can now be used both for modulators and for filters, it is possible to build a data transmitter almost entirely in the form of an integrated circuit. *Fig. 14* shows the block diagram of a vestigial-sideband transmitter which, as an integrated circuit, measures only 2.7 × 2.1 mm. *Fig. 15* is a much enlarged

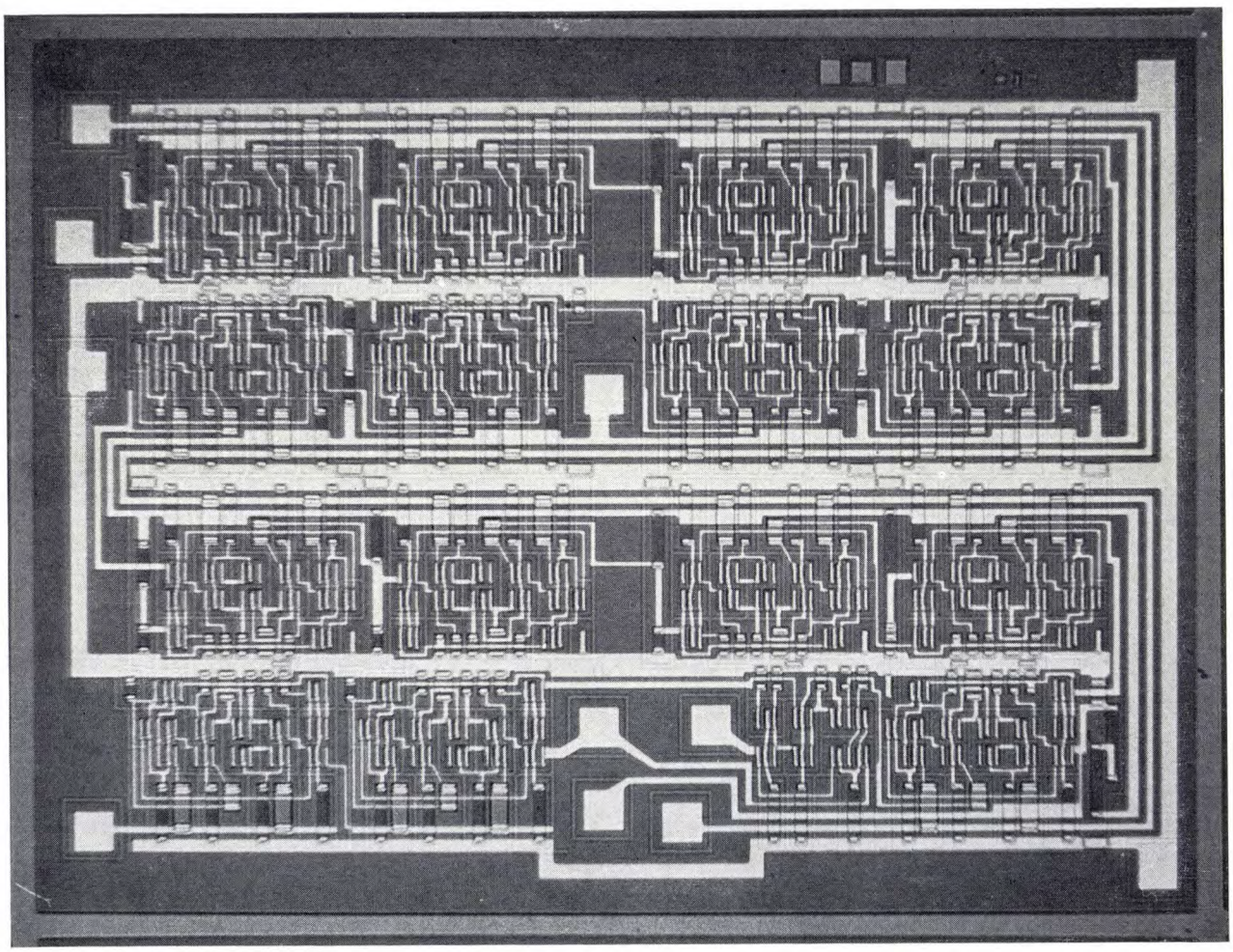


Fig. 15. The data transmitter of fig. 14 as an integrated circuit, measuring 2.7 × 2.1 mm and including 203 transistors and 172 resistors. The upper three rows each contain four bistable circuits, the lower row contains the modulator and the two frequency dividers.

picture of this monolithic circuit, which includes 203 transistors and 172 resistors. The transmitter comprises a modulator for phase modulation, a filter containing twelve bistable circuits, and two frequency dividers (step-down ratio 1 : 2) which derive the clock signal and the carrier from the control pulses of the filter. Bistable circuits with an exceptionally low power consumption are used [5], giving a total power consumption of only 120 mW. The control pulses come from an oscillator which is not included in the circuit. If the frequency of this oscillator is 9.6 kHz, a data signal can be transmitted at a rate of 2400 bits per second. The carrier frequency is then equal to the bit frequency, and the spectrum of the modulated output signal lies within the range from 600 to 3000 Hz. This transmitter can however also be used for a much higher transmission rate (a channel with a larger bandwidth is then required). If the oscillator frequency is increased by a factor of 20, the transmission rate and the carrier frequency are increased by the same factor. This means that the transmission rate is then 48 000 bits per second. The spectrum of the modulated signal then ranges from 12 to 60 kHz.

The versatility of the transmitter is also demonstrated by the relative ease with which it can be altered from amplitude modulation to phase modulation. We have already shown how this can be done by simply adding the inverted carrier to the phase-modulated signal (fig. 12). The result is a ternary signal, and we have seen that the spectrum of this signal cannot be limited with a digital filter made up from bistable circuits. What can be done, however, is to pass the phase-modulated (binary) signal through the filter before adding the inverted carrier to it.

Another block diagram (*fig. 16*) demonstrates further ways of increasing the versatility of a transmitter (at least, in so far as no limitations are set by the unit to be integrated). The clock pulses and the carrier in this case are derived from the oscillator that delivers the control pulses for the shift register, by means of three frequency dividers, one of which has a fixed step-down ratio of 2. (The carrier frequency is thus always an integral multiple of half the clock frequency.) By suitably choosing the other step-down ratios, m and k, the transmitter can be made to operate at different transmission rates with the oscillator frequency kept fixed. For example, with the same oscillator frequency, clock signals can be obtained at frequencies of 600, 1200 or 2400 bits/s, and there are various possibilities for the carrier frequency. With an oscillator frequency of say 28.8 kHz, and with $m = 8$ and $k = 3$, we obtain a clock signal of 1200 bits/s and a carrier frequency of 1800 Hz.

The transmitter of fig. 16 also allows a signal to be transmitted with vestigial-sideband modulation or with double-sideband modulation. For this purpose the shift register is provided with two resistance networks, which can be connected in by two switches S_1 and S_2. In one of these networks the resistors have values which allow the lower sideband to be transmitted, and in the other they are chosen to allow the upper sideband to be transmitted. If both switches are closed, both sidebands are transmitted. *Fig. 17* shows the corresponding attenuation characteristics, measured at a carrier frequency of 1800 Hz and a clock signal of 1200 bits/s. Finally, *fig. 18* shows some of the corresponding signals: the data signal, the modulated signal (with phase modulation), the output signal from the filter and the same signal after suppression of frequency-components in the higher passbands.

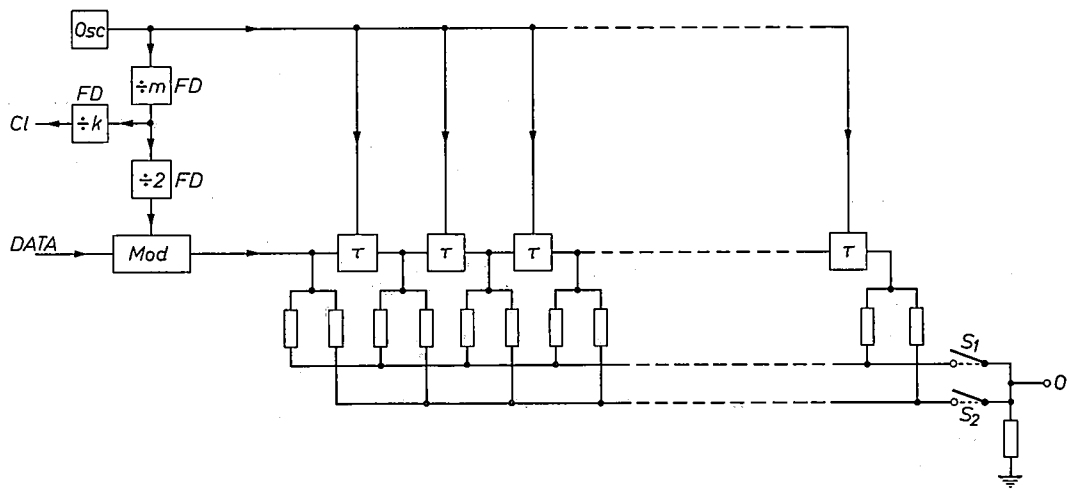


Fig. 16. Block diagram of a more versatile data transmitter. The switches S_1 and S_2 allow either the upper sideband, the lower sideband or both sidebands to be transmitted. The appropriate choice of the step-down ratios m and k gives a fairly wide margin of freedom in the choice of the clock frequency and the carrier frequency.

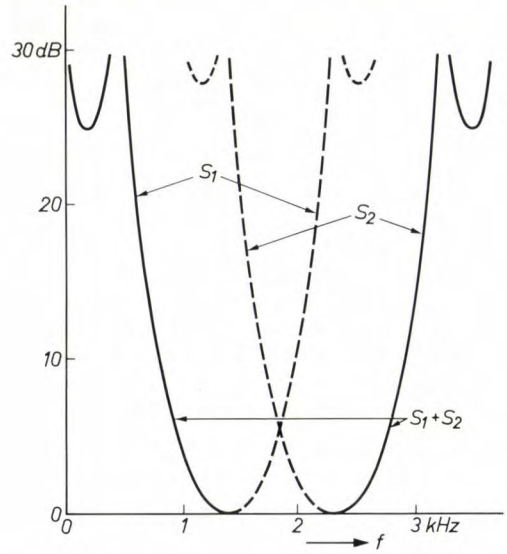


Fig. 17. Attenuation characteristics that can be obtained with the circuit of fig. 16 with a carrier frequency of 1800 Hz and a clock frequency of 1200 bits/s. When switch S_1 is closed, the lower sideband is transmitted and the upper sideband is strongly attenuated; when S_2 is closed the filter transmits the upper sideband. When both switches are closed, both sidebands are transmitted.

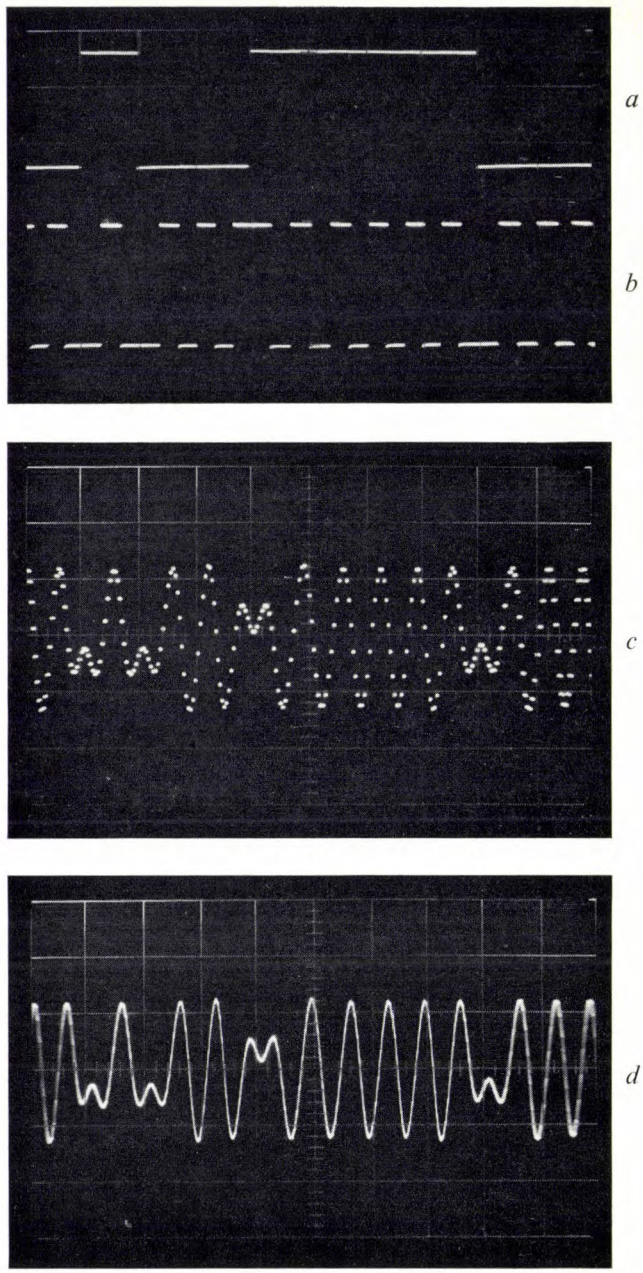


Fig. 18. Signals occurring in the circuit shown in fig. 16. *a* data signal, *b* phase-modulated signal before the filter, *c* signal after the filter with components in the higher passbands, *d* signal arising when the latter components are suppressed by means of an extra filter.

Digital circuits in data receivers

As we noted earlier, the signals reaching a data receiver via a telephone line are no longer binary (or ternary, etc.) because of the bandwidth limitation and the possibility of a frequency shift. For this reason conventional receivers generally use analogue circuits. However, the rapid advances in integrated-circuit techniques present new possibilities here as well. This applies particularly to the filters transmitting the signal before and after the demodulator (see fig. 2). A difficulty arises with the analogue code filters mentioned earlier, in which the analogue signal is fed to a digital filter through an analogue-digital converter and the output signal of this filter passes through a decoder. In order to obtain a sufficient accuracy in encoding the analogue signal digitally a fairly high frequency has to be used for the control pulses of the shift registers. This high frequency necessitates the use of a large number of shift register elements. However, when integrated circuits are used, this is no longer a real difficulty, and it therefore seems likely that the use of analogue code filters in data receivers will in due course become an economic proposition. Since digital circuits are already being used for other receiver functions, in particular for reconstituting the data signal, both data receivers and data transmitters will then consist almost entirely of digital circuits.

[5] See A. Slob, Fast logic circuits with low energy consumption, Philips tech. Rev. **29**, 363-367, 1968 (No. 12).

Summary. The circuits used in conventional data transmission systems are very similar to those used in other branches of electronics: filters, modulators, etc., which are mainly made up from inductors and capacitors. Since data signals consist of series of binary pulses, however, many of the circuit functions can be performed by digital circuits, which are easily made as integrated circuits. Digital filters can be made by using shift registers, and logic circuits such as AND, OR and NOT gates can be used as digital modulators. In this way it was possible to make a complete data transmitter in integrated form on a crystal chip measuring 2.7×2.1 mm. A particular feature of the circuits is their great versatility, since various characteristics of the data transmitter can be easily modified. Wider use of digital circuits in data receivers is also to be expected.

Stoichiometry

W. Albers and C. Haas

The initial outcome of the historic dispute on the composition of chemical compounds (Dalton 1766-1844; Berthollet 1748-1822; Proust 1754-1826; etc.) was that the composition of compounds is constant, with simple ratios between the numbers of the constituent atoms. Although thermodynamics later showed that, strictly speaking, the composition of every compound can vary inside a certain region (the existence region), the simple concept of the exact stoichiometric composition has lasted well. Modern research and experimental methods have provided deeper insight into the composition of compounds.

Many physical properties vary markedly with composition within the existence region. These include properties like colour, diffusion, electrical conductivity, thermoelectric e.m.f., Hall effect, photoconductivity, luminescence, magnetic susceptibility, laser action, etc. The marked variability of these properties — particularly in semiconductors and insulators — within an often extremely narrow existence region has caught the attention of industry, and as a result various interesting technical applications have been developed in the last twenty years.

The article which starts below discusses questions concerned with stoichiometry and the existence region. In the first part, it is shown that all compounds have an existence region; the second part will deal in greater detail with the related point defects in the crystal lattice and with the way in which their concentrations can be controlled; and the last part will discuss the determination of existence regions.

I. Existence region

It appears from ordinary chemical analysis that compounds that are poor electrical conductors (insulators, semiconductors) can be assigned an exact proportionality formula, in which the relative numbers of the constituent atoms are given by ratios between integral numbers, usually quite simple ones. This formula is known as the stoichiometric composition [1]. Compounds belonging to this group include all the liquid and solid organic compounds and some of the liquid and solid inorganic compounds. In addition it has been found that the group of metallic inorganic compounds and alloys do not have an exact stoichiometric composition, but are often stable over a wide range of compositions, in other words they have a wide existence region.

More exact experiments demonstrate, however, that the compounds that are insulators or semiconductors also possess an existence region, although it is only a very narrow one. From thermodynamic considerations it can be shown that all compounds have an existence region. Theoretically there is no difference between the two groups. This means that in any compound, say A_pB_q, a certain quantity of native atoms or molecules can be randomly dissolved (at least when $T > 0$ °K, see below). For crystalline substances, this means that at $T > 0$ °K there is always a fraction of the constituent atoms in disorder. This has very important consequences, which are also of significance in technical applications. Many physical properties are based on this solubility; they have their origin in the native point

Dr. W. Albers is with Philips Research Laboratories, Eindhoven; Prof. Dr. C. Haas, formerly with Philips Research Laboratories, is now Professor of Inorganic Chemistry at the University of Groningen.

defects, and may thus vary strongly with the composition inside the existence region. Because of this it is in many cases impossible to characterize the chemically pure compound by one specific value of a given property, such as the value of the electrical resistivity. However, it is possible to indicate the minimum and maximum values at a given temperature.

It is of great importance to know the physical potentialities of a substance. In the last ten years many investigators have for this reason attempted to determine the maximum solubilities of the constituent components as a function of temperature, paying particular attention to compounds with very narrow existence regions.

Determining the limits of existence regions, narrow and broad, is the subject of the third part of this article; the manner in which the concentrations of point defects and charge carriers can be controlled *inside* the existence region — thus controlling many properties of the substance — is the subject of the second part. We shall now start with a few fundamental equations.

Phase equilibria

At a given temperature T and pressure P the chemical equilibria in a closed system will be such that the Gibbs free energy (sometimes called the free enthalpy) G is at a minimum, so that:

$$(\partial G)_{P,T} = 0. \quad \ldots\ldots \quad (1)$$

Two situations may arise here. The atoms may form a single homogeneous body or may form into a number of homogeneous regions (solid, liquid and gaseous) which represent different states with, as a rule, different compositions. These different states, which are in equilibrium with each other and are in contact with each other at an interface, are called the phases of the system. A distinction between different phases has no significance unless these phases can coexist.

If two bodies are in equilibrium with one another, their temperatures are equal and also their pressures, since bodies in equilibrium exert equal and opposite forces on each other through the interface. The system can be made to undergo a small change by changing the temperature by dT, by changing the pressure by dP and also by the addition of a small quantity dn of a substance. The change in the Gibbs free energy is then:

$$dG = -S\,dT + V\,dP + \mu\,dn, \quad \ldots \quad (2)$$

where S is the entropy, V the volume, T the absolute temperature and μ the chemical potential of the substance.

If the substance contains atoms or groups of atoms of a dissimilar kind i, they can be added separately, so that equation (2) can be generalized to:

$$dG = -S\,dT + V\,dP + \sum_i \mu(i)\,dn(i), \quad \ldots \quad (3)$$

where $\mu(i)$ is the chemical potential of the atoms i. After integration at constant P and T we find from equation (3):

$$G = \sum_i \mu(i)\,n(i). \quad \ldots\ldots \quad (4)$$

From equation (1) a number of other equilibrium equations can be derived. The atoms i may occur in different states, for example in two states indicated by I and II. If the closed system in equilibrium containing states I and II is taken as the sum of two open subsystems I and II (between which there can be exchange of material), with Gibbs free energies G_I and G_{II}, then for the transfer of a fraction $dn(i)$ from I to II we have from equations (1) and (3):

$$\partial G_{tot} = 0 = \partial G_I + \partial G_{II} = \{\mu(i_I) - \mu(i_{II})\}\,dn(i),$$

from which it follows that:

$$\mu(i_I) = \mu(i_{II}). \quad \ldots\ldots \quad (5)$$

(∂ means that all the other variables remain constant during the variation.) In equilibrium the chemical potentials of component i in states I and II are thus the same.

For a two-phase equilibrium in a binary system A-B we may therefore write:

$$\mu(A_{\text{phase 1}}) = \mu(A_{\text{phase 2}}), \quad \ldots \quad (6)$$

$$\mu(B_{\text{phase 1}}) = \mu(B_{\text{phase 2}}). \quad \ldots \quad (7)$$

Different states may also occur within one phase. In a crystal of an element A most atoms will be situated at lattice sites (A_{latt}, state I), but a small fraction will not (*A, state II, see below). In equilibrium, therefore, from equation (5):

$$\mu(A_{latt}) = \mu(*A). \quad \ldots\ldots \quad (8)$$

An atom may also occur in a number of different states in liquid and gaseous phase. For example, in a gas in equilibrium an A atom may be bound to another A atom (state I) and in addition there may be free A atoms (state II), so that $\mu(A_I) = \mu(A_{II})$. If two atoms A_{II} are added to sub-system II with the simultaneous removal of two atoms A_I from sub-system I, which is equivalent to removing an A_2 molecule, we can write:

$$\mu(A_2) = 2\,\mu(A). \quad \ldots\ldots \quad (9)$$

In the same way in a gaseous equilibrium:

$$\mu(AB) = \mu(A) + \mu(B), \quad \ldots \quad (10)$$

---

[1] J. A. A. Ketelaar, Chem. Weekbl. **32**, 58 and 262, 1935; **35**, 852, 1938; **37**, 522, 1940 (in Dutch).

when A and B occur both bound to each other and free. In equations (9) and (10) $\mu(A_2)$ and $\mu(AB)$ are again defined as the change of G due to the separate addition of A_2 and AB, respectively. Equations (9) and (10) also apply to liquids in which there is dissociation and to solid compounds with dissolved native atoms (native disorder), since the derivation is a general one, based on the sole premise that in equilibrium atoms in different states are present side by side. For a solid substance with the stoichiometric formula A_pB_q and dissolved native atoms *A and *B we can thus write:

$$\mu(A_pB_q) = p\,\mu(*A) + q\,\mu(*B). \quad . \quad (11)$$

Every compound has an existence region

To find the stable range of concentrations in a multi-component system with the aid of equation (1) it is necessary to know G as a function of the composition. In a binary mixture A-B the composition may be given by the fraction

$$x = \frac{n(B)}{n(A) + n(B)}.$$

From equation (4) it follows that:

$$G = (1 - x)\,\mu(A) + x\,\mu(B). \quad . \quad . \quad (12)$$

For the special case of no interaction between the B atoms it may be deduced that:

$$\mu(B) = \mu^0(B) + RT \ln x, \quad . \quad . \quad (13)$$

with $\mu^0(B) = H(B) - TS(B)$,

where x is expressed as the atomic fraction, and H and S are respectively the enthalpy and the entropy per gramme-atom of the appropriate element. Varying G at constant P and T it follows from equations (4) and (3) that:

$$\sum_i n(i)\,d\mu(i) = 0 \quad . \quad . \quad . \quad . \quad (14)$$

(the Gibbs-Duhem relation), which, with the aid of equations (12) and (13), gives $\partial\mu(A)/\partial x = -RT/(1-x)$, so that:

$$\mu(A) = \mu^0(A) + RT \ln (1-x), \quad . \quad (15)$$

with $\mu^0(A) = H(A) - TS(A)$.

Solutions for which equations (13) and (15) apply over the whole range of concentrations are referred to as *ideal* solutions. Substituting equations (13) and (15) into (12) we then find the Gibbs free energy of the ideal mixture:

consisting of a part l which is linear in x and of a logarithmic part m, called the mixing term (see *fig. 1*). This mixing term is negative, has a minimum value $-RT \ln 2$ at $x = \frac{1}{2}$ and, since $\partial m/\partial x = RT \ln\{x/(1-x)\}$, has the slope $-\infty$ at $x = 0$ and the slope $+\infty$ at $x = 1$. These infinite slopes form the basis of the disorder in all substances at $T > 0$ °K.

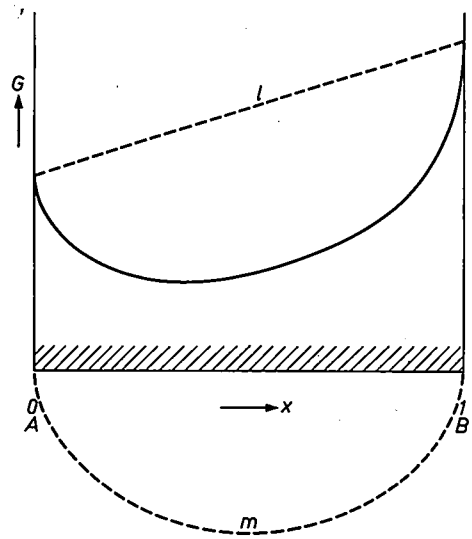


Fig. 1. Gibbs free energy (sometimes called the free enthalpy) G as a function of x, the atomic fraction B of an ideal mixture A-B, at constant pressure and temperature (solid line). This G-x curve consists of a part l which is linear in x, and a "mixing term" m (dashed lines). The hatching indicates the region in which complete mixing takes place, in this case over the entire range of compositions.

For *non-ideal* solutions the term l in equation (16) is not linear in x. If the particles of the same type attract one another more strongly than the dissimilar particles, the curve of l against x will bend upwards (mixing uses up heat; the enthalpy of the system increases), and as a result of this there may be a hump in the G-x curve (*fig. 2*). In this case the equilibrium condition (1) leads to separation into two coexisting phases (referred to as disproportionation). The compositions of the two phases (I and II) can be found as follows. From equations (12) and the Gibbs-Duhem relation (14) we have:

$$\frac{\partial G}{\partial x} = \mu(B) - \mu(A). \quad . \quad . \quad . \quad (17)$$

Substituting equation (17) into equation (12) gives:

$$\mu(A) = G - x\left(\frac{\partial G}{\partial x}\right), \quad . \quad . \quad . \quad (18)$$

$$\mu(B) = G + (1 - x)\left(\frac{\partial G}{\partial x}\right). \quad (19)$$

$$G = \underbrace{\{H(A) - TS(A)\}(1-x) + \{H(B) - TS(B)\}x}_{l} + \underbrace{RT\{(1-x)\ln(1-x) + x \ln x\}}_{m}, \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (16)$$

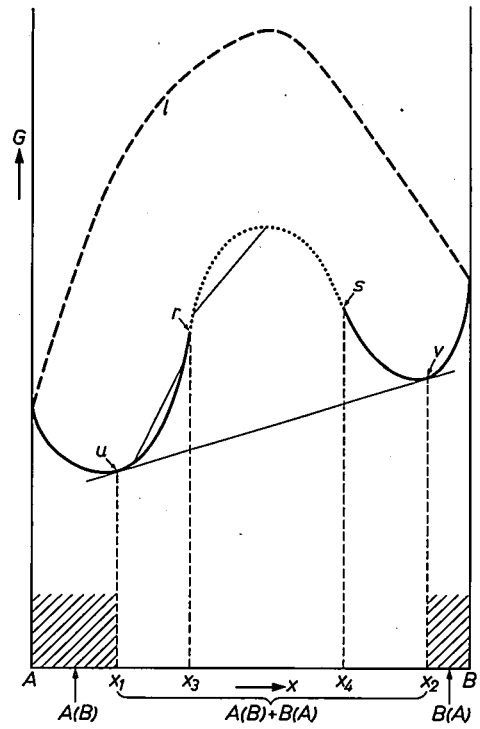


Fig. 2. Figure corresponding to fig. 1 for a non-ideal system. A and B have the same symmetry, and no compound is formed. The location of the tangent *uv* determines the boundaries of the regions of miscibility and immiscibility. A(B) is the phase in which B is dissolved in A, and B(A) is the phase in which A is dissolved in B. The significance of *r* and *s* and the dotted part of the *G-x* curve is explained in the text.

By combining equations (6), (7) and (17) we obtain:

$$\left(\frac{\partial G}{\partial x}\right)_{x=x_1} = \left(\frac{\partial G}{\partial x}\right)_{x=x_2}, \quad \ldots \quad (20)$$

where x_1 and x_2 are the compositions of the coexisting phases I and II. From equations (18) and (20) it follows that:

$$\left(G - x\frac{\partial G}{\partial x}\right)_{x=x_1} = \left(G - x\frac{\partial G}{\partial x}\right)_{x=x_2} \quad \ldots \quad (21)$$

From equations (20) and (21) it follows that x_1 and x_2 can be found by constructing a tangent, such as *uv* in fig. 2.

Between x_1 and x_2 the vertical distance between the *G-x* curve and the tangent *uv* gives the difference in Gibbs free energy between the mixed and the disproportionated system. The *G-x* curve between the deflection points *r* and *s*, with compositions x_3 and x_4, is shown dashed to indicate that in this region the mixed system A-B is unstable because local disproportionation on a microscopic scale causes a decrease in Gibbs free energy (see thin line between *r* and *s*). Consequently, once a disproportionation has occurred, no matter how small it may be, it will persist. In the regions x_1-x_3 and x_4-x_2 the system is metastable, since in this case a small disproportionation may disappear (see thin line between *u* and *r*). In these concentration regions the mixed system can thus exist for a shorter or a longer time, e.g. upon freezing-in.

At $T = 0$ °K the mixing term *m* in equation (16) vanishes, so that there is disproportionation into pure A and B. At the same time A and B are built up in an entirely regular fashion. There is thus no *A present with A_{latt} (equation 8), since owing to the presence of randomly dissolved *A, the entropy in that case would contain a mixing term of finite value, whereas the Nernst heat theorem requires that $S = 0$ at $T = 0$ °K. The mixing term does however make a contribution for $T > 0$ °K, causing an infinite slope in the *G-x* curve at $x = 0$ and $x = 1$, which is always greater than the (finite) increase due to *l*. As a result there is disproportionation into an A-rich phase in which B is dissolved: A(B), and a B-rich phase in which A is dissolved: B(A). It therefore follows from this that for $T > 0$ °K every substance is soluble to some extent in every other. This is always true, for both ideal and non-ideal solutions, since equations (13) and (15) are generally valid for small values of *x*.

In figs. 1 and 2 the same *G-x* curve describes both A(B) and B(A), which is only possible if they have the same symmetry [2]. Generally speaking, coexisting phases in the solid state will have different symmetry. The conclusions just drawn, however, remain valid. Suppose that at a given pressure *P* and temperature *T* the phase A is stable in a crystal structure with space group I, and phase B is stable in a structure with space group II, so that $G(A_I) < G(A_{II})$ and $G(B_{II}) < G(B_I)$. Let us consider, for example, with reference to *fig. 3*, the case where A and B mix ideally in the same space

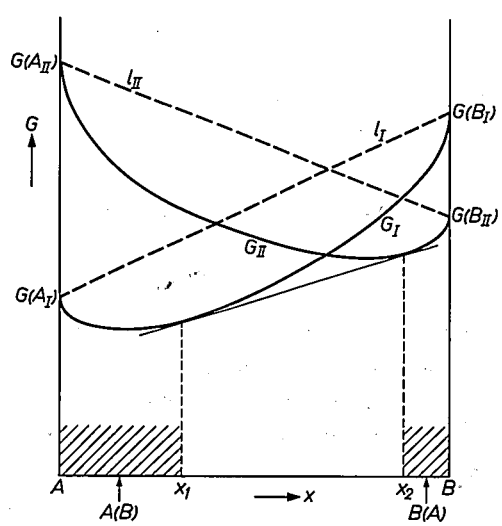


Fig. 3. Figure corresponding to fig. 1, but with different symmetries for the solid phases, denoted by I and II.

[2] This subject will be dealt with at greater length in a future article in this journal on "Phase transitions and symmetry".

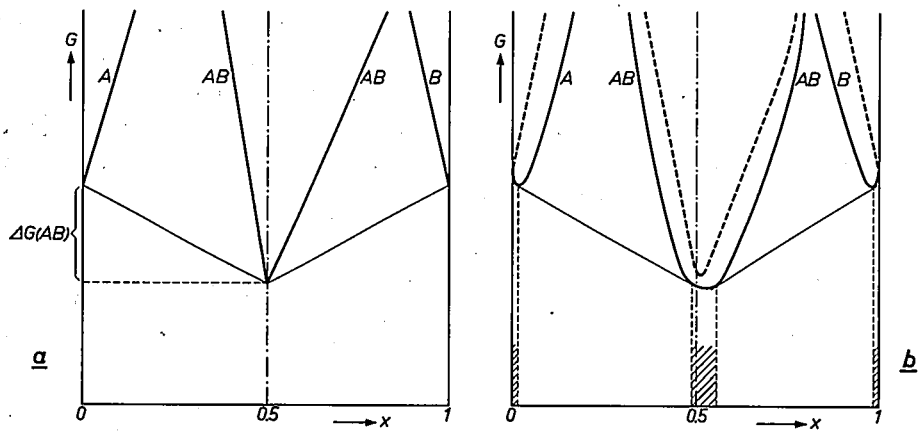


Fig. 4. G-x curves for a system A-B where the compound AB is stable, *a*) at $T = 0$ °K, *b*) at higher temperature. The dashed lines are again the l terms; the chain-dotted lines indicate the stoichiometric composition of AB; ΔG(AB) is the Gibbs free energy of the formation of AB per gramme-atom.

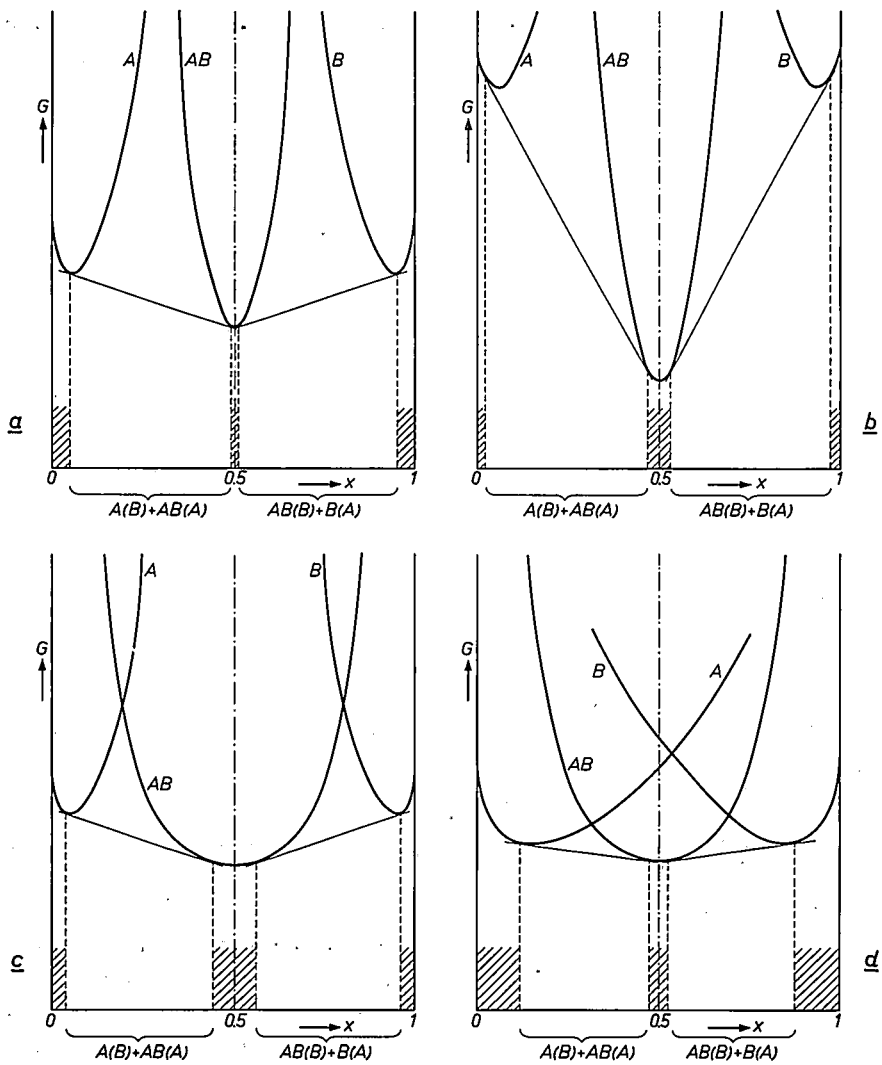


Fig. 5. The width of the existence region depends on the magnitude of the Gibbs free energy ΔG(AB) of the formation of AB (shown for a high value of ΔG in *b*), on the steepness of the G-x curve of the compound AB (with marked curvature shown in *a* and *b*), and on the G-x curves of the adjacent phases A and B (with smaller curvature drawn in *d*).

groups: The dashed lines l_{I} and l_{II} are linear in x, and the solid curves represent $G(x)_{\mathrm{I}}$ and $G(x)_{\mathrm{II}}$ respectively. Using the tangent construction one can immediately see that in this case there is also disproportionation at $T = 0$ °K into pure A and B, but that at $T > 0$ °K both A and B are always soluble in each other.

If the molecules of different kinds attract one another more strongly than those of the same kind, the l-line bends downwards, since heat is now liberated in the mixing process. The G-x curve will therefore have a greater curvature than in the case of ideal mixing (fig. 1) and here also this will result, at $T > 0$ °K, in a continuous region of random mixing from A to B, and this appears to be stable even at $T = 0$ °K. However, at $T = 0$ °K the Nernst heat theorem requires zero entropy ($S = 0$), and this rules out a random building-up of matter; consequently fully ordered solid phases will also occur in this case, possibly in great numbers, and they will be separated by miscibility gaps, each of the solid phases having an exact composition (and hence no existence region) at $T = 0$ °K.

Fig. 4 illustrates schematically the relevant parts of the G-x curves for a binary system A-B with immiscibility in the solid state, where one compound is stable, the one with the stoichiometric composition AB. Fig. 4*a* shows the curves for $T = 0$ °K, and fig. 4*b* the curves for $T > 0$ °K. ΔG(AB) is the Gibbs free

energy of the formation of AB per gramme-atom. For $x < \frac{1}{2}$, A dissolves in AB, and for $x > \frac{1}{2}$, B dissolves in AB. This corresponds to the occurrence of a sudden change in slope in the G-x curve at $x = \frac{1}{2}$, $T = 0$ °K. For $T > 0$ °K this curve appears to have a simple minimum at the stoichiometric composition; in other words, $\partial^2 G/\partial x^2$ is positive in the neighbourhood of the compound AB.

This can be deduced as follows for the compound of stoichiometric composition $A_{1-\xi}B_\xi$. Suppose that in a particular composition region in the neighbourhood of $x = \xi$ the mixture contains a large amount of "solvent" with the composition $A_{1-\xi}B_\xi$ and a small amount of dissolved *A and *B, i.e. that n(*B) and n(*A) are much less than $n(A_{1-\xi}B_\xi)$. The composition x is then given by:

$$x = \frac{\xi n(A_{1-\xi}B_\xi) + n(*B)}{n(A_{1-\xi}B_\xi) + n(*B) + n(*A)}. \qquad (22)$$

If we write [*B] for

$$\frac{n(*B)}{n(A_{1-\xi}B_\xi) + n(*B) + n(*A)}$$

and similarly use [*A], we then have:

$$\mu(A) = \mu^0(*A) + RT \ln [*A], \quad . \quad (23)$$

$$\mu(B) = \mu^0(*B) + RT \ln [*B]. \quad . \quad (24)$$

Combining equations (11), (14) and (17) with (22)-(24) gives the following expression for the curvature of the G-x curve in the neighbourhood of ξ:

$$\left(\frac{\partial^2 G}{\partial x^2}\right)_{x \approx \xi} = \frac{RT}{(1-\xi)^2[*B] + \xi^2[*A]}, \quad . \quad (25)$$

which is positive for all values of $x \approx \xi$ (see fig. 4*b*).

It also follows from equations (11), (23) and (24) that:

$$\exp \frac{\mu(A_{1-\xi}B_\xi) - (1-\xi)\mu^0(*A) - \xi\mu^0(*B)}{RT} = [*A]^{1-\xi}[*B]^{\xi} = K. \quad . \quad (26)$$

Here the disorder equilibrium constant K is to a good approximation independent of the concentration because $\mu(A_{1-\xi}B_\xi)$ varies very little when the small concentrations [*A] and [*B] vary. At $x = \xi$ we have $\xi[*A] = (1-\xi)[*B]$, which, with equation (26) substituted into equation (25), gives:

$$\left(\frac{\partial^2 G}{\partial x^2}\right)_{x=\xi} = \frac{RT}{K\xi^{1-\xi}(1-\xi)^{\xi}}. \quad . \quad (27)$$

At the stoichiometric composition $x = \xi$ the curvature of the G-x curve increases as the degree of disorder decreases. Equation (26) shows that K goes to zero when T approaches 0 °K. The curvature of the G-x curve at the stoichiometric composition (equation 27) thus increases with decreasing T and is infinite at the absolute zero point (fig. 4*a*). At absolute zero a compound therefore has an exact stoichiometric composition. At temperatures above absolute zero a compound has an existence region.

The width of the existence region

The following factors determine the width of the existence region (see *fig. 5*). The existence region Δx is wide when:

1) The curvature of the curve G(AB) is small. This is the case, as has just been shown, when the disorder equilibrium constant K is large, that is to say when the native disorder is large (see figs. 5*a* and 5*c*).

2) The Gibbs free energy ΔG(AB) of the formation of AB is high — cf. figs. 5*a* and 5*b* (in general terms, when the free energy of formation of the compound is high compared with that of the neighbouring phases). *Fig. 6* shows the extreme case where $\Delta G(AB) < 0$, and yet the compound AB is nevertheless metastable.

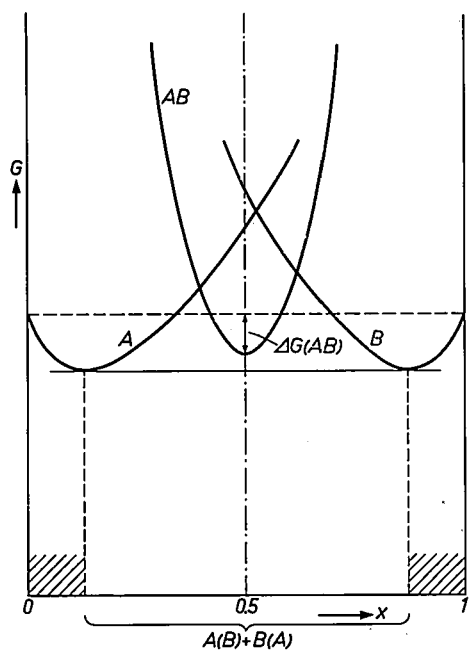


Fig. 6. The G-x curves for the special case where ΔG(AB) is negative yet the compound AB is nevertheless metastable.

3) The mutual solubility energy of A in B and of B in A is high — cf. figs. 5*c* and 5*d* (in general terms, when the solubility energy of A in the B-rich adjacent phase and that of B in the A-rich adjacent phase are high).

It is evident from the foregoing that, depending on the system, the width Δx of an existence region at temperatures above 0 °K can have all values between 0 and 1. This has been confirmed experimentally.

Examples are to be seen in *Table I*, where the order of magnitude of Δx is given, increasing from left to right, going from insulators and semiconductors to metallic alloys or compounds.

Table I. Order of magnitude of the existence region Δx of some binary compounds and alloys [3], indicated in atomic fractions.

CdTe	PbS	SnS	PbTe	MnTe	Cr_3S_4	SnTe	MnSb	CuAu
10^{-5}	10^{-4}	10^{-4}	10^{-4}	10^{-3}	10^{-2}	10^{-2}	10^{-1}	1

The fact that insulators and semiconductors usually have a narrow existence region and metallic compounds a wide one will be discussed in more detail in part III.

General form of the existence region (cross-section)

The phase diagram shown in *fig. 7* illustrates schematically the general shape of the existence region of AB at a given pressure P between $T = 0$ °K and the congruent melting point T_{congr}, for the binary system A-B with one solid compound AB, a miscibility gap in the solid phase and complete mixing in the liquid phase. This T-x cross-section at constant pressure of the three-dimensional P-T-x figure can be constructed from a series of G-x curves for the solid phases A, AB and B (see fig. 4), and the G-x curves of the liquid phase (*liq*), at a given pressure and at a series of different temperatures. The compound AB is stable at the given pressure P between the temperatures $T = 0$ °K and $T = T_{congr}$. Between these temperatures the width of the existence region Δx has a finite value. At $T = 0$ °K the width of the existence region is zero (equation (27) and fig. 4*a*) and the compound is exactly stoichiometric and completely ordered (Nernst heat theorem). The width Δx is also zero at T_{congr}, but here the native disorder is at a maximum (eq. 26). The boundary of the existence region is a continuous line, except at the eutectic temperatures T_{e1} and T_{e2}. Owing to the asymmetry of the G-x curves (fig. 4) the existence region is asymmetric with respect to the stoichiometric composition and, except at $T = 0$ °K, it may even lie completely outside it. Because of this, congruent melting also occurs outside the stoichiometric composition, while the exactly stoichiometric AB has a melting range at temperatures below T_{congr}.

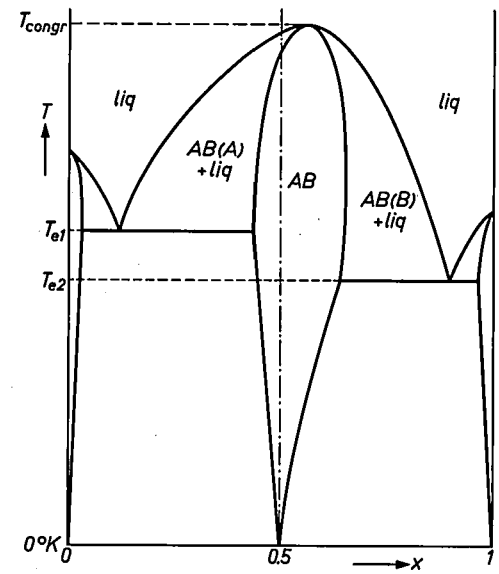


Fig. 7. Cross-section at constant pressure of the three-dimensional P-T-x phase diagram of a binary system A-B in which a stable compound AB is present, there is a miscibility gap in the solid phase and the liquid phase is fully miscible. T_{e1} and T_{e2} are eutectic temperatures, and T_{congr} is the congruent melting point.

The above properties of the existence region follow from general thermodynamic considerations. They therefore apply to all organic and inorganic liquid and solid phases.

Summary. This article, the first of a series of three, gives a rather detailed discussion of some of the questions concerned with stoichiometry and the existence region of compounds. It is shown that each compound has an existence region of a certain width at $T > 0$ °K, and a general indication is given of the factors determining the width of the existence region and its general form.

[3] The data given in the table were obtained from the following publications:
for CdTe: D. de Nobel, Philips Res. Repts. **14**, 361, 430, 1959;
for PbS: J. Bloem and F. A. Kröger, Z. phys. Chemie Neue Folge **7**, 1, 1956 and W. Albers, C. Haas and H. J. Vink, Philips Res. Repts. **18**, 372, 1963;
for SnS: H. Rau, J. Phys. Chem. Solids **27**, 761, 1966;
for PbTe: R. F. Brebrick and E. Gubner, J. chem. Phys. **36**, 1283, 1962;
for MnTe: J. van den Boomgaard, to be published shortly in Philips Res. Repts.;
for Cr_3S_4: F. Jellinek, Acta cryst. **10**, 620, 1957;
for SnTe: R. F. Brebrick, J. Phys. Chem. Solids **24**, 27, 1963;
for MnSb: Ch. Guillaud, Ann. Physique **4**, 671, 1949 and I. Teramoto and A. M. J. G. van Run, J. Phys. Chem. Solids **29**, 347, 1968;
for CuAu: N. S. Kurnakov and S. F. Zemczuzny, Z. anorg. Chemie **54**, 158, 1907.

PHILIPS TECHNICAL REVIEW

VOLUME 30, 1969, No. 4

Crystal growth and chemical synthesis under hydrothermal conditions

A. Rabenau and H. Rau

Nuggets of gold have recently been grown in the laboratory for the first time. They are grown under conditions which are believed to reproduce those that occur in nature, and the process has provided deeper knowledge of the way in which gold deposits are formed under natural conditions. This is in fact an offshoot of a study, made at the Philips Aachen laboratory, of crystal growth and chemical synthesis under hydrothermal conditions. Such processes can be used to prepare a great diversity of materials of importance in solid-state physics.

The hydrothermal method of crystal growth and hemical synthesis is modelled on nature. In many arts of the world, deposits of minerals that under normal conditions are insoluble in water have been formed in the presence of water at high temperature and under high pressure.

In geology and mineralogy a distinction is made between hydrothermal and pneumatolytic conditions: the first are found below the critical temperature and the second above it. Apart from the fact that there is a gen-

rof. Dr. A. Rabenau (deputy director) and Dr. H. Rau are both with he Aachen laboratory of Philips Zentrallaboratorium GmbH. rof. Rabenau is also a Visiting Professor at the Rhine-Westphalia echnical University at Aachen.

eral lack of data on the critical point, the solubility near this point seems to show no discontinuity. In the laboratory hydrothermal conditions are therefore taken to refer to an aqueous environment above 100 °C and above atmospheric pressure. In the past most of the attempts to reproduce the hydrothermal process of mineral formation were made by mineralogists and geologists. They enclosed the materials in a reaction vessel (an autoclave), which was heated to the required temperature and in which the pressure depended on the extent to which it was filled.

During the second world war the problem became highly topical because of the need for oscillator crystals, particularly in Germany and the United States. Crystals of the low-temperature modification of quartz were required, and these had to be prepared hydrothermally since the demand far exceeded the supply of natural crystals. Since that time an increasing amount of work has been done on the hydrothermal growth of crystals, and the number of materials that have been made in this way is already over a hundred. The hydrothermal method is no longer the exclusive property of mineralogists and geologists, but has now found its way into solid-state physics. The merits of the method account for this very readily: crystals can be grown hydrothermally in the laboratory to standards of purity and perfection that are seldom matched in nature. The substitution of related elements (isomorphous substitution) enables homologous compounds to be obtained, and it is even possible to synthesize new compounds that have no natural analogues.

Conversely, the experience gained in solid-state research had yielded an explanation for a problem that has long puzzled geologists, the problem of how gold deposits are formed in nature. At the end of this article we shall deal with this point in more detail.

The conventional method [1]

The special aspects of the hydrothermal method stem directly from the physico-chemical properties of the solvent water at the applied temperature and pressure. *Fig. 1* shows the *p-T* diagram of pure water, with the filling factor (degree of fill) of the autoclave as a parameter. Since the filling factor is defined as the ratio of the volume of liquid and reaction vessel (at 20 °C) it is also correlated with the "total density" (above the critical point this is the density of the fluid state). With the hydrothermal method the filling factor is usually between 50 and 80%, and the pressure between 200 and 3000 bars. Under these conditions the value of the dielectric constant is between 10 and 20, which means that water under hydrothermal conditions is still to be regarded as a polar solvent. Another very important point is that the ion product of water (i.e. the product of the concentration of the H^+ and OH^- ions into which the water has partly dissociated) increases with rising temperature and density, and at 600 °C and 2000 bars it is about 10^5 times greater than at room temperature. These conditions therefore give rise to hydrolytic reactions.

Many substances that are not soluble in water at room temperature and atmospheric pressure will show increased solubility under these conditions. Often, however, the solubility may still not be sufficiently high: for crystal growth and synthesis it should not in practice be less than about 1%. The usual procedure is therefore to make the environment alkaline by adding NaOH or Na_2CO_3, which causes many substances to dissolve sufficiently by the formation of complex ions.

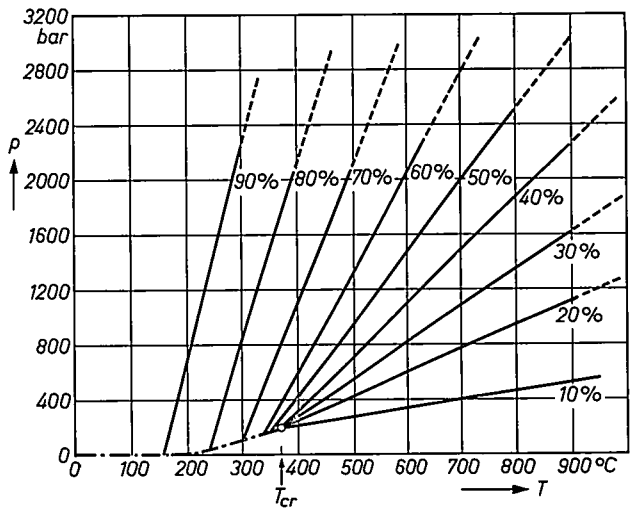


Fig. 1. In the hydrothermal method crystal growth and chemical synthesis take place in an aqueous solution at high temperature and pressure in a closed vessel (an autoclave). The diagram relates to pure water, and shows pressure as a function of temperature, with the filling factor (degree of fill) of the autoclave as parameter. The filling factor is usually between 50 and 80% and the pressure between 200 and 3000 bars. The chain-dotted line is the equilibrium line of vapour and liquid, T_{cr} is the critical temperature.

The solution equilibrium is in general temperature-dependent. This dependence is made use of by placing the material in a temperature gradient, so that it is transported from places with a high solubility to places with a low solubility [2]. This results in the growth of relatively large crystals. Moreover, a seed crystal can be placed in the low solubility region, as in growing of quartz crystals (*fig. 2*).

As a result of the low viscosity of the hydrothermal solution and of the strong convection due to the temperature gradient the dissolved substance is transported rapidly and consequently the reaction and growth rate are generally high with this method.

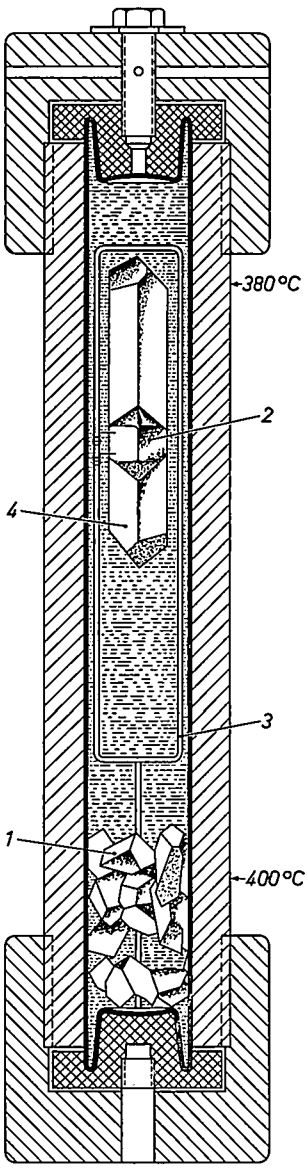


Fig. 2. Autoclave for growing quartz crystals in an alkaline solvent (1-molar NaOH). With a pressure of about 1500 bars and a temperature gradient of 400→380 °C, material is transported from the mother liquor *1* to the seed crystal *2*. *3* frame to which the seed crystal is fixed. *4* grown crystal. The cap is provided with a safety device. (After A. C. Walker and E. Buehler, Ind. Engng. Chem. **42**, 1369, 1950.)

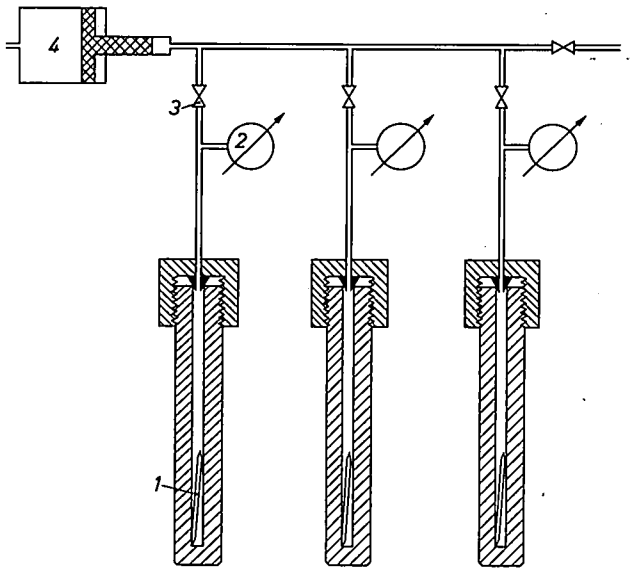


Fig. 3. "Tem-Pres" apparatus used for preliminary experiments on small quantities of material. The autoclaves are made of steatite. The reaction vessel, which must be resistant to hot alkaline solutions, is a sealed gold capsule *1*, which may have a volume between 0.25 and 2 cm^3. *2* pressure gauge. *3* valve. *4* pressure intensifier.

The autoclave must be made of a special type of steel to withstand the high temperature and pressure to which it is subjected. To protect it from the corrosive action of alkaline solutions the steel often has to be coated with noble metals such as gold, silver or platinum.

For preliminary investigations on small quantities of material it has been found that sealed gold capsules can be used (*fig. 3*). In the equipment illustrated here a pressure intensifier is used to provide an external counter-pressure on the capsules, which would otherwise burst as a result of the internal pressure.

The relatively high cost of the equipment and the complicated procedure limit the usefulness of the method, however elegant it may be. We shall now describe a method developed at the Philips Aachen laboratory which does not have these disadvantages.

Crystal growth and chemical synthesis in an acid solvent

The conventional hydrothermal method has another disadvantage which we have not yet mentioned. It is very suitable for oxidic compounds, in particular quartz and silicates, but not very suitable for metallic compounds of sulphur, selenium and tellurium, which are of particular interest in modern solid-state research. This is rather odd, since there are many natural examples of sulphide ores formed under hydrothermal conditions. These include sulphides which are among the less soluble compounds. The usual method of increasing the solubility by working in an alkaline solution does not — apart from a few chalcogenides of zinc and cadmium — lead to the desired result since hydrolysis takes place. An attempt to use solutions containing large quantities of sulphur as solutions of alkali polysulphides, thiosulphates, etc., in the hope of forming sulphide complexes, was equally unsuccessful. Quite apart from the practical difficulties in these experiments, due to the highly corrosive action of such solutions, the results were disappointing.

A direct incentive to look for other methods arose in the Philips Aachen laboratory when it became necessary to grow crystals of the mineral CuS. We then investigated whether concentrated *hydrohalic acids* (HCl, HBr, etc.) could be used as solvents. There

[1] See A. A. Ballman and R. A. Laudise, in: The art and science of growing crystals, editor J. J. Gilman, Wiley, New York 1963, p. 231.

[2] See A. Rabenau, Philips tech. Rev. **26**, 117, 1965.

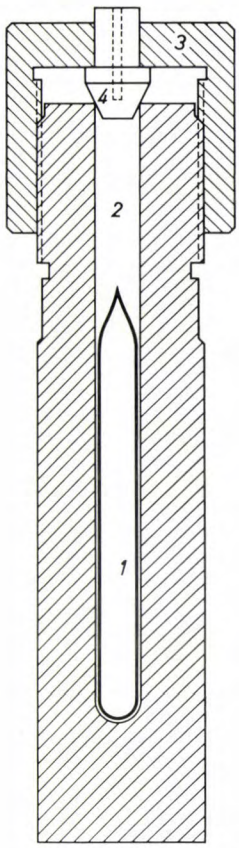


Fig. 4. Equipment for the hydrothermal method developed in the Philips Aachen laboratory, using hydrohalic acids as solvent. The sealed quartz ampoule *1* (see also fig. 5) is filled with the material to be converted and the solvent. A calculated quantity of solid CO_2 is added to the remaining volume *2* of the autoclave to build up a pressure high enough to prevent the ampoule from being exploded by the internal pressure. *3* screw cap. *4* sealing cone with hole for thermocouple.

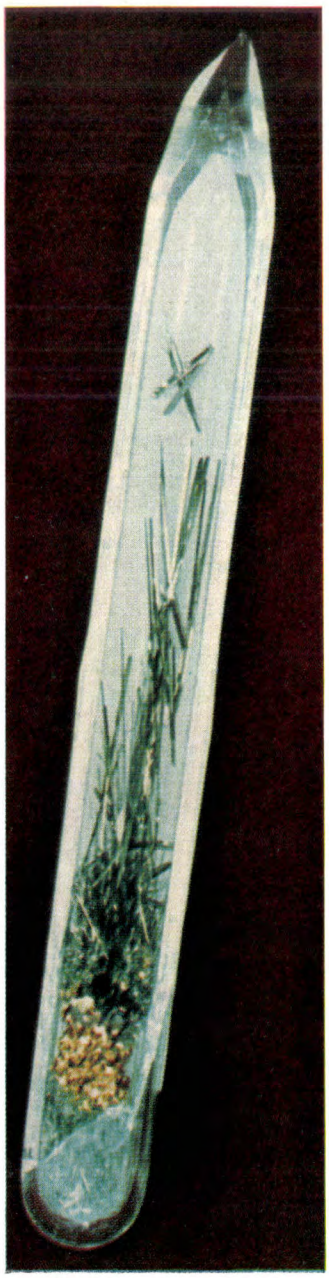


Fig. 5. Quartz-glass ampoule in which crystals of $AuITe_2$ together with gold have been grown. This result was obtained by bringing together gold and tellurium in equal numbers of gramme-atoms in 10-molar HI at a temperature of 450 °C and then allowing it to cool for 10 days to 150 °C. (Actual scale.)

were two reasons for this: the tendency of sulphides to dissolve in acids with the formation of H_2S, and further the tendency of heavy metals to form complexes with halogen ions. The first tendency can be understood from the fact that an increase in the hydrogen-ion concentration gives a decrease in the activity of the sulphide ions. (With reasonably soluble sulphides gaseous H_2S may in fact be liberated, but this does not matter since the materials are part of a closed system and continue to take part in the equilibrium.)

These investigations have considerably widened the potential applications of the hydrothermal method. Moreover, the technique of the method in its new form is so much simpler that it can quite easily be employed in small laboratories [3].

An important point is the effect of the acid solution on quartz glass. At temperatures up to 500 °C and pressures up to 3000 bars quartz glass is virtually unaffected, even in the presence of free halogen, sulphur, selenium, etc. This means that the experiments can be carried out in sealed ampoules of quartz glass [4]. These ampoules can withstand exceptionally high external pressures and consequently the danger of explosion due to the (generally unknown) internal pressure developed during the heating can be ruled out by simply ensuring that the ampoules are subjected to a sufficiently high external pressure say 2500 bars at the working temperature. However great the internal pressure then becomes, provided it remains lower than the external pressure the ampoules will not explode.

Some further details of our method can be seen in *figs. 4* and *5*. A quartz-glass ampoule with a diameter of about 15 mm is filled up to between 50 and 70% with one of the hydrohalic acids. (These values are determined empirically.) The acid is frozen with the aid of liquid air, and the substances to be transformed are then introduced into the ampoules. Next, the ampoule is evacuated, sealed and placed in the autoclave. An autoclave for the temperatures and pressures quoted above can be made of low-alloy steels of the type DIN 24CrMoV55 (0.45 Mn; 1.35 Cr; 0.55 Mo; 0.20 V), which can be machined by conventional methods. The volume in the autoclave after deducting the volume taken up by the ampoule is then measured, and as much solid CO_2 is added as is needed to obtain a sufficiently high external pressure at the reaction temperature. The required quantity of solid CO_2 is cut from

[3] H. Rau and A. Rabenau, Mat. Res. Bull. **2**, 609, 1967.

[4] Up to a temperature of about 400 °C borosilicate glasses can also be used for this purpose.

[5] The chemical structure is not well expressed by the formula CuS. The crystal structure shows that the compound in question should be described as $Cu_2^{+}S^{2-}Cu^{2+}(S_2)^{2-}$.

[6] See J. L. Meijering, Philips tech. Rev. **26**, 52, 1965, in particular page 54 and fig. 6.

a bar of suitable diameter. The autoclave is then sealed by means of a screw cap and placed in an electric furnace. As in the conventional method, a temperature gradient is produced, causing material to be transported in a direction — from cold to hot or *vice versa* — that depends on the sign of the reaction enthalpy.

The upper left-hand sample in fig. 6 (crystal *1*) is a CuS crystal [5], grown in 9-molar HBr. This crystal was grown at a relatively low temperature, namely below the peritectic [6] temperature 508 °C, because CuS would melt at this temperature and decompose. For the homologous CuSe we had to keep the temperature below 379 °C for the same reason. In view of the relatively low temperature required, any other method of growing these crystals would be barely feasible.

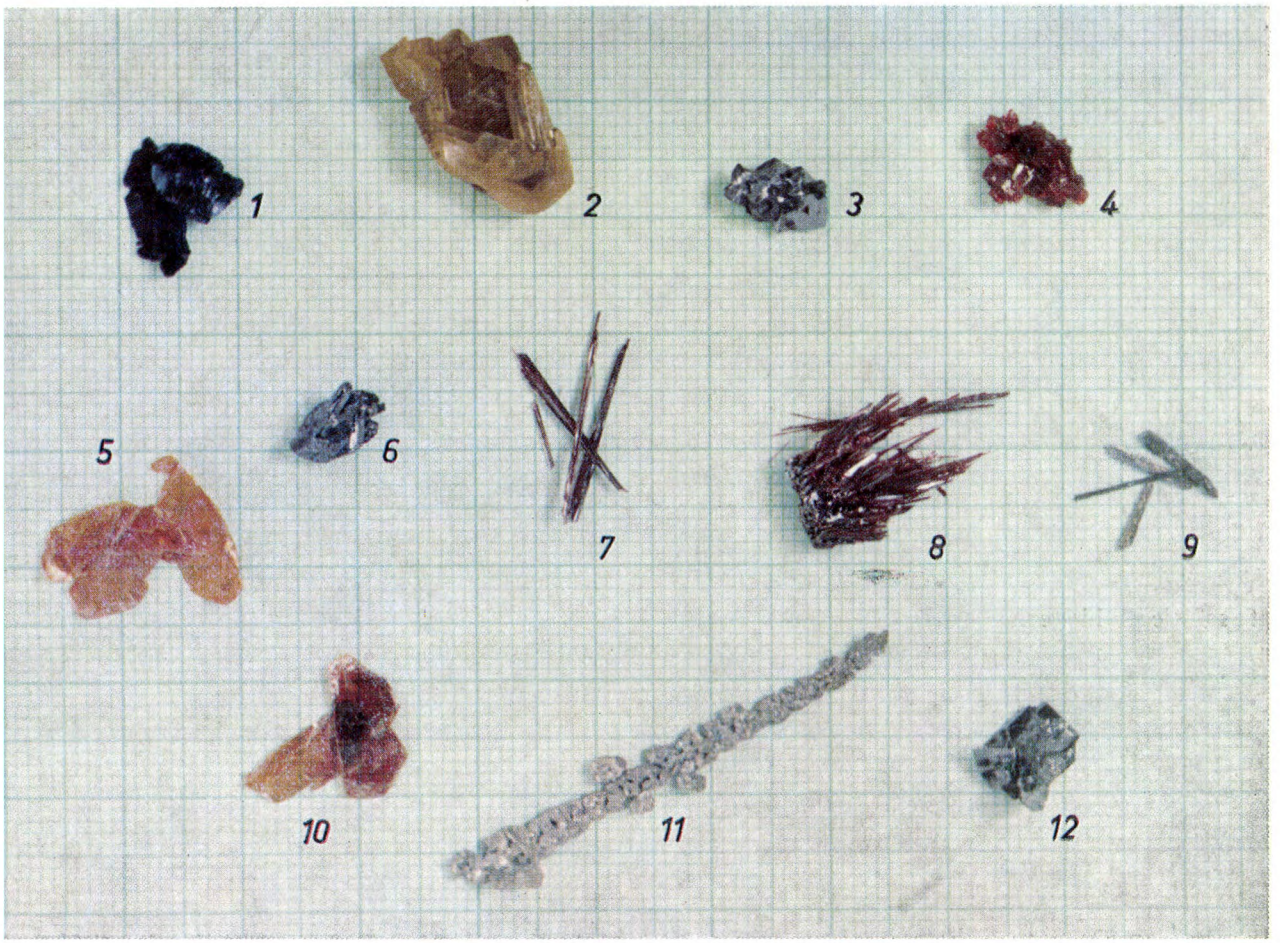


Fig. 6. Crystals grown by the new hydrothermal method. The name of each compound is followed by the solvent used and, where applicable, the temperature interval.
1 CuS, from 9-molar HBr, 450-420 °C.
2 CuI with zinc-blende structure; the crystal precipitated from 2-molar HI on slow cooling from 365 °C to 170 °C.
3 PbSe, from 9-molar HBr, 350-335 °C, in the presence of $PbBr_2$.
4 CdS, from 9-molar HBr, 450-430 °C.
5 GeI_2, after a few days grown from Ge in 10-molar HI at 480 °C.
6 Sb_2S_3 (stibnite), from 9-molar HBr, 455-440 °C.
7 BiSBr, formed by the conversion of Bi_2S_3 in 9-molar HBr at 500 °C and slow cooling to 200 °C.
8 $Pb_5S_2I_6$, carmine red needles formed by the conversion of PbS in 10-molar HI at 450 °C and slow cooling to 200 °C.
9 tellurium crystals in the form of needles, formed by the conversion of PbTe in 10-molar HI at 450 °C and cooling to 200 °C.
10 PbI_2, often obtained as a by-product in the conversion of lead chalcogenides in 10-molar HI.
11 platinum, from 10-molar HI, 480-500 °C, in 10 days.
12 SnSe, formed in 10-molar HI on cooling from 500 °C to 300 °C in 10 days.

Applications

We shall now present a selection from the many kinds of crystal that have been grown by the new method (*fig. 6*). These include materials which have never previously been synthesized. The selection is also intended to give an idea of the wide variety of properties shown by the crystals.

The CuS compound is a metallic conductor and becomes superconducting below 1.67 °K; CuS is also noted as the mineral with the greatest optical dispersion ever observed.

We have just touched upon one of the important advantages of the method, the possible establishment of subsolidus equilibria because of the relatively low temperature at which crystal growth and synthesis can take place. This also played a part in growing CuI crystals with a cubic zinc-blende structure (*2* in fig. 6). This modification is stable only below 369 °C. Above this temperature the modification changes to a hexagonal wurtzite structure, and between 407 °C and the melting point at 605 °C another cubic modification is stable. The solubility of this compound in 2-molar HI is so great and so highly temperature-dependent that in this case crystals can be obtained by the simplest method, i.e. by the cooling of a saturated solution.

Low growth temperatures are also important for PbSe and CdS (*3* and *4* in fig. 6). Both can also be grown by other methods, but growth at a relatively low temperature can give crystals of greater perfection.

CuI has attracted some attention because of its opto-electronic properties. PbSe is used in infra-red detectors and CdS in photoelectric cells.

Chalcogen halides

In some experiments with sulphides it was observed that the sulphide itself was not transported but that, at a sufficiently high concentration of the hydrohalic acid a reaction takes place in which a thiohalide is formed. For example, in 9-molar HBr stibnite, Sb_2S_3, grows in the form of beautiful crystals (*6* in fig. 6), but in 10molar HI, the compound SbSI is formed, in accordance with the equation:

$$Sb_2S_3 + 2\ HI \rightleftarrows 2\ SbSI + H_2S.$$

In reactions of this type, liquid H_2S has been formed as a separate phase in capsules cooled to room temperature. Similarly BiSBr is formed from Bi_2S_3 (*7* in fig. 6). This group of substances, in particular SbSI, has recently been the subject of numerous investigations. SbSI combines ferroelectric with photoconductive properties.

In a similar way γ-$Hg_3S_2Cl_2$ crystals are formed starting from HgS in concentrated hydrochloric acid.

In growing PbS from hydrohalic acids the formation of red crystals has been observed, which upon analysis have been found to be thiohalides of lead. Compounds of this type have been known in the literature for more than a hundred years. They are familiar in particular as the bright red precipitate obtained when acid lead-salt solutions are allowed to react with H_2S. Until recently, however, the chemical composition of these compounds was still uncertain. Owing to their chemical instability it was not possible to isolate them in a pure form. The hydrothermal method has helped to provide the answer here, leading directly to isolated single crystals which can be investigated chemically and by X-ray analysis.

Two lead thiohalides were obtained: $Pb_5S_2I_6$ (*8* in fig. 6) and $Pb_7S_2Br_{10}$. The red colour of the crystals corresponds to an energy gap of about 2 eV at room temperature. In a recent study of pseudobinary systems that form PbS with $PbCl_2$, $PbBr_2$ and PbI_2 it was found that the two lead thiohalides are the only two stable ternary compounds occurring in these systems.

We took this as an indication that our hydrothermal method was especially useful for synthesizing chalcogen halides. We made use of this to investigate whether such compounds exist of the transition elements of the first group of the periodic system: Cu, Ag and Au. Until recently the only compounds of this kind that had been described were the silver compounds Ag_3SI and Ag_3SBr, which have a perovskite-like structure. It is a remarkable thing that no references to the existence of corresponding copper compounds are to be found in the literature. In fact, both the chalcogenides and the halides of copper are stable, and the transport properties of these compounds have been extensively studied. The application of our hydrothermal method resulted in the discovery of both copper and gold compounds (*fig. 7*). All of the new compounds are located at cuts [7] of the ternary phase diagram that run from the metal halides M^IX^{VII} to selenium or tellurium. We were able to prove that the copper systems are pseudobinary systems. One might be inclined to deduce

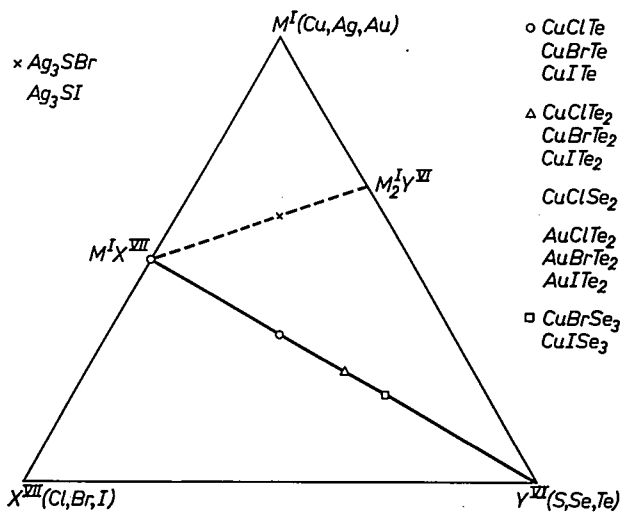


Fig. 7. Ternary systems with the following components: the transition metals of the first group Cu, Ag, and Au, the chalcogens S, Se, and Te, and the halogens Cl, Br, and I. They are indicated as M^I, Y^{VI} and X^{VII} respectively. The copper and gold compounds formed in these systems are found to occur on cuts running from the composition of the metal halides M^IX^{VII} to that of the chalcogens Y^{VI}. The silver compounds indicated on the left were until recently the only known stable compounds of these systems. It has since been found possible to synthesize the compounds named on the right by hydrothermal synthesis in an acid solution.

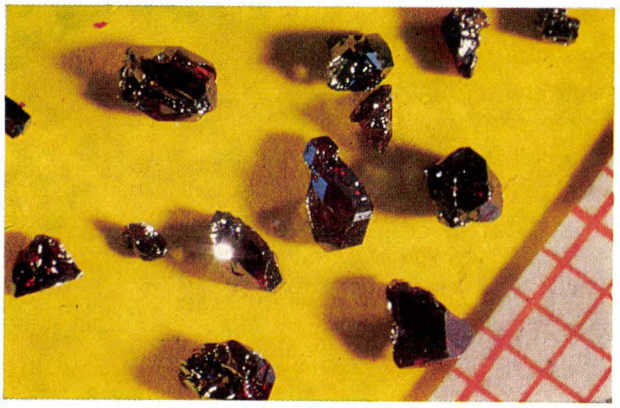


Fig. 8. Crystals of $CuISe_3$ obtained from CuI and Se in 5-molar HI, after cooling from 390 °C to 200 °C.

from the diagram that the copper and gold compounds show a greater resemblance to one another than to the silver compounds. It turned out, however, that the copper and gold compounds differ considerably from each other in physical properties: whereas the copper compounds (*fig. 8*) are semiconductors or insulators, the bright metallic needles of the gold compounds (fig. 5) exhibit metallic conductivity.

Gold

The hydrothermal method in an acid solution was also found to be suitable for growing crystals of elements. For example, crystals of tellurium (*9* in fig. 6) of platinum (*11* in fig.6) and of arsenic (*fig. 9*) were grown in this way from solutions of HI.

In the following we shall deal more in detail with the growing of gold crystals [8]. This brings us to a problem that has long preoccupied geologists. They have assumed that the greater part of the primary gold strata originated under hydrothermal conditions [9]. In the laboratory, however, as we noted earlier, gold had previously only been used as a material for coating the autoclave to protect it against alkaline corrosion. This indicates that it is unlikely that the gold strata originated under such conditions.

In an earlier stage of our experiments we had already noticed that HI is an excellent solvent for gold. When 10 grammes of gold, in the form of wire or foil, are kept in 10-molar HI at a temperature of 500 °C (with no temperature gradient), then in 10 days a group of gold crystals like the one shown in *fig. 10* is formed. If a temperature gradient is applied, for example from 480 to 500 °C, the gold is transported from cold to hot regions, giving crystals like the ones in the title photograph.

In our further investigations we tried to find an answer to the following two questions.:

1) Can gold transport of this kind also take place under conditions actually found in nature?
2) Is transport from cold to hot regions a reasonable assumption for these conditions?

The second question is important because all previous discussions and experiments on this subject have started from the assumption that gold is a deposit formed in the course of time from a gradually cooling solution.

It has been found that gold is also transported in solutions containing hydrogen chloride and bromide, provided an oxidizing agent is present, such as free halogen or oxygen. Where the solvent is HI, the free halogen is formed by dissociation from the solvent itself.

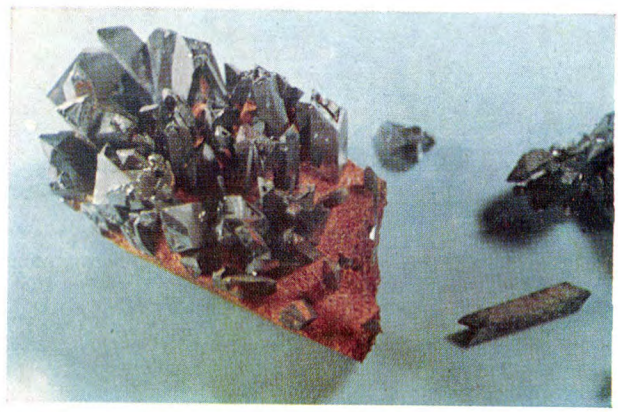


Fig. 9. Arsenic grown together with AsI_3; obtained by heating arsenic powder in 10-molar HI at 450 °C for 9 days. (Twice full size.)

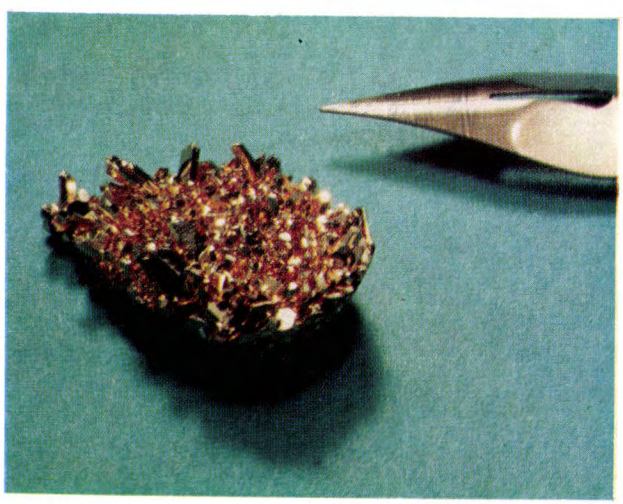


Fig. 10. Group of gold crystals of about 10 grammes, obtained by heating gold (without temperature gradient) in 10-molar HI for 10 days at 500 °C. The title photograph shows gold crystals grown in a temperature interval of 480→500 °C.

[7] See J. L. Meijering, Philips tech. Rev. **27**, 213, 1966, in particular pages 214 and 219.
[8] A. Rabenau and H. Rau, Naturwiss. **55**, 336, 1968 (No. 7).
[9] See K. B. Krauskopf, Introduction to geochemistry, McGraw-Hill, New York 1967.

A current view of geologists is that natural gold deposits must have been deposited at temperatures between 50 and 550 °C and at pressures up to 2000 bars [9]. In natural conditions, weakly acid salt solutions are available as solvent, with rock salt as the main constituent, in concentrations greater than 2-molar. It has been established by experiment that gold is

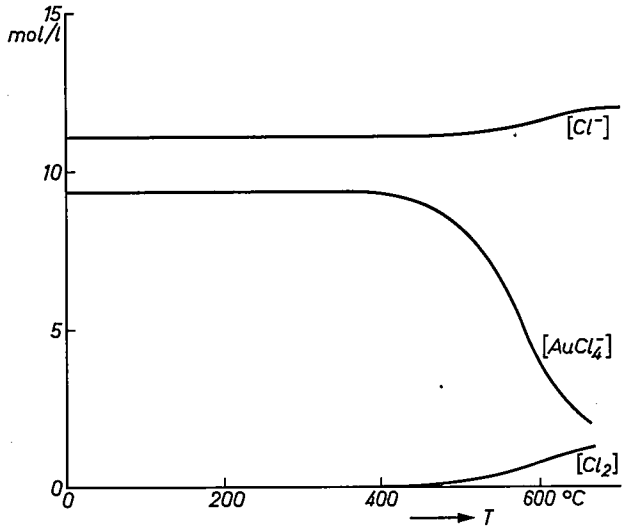


Fig. 11. Calculated concentration-temperature diagram, based on a solution equilibrium which we assume to occur in the hydrothermal growth of gold. The complex ion $AuCl_4^-$ is significant here (see text). Gold transport from cold to hot could take place in the temperature range in which the concentration of $AuCl_4^-$ shows marked variations. (For clarity the concentration of $AuCl_4^-$ is shown on a scale 10 times larger than the calculated value.)

in fact transported in weakly acid 10-molar NaCl solutions. The oxidizing agent in this case was oxygen, added in the form of H_2O_2.

We now assume that the transport of gold is based on the following equation:

$$Au\ (solid) + 1.5\ Cl_2\ (gas) + Cl^-\ (aq) \rightleftarrows AuCl_4^-\ (aq).$$

$AuCl_4^-$ is a fairly stable complex, and its thermodynamic data are known. These data relate to normal conditions. The values applicable to hydrothermal conditions, obtained by extrapolation, must of course be regarded as a very rough approximation. *Fig. 11* shows the result of a calculation based on extrapolated values. At low temperature there is virtually no free halogen present. As the temperature rises, the complex dissociates. In the region in which the concentration of the complex varies strongly as a function of temperature we may expect a transport of gold from cold to hot. Considering the approximation used, the agreement between experiment and calculation may be regarded as satisfactory.

Summary. Problems connected with materials in solid-state research are increasingly being solved with the aid of the hydrothermal method, in which materials are converted in aqueous solutions at temperatures above 100 °C and pressures above atmospheric. This method, originally borrowed from nature permits the crystal growth and chemical synthesis at relatively low temperatures. Substances that are not readily soluble under normal conditions can often be sufficiently dissolved in an alkaline solvent, resulting in the formation of complex ions. Experiments are done by placing the material in a temperature gradient, which causes materials to be transported from places of high solubility to places of low solubility. The method is mainly used for oxidic compounds, quartz and silicates in particular. The expense of the equipment and the complicated procedure is a bar to its wider application. At Philips Aachen laboratories a hydrothermal method has now been developed which uses concentrated *hydrohalic acids* as solvent. This method requires relatively simple equipment, and it is suitable for crystal growth and synthesis of sulphides, selenides and tellurides, which can rarely be grown in alkaline solutions. The method is particularly suitable for synthesizing chalcogen halides. It has also been used for growing crystals of elements such as tellurium, arsenic, platinum and gold. Some details are given of the growth of gold crystals in connection with a hypothesis on the origin of natural gold deposits.

Prospecting with neutrons

A. W. Wylie and P. L. Eisler

In recent years, pioneering work in the development of very compact and reliable sealed-off neutron generator tubes has been carried out by Philips Research Laboratories. An interesting application of such neutron tubes is the "logging" of bore holes for the detection of useful minerals. The state of this rapidly developing field is reviewed and current experimentation is described in this article.

General introduction to nuclear methods of bore-hole logging

The principles of gamma spectrometry

The most generally applied method of prospecting for minerals is diamond-core drilling: cores of the layers cut through in the drilling are retrieved and examined in the laboratory. Every year in countries with an active mining industry many thousands of metres of drill holes are sunk in this way to explore the extent of mineral occurrences. In Australia in 1967 the total length of diamond-core drilling exceeded 520 000 metres; in Canada in 1965 the corresponding length amounted to 2 500 000 metres [1].

The cost of core drilling considerably exceeds that of simple percussion drilling. Whereas in Australia the latter costs from 1 to 5 Australian dollars per foot, the former will cost from 7 to 18 dollars per foot, depending on depth, nature of rock and locality. In addition, percussion holes can be put down more speedily. These advantages make it very attractive to replace the time-honoured technique of core drilling by a method based on percussion holes and "logging", which means that a sensing instrument is lowered into the hole to provide data on the nature of the strata cut through. Logging of course will also cause some extra cost but the total operation will remain much cheaper than core drilling.

The problem of collecting sufficient down-hole information to guide the mining technologist is considerable in view of the restricted size of the holes and the nature of the environment. The latter is hostile to the performance of precise physical measurements. Thus the holes are frequently full of water to within a few metres from the surface. Water pressure increases progressively with depth and amounts to about 300 N/cm^2 (450 lbs./sq. in.) at 300 m. The sensing instrument must be robust enough to withstand both this pressure and the mechanical buffeting received when it is wound up and down rough-walled holes on the end of a cable between a hundred to more than one thousand metres in length. In addition, it must be able to cope with temperatures well above those at the surface.

The procedure adopted and discussed in this article consists in lowering down the drill hole a source of radiation which causes the material of the walls to emit signals (secondary radiation) carrying information picked up by a detector mounted on top of the source. Very penetrating gamma radiation produced in nuclear processes is chosen as the most appropriate carrier of signal information, since softer radiations are largely absorbed by the walls of the logging tool, if not by the water or slurry surrounding it, and therefore are useless for a general analytical method.

To excite gamma rays of sufficient intensity strong irradiation of the rock matrix is essential. Fortunately, compact sources of fast neutrons capable of generating gamma rays in substantial volumes of rock are now available. It is the spectrometry of this gamma radiation, produced by a source of fast neutrons drawn past the various formations, that forms the main basis of element analysis. The fact that different types of available sources will excite different nuclear processes giving rise to various gamma emissions (page 100)

Dr. A. W. Wylie and P. L. Eisler are with the Division of Mineral Chemistry of the Commonwealth Scientific and Industrial Research Organisation (P. O. Box 124, Port Melbourne, Victoria 3207), Australia.

[1] Statistical Bulletin No. 3, Mineral Exploration, Commonwealth Bureau of Census and Statistics, Canberra, 1967. J. W. McBean and G. H. Charlewood, Canadian Min. metall. Bull. **60**, No. 657, 63, 1967.

adds to the appeal of the method, since each type of excitation has its own advantages for particular elements and for particular situations.

Although the element analysis achieved will often be less quantitative in the geophysically complicated environment of a bore hole than in the favourable "linear geometry" situation that applies when cores are analysed in the laboratory, compensating advantages are to be expected from the greater volume of rock sampled and from the broad picture of element distribution which can be obtained as the log goes down the drill hole. Direct computer processing of logging data, besides greatly facilitating interpretation of the information gathered, will, moreover, permit its ready storage and rapid comparison with information from other drill holes.

Brief historical survey

For a number of years the oil industry has used nuclear techniques to supplement other methods of logging oil wells. Early work [2] began with counting the mixed radiations from naturally radioactive elements, and with measurements of the intensity of scattered gamma radiation, both of these logs being made essentially to explore the stratigraphy of sedimentary formations [3]. Energy analysis of neutron-induced gamma radiation was not feasible until the invention of suitable detectors and was limited in its early achievement by inferior equipment. By 1958-1963, the radiative capture spectra of H, Si, S, Cl, Ca and Fe generated in various formations by neutrons from a radioactive source were recorded with the progenitors of the modern pulse-height analyser [4]. Resolution in these spectra, however, still fell short of that required for general application of the technique.

The advent, in 1957-1960, of the pulsable 14 MeV neutron generator [5] (*fig. 1*) improved the prospects for better spectral resolution in gamma spectrometry, as will be discussed below, and in oil-well logging enabled the radiation peaks due to inelastic scattering by C, O and Mg to be detected under bore-hole conditions [6]. But with the single exception of iron, no element heavier than calcium (atomic number 20, atomic weight 40) was reported in these investigations, for the sufficient reason that the "light" elements are the elements most characteristic of the geological structures of interest in oil technology.

The geological situation normally encountered in the mining industry is substantially different and more complex. Instead of an often considerable depth and relatively uniform composition of sedimentary limestones and sandstones for example, very variable lithology may be encountered, especially in the vicinity of an ore body. The ore body itself may be complex in composition and more complex in its interactions with neutrons than sedimentary strata. Moreover, the number of elements of geological interest or economic significance is not only large but is rapidly increasing. Detailed spectral analysis of a variety of mainly heavy elements is therefore likely to be needed in each specific case, at least with a considerable proportion of the hole or holes drilled, and multichannel analysis seems essential in view of the complexities of the spectrum. No such work has been attempted by the industry itself in Western countries. The U.S.S.R., however, notably through the work of the Ghubkin Institute and the All-Union Research Institute of Nuclear Geophysics and Geochemistry in Moscow, have trained many nuclear and other well-logging specialists and have many field units engaged in the search for minerals. Both activation [7][8] and radiative capture processes [8][9][10], discussed below, have been utilized in these investigations.

Some general aspects of gamma spectrometry used in logging

The available neutron sources will be discussed presently, and so will the detectors lowered down the drill hole for analysing the gamma radiation obtained. A few preliminary remarks may however be useful.

The down-hole equipment must be capable of transmitting the detector pulses over considerable distances without degrading their shape, so that the pulses can be accurately sorted at the surface by a multichannel pulse-height analyser. Alternatively, for example with very long cables, the down-hole equipment should be able to transmit digital information about pulse heights to the surface instrumentation. When using different detectors in one rig, multiplex operation may also be desired. In each case, the result is a digital read-out of intensity (count rate) versus channel number, i.e. gamma photon energy (after calibration of the analyser). This form of read-out is very suitable for handling the large amount of information liable to be acquired in logging a drill hole.

The power of the spectroscopic technique is clearly determined by the resolvability and by the recording rate of the spectrum excited. Whereas the performance of a laboratory spectrometer is only limited by parameters governing the response function of the detector and by the source intensity, the performance of a bore-hole spectrometer is influenced by other important factors. There are, for instance, spectral interferences caused by radiation from the detector and its housing, and spectral interferences from different nuclear reactions. These two problems can be substantially overcome by procedures shortly to be described. In addition, owing to scattering and other effects, the true spectrum

of gamma radiation liable to be induced in the rock matrix is not transmitted to the detector. Indeed, the actual size of the gamma "kernel" from which useful spectroscopic information can be obtained depends on a variety of factors including the distribution of neutron flux densities of various energies as a function of time and of distance from the source, and on the attenuation coefficients for gamma radiations of various energies. Both these latter factors depend heavily on the nature of the elements present in the rock and have been the subject of many theoretical investigations comprehensively reviewed by Koshevnikov[8]. No general statement can be made but some idea of the volume of rock providing a resolvable gamma signal can be obtained from experiments with granitic type rocks of average composition and density (2.5-2.9 g/cm^3) and for ^{24}Na

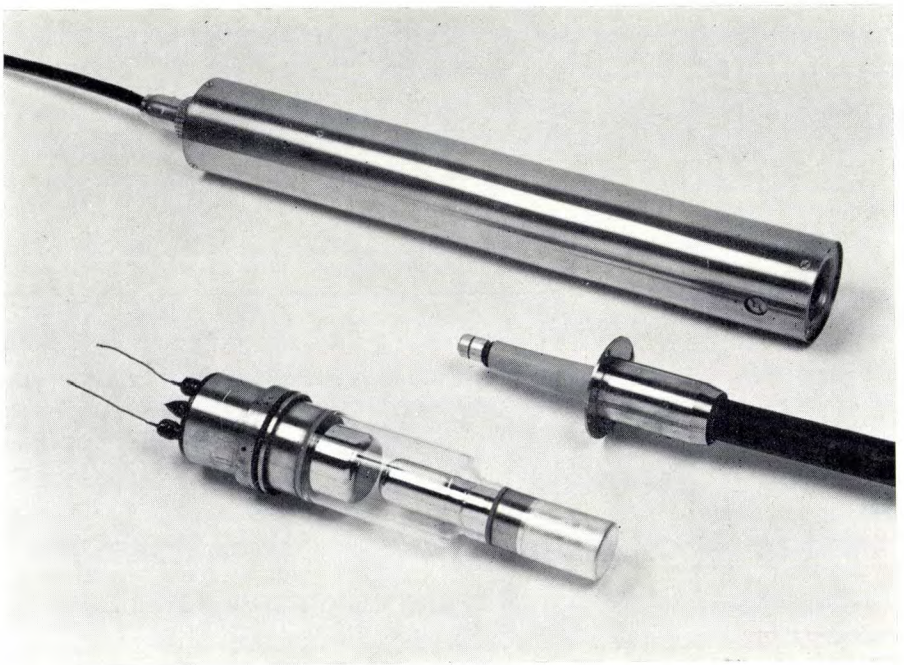


Fig. 1. Philips sealed neutron generator tube (lower left). The tube as delivered (top) is enclosed in a metal cylinder with an outer diameter of 7 cm. (Photograph from O. Reifenschweiler and K. Nienhuis [5].)

[2] H. R. Brannon and J. S. Osoba, J. Petrol. Technol. **8**, 30, 1956.
P. E. Baker, J. Petrol. Technol. **9**, 97, 1957.
N. L. Muench and J. S. Osoba, Petrol. Trans. A.I.M.E. **210**, 89, 1957.
A. A. Fedorov, M. M. Sokolov and A. P. Ochkur, Soviet J. Atomic Energy **8**, 555, 1960.

[3] For completeness it should be mentioned that gamma spectrometry is not the only means of obtaining information by nuclear methods in logging. Again using a source of fast neutrons, there are a number of possibilities for detecting the presence of neutron-capturing elements by logging the concentration of slow neutrons produced in the rocks. Average neutron lifetime can also be measured in favourable situations. Both methods have been widely employed for the solution of special problems in the oil industry, but neither of them gives specific information on element occurrence in the complicated environments usually encountered in the mineral field, so these methods will not be discussed here. See for instance R.L. Caldwell, W. R. Mills and W. W. Givens, Proc. I.A.E.A. Symp. on Use of nuclear techniques in the prospecting and development of mineral resources, Buenos Aires 1968, paper SM-112/25.

[4] R. L. Caldwell and R. F. Sippel, Bull. Assoc. Amer. Petrol. Geol. **42**, 159, 1958.
J. T. Dewan, R. L. Stone and O. L. Morris, J. Petrol. Technol. **13**, 531, 1961.
R. L. Caldwell, W. F. Baldwin, J. E. Bargainer, J. E. Berry, G. N. Salaita and R. W. Sloan, Geophysics **28**, 617, 1963.

[5] O. Reifenschweiler and A. C. van Dorsten, Phys. Verh. (Mosbach) **8**, 163, 1957.
O. Reifenschweiler, Philips Res. Repts. **16**, 401-418, 1961.
O. Reifenschweiler and K. Nienhuis, Philips tech. Rev. **23**, 325-337, 1961/62.

[6] A. H. Youmans and W. Zimmerman, Fifth World Petrol. Congr., 1959, section 10, paper 17, p. 207.
A. H. Youmans, E. C. Hopkinson and R. M. Stewart, Soc. Petrol. Engrs., 34th Fall Meeting, Dallas 1959, paper 1304G.
J. Tittman and W. B. Nelligan, Petrol. Trans. A.I.M.E. **219**, 375, 1960.
I. V. Popov, Portable neutron generators in nuclear geophysics, Publ. 127 Gosatomizdat, Moscow, 1962.
F. A. Alekseev, D. F. Bespalov, G. V. Jorbunov and D. N. Frebrodol'skii, Proc. I.A.E.A. Symp. on Radio-isotope instruments in industry and geophysics, Warsaw 1965, Vol. 2, p. 139.

[7] E. M. Lobanov, A. P. Novikov, G. S. Nikanorov, O. M. Romanov and A. A. Khaidarov, English translation in Min. Minerals Engng. July 1966, p. 261, and August 1966, p. 303.

[8] D. A. Koshevnikov, Proc. I.A.E.A. Symp. on Radio-isotope instruments in industry and geophysics, Warsaw 1965, Vol. 1, p. 43.

[9] K. I. Yakubson and K. Eife, Proc. I.A.E.A. Symp. on Radio-isotope instruments in industry and geophysics, Warsaw 1965, Vol. 2, p. 159.

[10] A. M. Blyumentsev, I. I. Feld'man, G. A. Nedostup and A. N. Zheltikov, Proc. I.A.E.A. Symp. on Radio-isotope instruments in industry and geophysics, Warsaw 1965, Vol. 2, p. 187.

radiation (2.75 MeV), when the kernel radius was found to measure approximately 28 cm for a 7.5 cm (3 in.) diameter bore hole. That is, useful gamma radiation can be obtained from ^{24}Na sources up to approximately 10 cm from the walls. Beyond this radial distance the great bulk of the signal radiation is scattered by the Compton process and thereby degraded in energy, so that it will contribute only to the background in the lower energy portions of the spectrum. In the same rock the useful distance would be smaller for lower energy radiation and substantially larger (about 15 cm) for 7.6 MeV iron radiation.

Neutron and photon physics of the logging process

The release of fast neutrons within an extensive rock matrix results in a number of competing processes of gamma-ray excitation:

a) Inelastic collisions, the (n, n'γ) process. A nucleus is excited to one of its higher states by a neutron which loses the corresponding amount of its energy, and the nucleus rapidly returns to its normal state by emitting the excitation energy as gamma radiation.

b) Radiative neutron capture, the (n,γ) process. A nucleus undergoes transmutation by capturing a neutron. This is accompanied by gamma emission, and more gamma photons are emitted immediately after the capture in the de-excitation process of the product nucleus.

c) Activation. An unstable compound nucleus is formed by a neutron capture, and in the decay of this radioactive nuclide different gamma photons are emitted. In this case a delayed gamma emission is observed; the intensity of this radiation varies with time according to the half-life of the nuclide or of its isomeric states.

d) Particle reactions such as (n,2n), (n,α), (n,p), which may yield gamma rays by the subsequent decay of radioactive reaction products.

The participation of wanted and unwanted constituents of bore-hole surroundings in these processes differs greatly for different materials, and it also depends in different ways on neutron energy and results in very different spectral, spatial and time distributions of the gamma radiation produced. A few details of this complex situation are given here.

a) The inelastic collision cross-section of a nuclide ("reaction probability") is a function of the excitation energy for its individual states. It is usually less than 0.5 barn for one state, and the total cross-section of the metallic elements sought frequently exceeds 1.5 barn. The greatest density of transitions of nuclei occurring in minerals is in the 0.7-2.0 MeV range. For the light and light-to-medium weight nuclei (atomic weight < 70) constituting the vast majority of rock and ore-matrix elements the transition energy of the lowest excited state, i.e. the threshold energy for inelastic collision excitation, is above 2 MeV. For oxygen, which is the most abundant rock element, it is 6.1 MeV; for carbon, which occurs in coal and limestones, it is 4.43 MeV. When the high-energy photons of these elements are produced, as will be the case with many of the commonly available neutron sources (see below), these spectral components complicate the resolution of peaks in this spectral region. They may, however, be relevant to the mineral prospector as an indirect guide to the quality of, for example, dense sulphide ore bodies or coals.

b) Radiative neutron capture occurs at a significant rate only after the neutrons in the rock or ore matrix have been slowed down to thermal (or nearly thermal) velocities. This gamma radiation therefore is produced with a certain delay — relative to the moment of neutron production — and at a larger mean distance from the bore hole, as contrasted with the "prompt spectrum" of the inelastic collisions. Most of the neutron-capture gamma rays emitted and useful for identification have energies in the range 2-10 MeV, and a fortunate circumstance is that many of the base metallic elements, for example Ni, Fe, Cu, Ti, Mn, Cr and Pb, are distinguished from those of common rock (Si, Ca, Mg) by the fact that their most important gamma radiations occur in the high-energy range of 7-9 MeV. The main source of interference, besides oxygen, is iron, because its 7.65 MeV radiation greatly exceeds in intensity neighbouring radiations of other nuclei and because it is fairly abundant in most rocks as well as in many ore bodies.

c) Gamma rays emitted in the decay of a radioactive nuclide formed by neutron capture have energy distributions which are largely below 3 MeV. The time dependence of their intensity is a very useful feature: when a nuclide has a relatively short half-life ($<$ 10 min), identification of its characteristic radiation by measurement of the half-life time may provide a powerful, and a sometimes necessary confirmatory test of a suspected element. Interferences in these measurements can arise only from products of particle reactions and from isomeric states which have very short half-lives ($<$ 1 second).

d) Radioactive nuclides resulting from particle reactions also may be useful for confirmatory identifications, although they are unsuitable for a general identification technique. For instance, the radiative neutron-capture cross-section of lead is very small; the main alternative to inelastic collisions for its identification is the (n, 2n) reaction which results in $^{203}Pb^{m}$, $^{204}Pb^{m}$, $^{205}Pb^{m}$ and $^{207}Pb^{m}$ isomers of very short half-life.

It will be clear from the above that the diversity of competing neutron induced reactions makes it desirable to carry out a separate spectral analysis for each type of reaction. This is feasible because the slowing-down of fast neutrons by inelastic collisions, the absorption of slow neutrons by capture, and the decay of unstable nuclides formed are successive, though overlapping, stages of neutron behaviour.

The slowing-down of the neutrons

The mean slowing-down time of fast neutrons to the thermal energy of 0.025 eV is a characteristic parameter of the overall chemical composition of the medium. It is an appreciable quantity in any medium owing to the long free path and to the large number of collisions required for such a degradation of the initial neutron energies. For a 3 MeV neutron, for instance, the first collision mean free path is about 6.5 cm in granite (with 1.5% water). The mean slowing-down time to 1 eV in granite is about 45 μs, and to 0.025 eV about 300 μs. There is therefore a significant delay before the level of gamma radiation due to radiative neutron capture can build up appreciably. This delay is augmented by the diffusion of the thermal neutrons: the "mean lifetime before capture" is roughly 1000 μs in most rocks, although it is substantially smaller than this in some ore bodies, haematite with 50 μs being an extreme example.

The spatial distribution of radiation processes is also greatly influenced by this behaviour. Reverting to the example of 3 MeV neutrons entering a granite rock volume, the long mean free path, whose effect is enhanced by the strong forward scattering of fast neutrons and the small logarithmic energy decrement in granite ($\approx$ 0.10), entails that the flux of fast neutrons would still be significant as compared with the flux of thermal neutrons at distances of up to 30 cm from the source.

In a bore-hole configuration there is usually a thin sheath of water (about 2 cm) between the logging tool and the surrounding rock. Approximate calculations, again for 3 MeV neutrons, show that this does not cause an appreciable immediate slowing-down effect on the neutrons emitted, although water has a slowing-down time of only about 0.6 μs to 1 eV and of about 4 μs to 0.025 eV. The reason is that the first collision mean free path in water is still quite large, namely about 3.5 cm.

The above-mentioned mean slowing-down times of 45 μs and 300 μs respectively will of course be somewhat reduced by the presence of an inner sheath of water in the logging configuration.

The situation described provides the basis of two techniques leading to improved spectral resolvability. The more powerful of these involves the use of a *pulsed neutron source* linked with a *gated pulse-height analyser*. The onset and duration of periods of gamma-ray energy analysis are preadjusted with a suitable delay to those of the neutron bursts to accumulate the spectra of gamma rays emitted by the preselected process of neutron excitation. The other, more limited technique applies to the use of a continuous neutron source and depends on the difference between the spatial distribution of the components of the fast-neutron and the slow-neutron flux. This difference can be exploited to vary the intensity of the spectral component from inelastic collisions relative to that from slow-neutron capture by appropriately adjusting the source-to-detector spacing.

This brings us to a discussion of neutron sources, detectors and other parts of the equipment used.

Equipment for logging by gamma spectrometry

Neutron sources

There are many types of neutron sources, but only those are considered which would fit into a small bore-hole logging tool (7-10 cm diameter), and which are characterized by long life, high yield and simple operation. (These requirements in fact largely restrict our choice to fast-neutron sources.)

Radioactive nuclide sources emit neutrons as the reaction products of either (α,n) or (γ,n) processes. One such source is beryllium activated by the alpha radiation of plutonium (half-life 24 000 years). With a mean neutron energy of 4.1 MeV, the energy spectrum of this Pu-Be source is mainly in the high-energy region (see *fig. 2*), so that even the inelastic collision spectra of carbon and

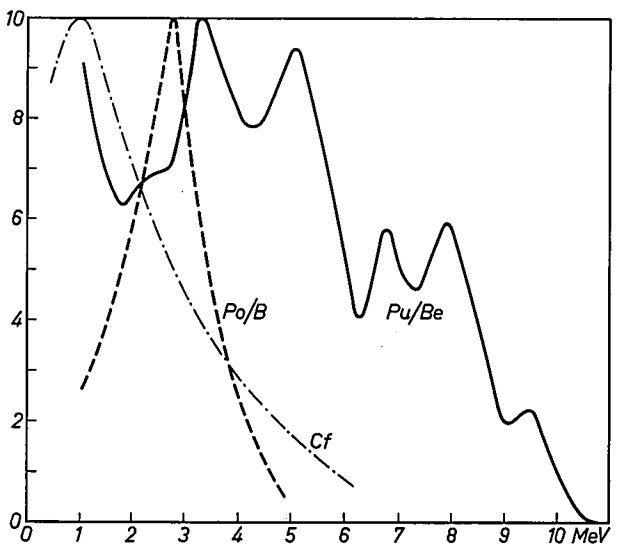


Fig. 2. Energy spectrum of neutron sources using radioactive nuclides. The solid curve applies to a plutonium-beryllium source, the dashed curve to a polonium-boron source, and the chain-dotted curve to the much stronger californium fission source. The relative neutron flux is plotted against the energy of the emitted neutrons.

oxygen can be weakly excited. Outputs of 10^7 neutrons/second are readily available. A continuously operating source with a lower neutron-energy distribution, such as boron activated by the alpha radiation of polonium (half-life 138 days), may enable easier resolution of prompt photon spectra. The mean neutron energy from the latter source is about 2.8 MeV (fig. 2) so that the nuclei of many of the common rock and metallic ore elements will also be excited by inelastic collisions with these neutrons. Since most of these resulting gamma photons are also below 2 MeV, there will be negligible interference in the high-energy region where many of the most important components of the photon spectrum of neutron capture are situated.

This type of source is also useful in studying the lower-energy range activation spectra, as interferences due to particle reactions caused by high-energy ($>$ 4 MeV) neutrons are largely absent.

Very recently a new and powerful type of neutron source has been announced [11] although these sources are not yet freely available. The neutrons are produced by spontaneous fission of the californium nuclide ^{252}Cf (half-life 2.65 years) and outputs of 10^{10} neutrons/second are reported from a few milligrammes of californium. The energy spectrum of this source (also shown in fig. 2) is mainly in the low-energy region.

Neutron tubes produce neutrons by the reaction of accelerated charged particles with suitable nuclide targets. Their advantages over available radioactive nuclide sources of the (α,n) type are a greater yield and the possibility of pulsed operation. The tubes of the greatest practical importance use a tritium target bombarded by deuterium ions. The reaction taking place is ^3H (d,n) ^4He + 17.6 MeV, which produces 14 MeV neutrons. Its effective cross-section is by far the highest of the charged particle reactions occurring at the relatively low accelerating voltages feasible with a down-hole tube (about 100 keV). Among the current range of commercially available sealed tubes producing 14 MeV neutrons, only two types of tube (both marketed by Philips) are suitable for logging applications and these have a continuous output restricted to a maximum of approximately 10^8 neutrons/s. Both types have an expected life of greater than 1000 hours and they can be pulsed to provide neutron burst lengths of between 3 microseconds and 3 milliseconds or longer at a much greater output rate than in continuous mode. The attainable output of the first type of tube, pictured in fig. 1 [5], during the pulse is 10^9 neutrons/s, at a duty cycle of 10 per cent. With the second type of tube [12], embodying an improved ion source, pulse-yield rates of 3×10^{11} neutrons/s are attainable, with the same continuous output and therefore at a proportionately smaller duty cycle.

Sealed tubes which utilize the ^2H(d,n)^3He reaction and yield 2.8 MeV neutrons are also available. Despite the advantage of this lower neutron energy the yield rates from currently available tubes of this type are at present too low for logging purposes at comparable accelerating voltages, certainly for pulsed output. Their yield rate is about 100 times smaller than that of the deuterium-tritium tubes.

Gamma radiation detectors

Our experience so far has been restricted to the well-known NaI(Tl) scintillation detector used with a photomultiplier [13] which has proved sufficiently rugged for oil-well logging applications.

The solid-state Ge(Li) detector has distinct advantages over the scintillation detectors in applications of high-energy gamma ray spectrometry [14], notably because of its energy resolution, which is at least 10-20 times better, and which largely offsets its lower detection efficiency due to the smaller feasible sensitive volume. The fragility of the Ge(Li) detectors, however, and the requirement of cooling to liquid-nitrogen temperature, present technological obstacles to their immediate development as logging tool detectors.

The NaI(Tl) scintillator has an adequate intrinsic efficiency at high photon energies, i.e. a large percentage of the incident photons are absorbed by liberation of K-electrons. Each of these absorption processes in a scintillator will by further ionization processes, either directly or via a fluorescent photon, produce a number of ion pairs, i.e. a pulse amplitude, which should be proportional to the incident photon energy (photo peak). In a certain portion of these avalanche processes, however, the fluorescent photon will escape from the crystal, so that a pulse amplitude smaller by a certain amount is obtained, and the spectral peak due to these pulses (the escape peak) clearly will be more intense when more photons can escape, i.e. when the crystal is of a smaller size [15]. The bore-hole configuration unfortunately restricts the maximum crystal diameter to about 5 cm, and in this case the photo peak obtained in the NaI(Tl) crystal is relatively weak. For the same crystal geometry, CsI(Na) has a larger photo-peak efficiency than NaI(Tl) (30 % greater at 2.5 MeV), and the relative difference increases with photon energy, but this advantage is offset by a slightly inferior energy resolution.

There are other problems associated with the use of NaI and CsI crystals. One of them arises from the high temperature sensitivity of photomultiplier tubes and from the considerable environmental temperature change expected during some logging operations [16]. This is surmountable, but large temperature changes may require a wider range of gain stabilization than is

offered by commercial equipment currently available. Another problem is the required shielding between detector and neutron source. This is an important problem which will now be discussed at some length.

Detector shielding

With a fast neutron source, two types of detector shielding are necessary. One of these is a cylindrical metallic shield to shut off the direct beam of source radiation. This takes the form of a multicomponent unit because a single element providing the required effective absorption of both neutrons and gamma rays emanating from the source does not exist. Shields fabricated from materials such as tungsten-rich alloy or copper can be located adjacent to the source for their effective fast-neutron attenuation and for their slow-neutron capture characteristics. Lead is usually preferred as the gamma-ray shield adjacent to the detector because of its low capture cross-section and its excellent photon-attenuation properties.

A shield is also required around the detector to protect it from the neutrons scattered from the bore-hole surroundings. The best materials are ^{10}B and ^{6}Li with their large cross-sections for neutron removal by the (n,α) reaction: $\sigma = 3837$ barns for ^{10}B and $\sigma = 950$ barns for ^{6}Li. However, the 477 keV photon emission from the excited state of ^{7}Li may be a disadvantage of the $^{10}B(n,\alpha)^{7}Li$ shielding reaction in some applications of low- and intermediate-energy gamma-ray spectrometry; also the high-energy photon measurements may be affected, since after each pulse, irrespective of its amplitude (i.e. photon energy), the detector and analyser have a certain dead time, so that swamping the detector with unwanted low-energy pulses will distort the response of the system to high-energy pulses as well.

Analysis of bore-hole logging output data

A logging operation might be conducted in two stages. The first stage would be a rapid continuous "scanning" descent of the tool (up to 250 m per hour). During this descent, a large number of radiative capture spectra would be consecutively recorded and analysed, each spectrum representing a small length of the bore hole (15-30 m). A single day's operation would thus provide preliminary information about the existence of mineral bearing zones even in the deepest holes. This analysis could then be followed by more detailed capture, inelastic scattering, or activation spectrometry in the relatively short regions of interest indicated by the first survey.

For a meaningful analysis of the data obtained during the first run the statistical precision should be adequate to reveal variations in spectral shape that correspond to significant chemical differences in the matrix. The "channels ratio" provides a simple and sensitive test which is suitable for the expected rate of accumulation of output data. The basis of the test is that the ratio of the counts recorded in any two given groups of consecutive pulse-height analyser channels is constant only in cases of similarly shaped spectra. The unavoidable loss of input signals during the digital readout mode of the analyser can be minimized by using high-speed output devices to furnish the operator with the information required for calculation of channels ratios.

Our experience with detailed gamma-ray spectra indicates that a data convolution technique somewhat similar to that of Yule [17] is necessary for a qualitative analysis. For a quantitative analysis it is probably essential to combine prior calibrations with programmed curve-fitting procedures utilizing non-linear least-squares techniques for both the gamma peaks (which have a Gauss distribution) and the background. A sophisticated approach to this problem is demanded because the spectrum given by a scintillation detector (particularly a capture or inelastic-collision spectrum) consists of lines which are often partially coalesced; because these lines are frequently of not much greater intensity than the background; and because the background cannot be easily represented by a simple analytical function. Other information which seems essential to a solution at this stage consists of a) the likely chemical composition of the rocks to be encountered and of the ores sought — which is usually available from the geological evidence — and b) the relative intensities of photo peaks and escape peaks as a function of energy. The latter information is available in the literature [18] from probability calculations. (These are usually based on the Monte Carlo method, which although not applicable to cylindrical geometry gives results sufficiently accurate for our purpose.)

[11] W. C. Reinig, Nucl. Appl. **5**, 24, 1968 (No. 1).
[12] C. W. Elenga and O. Reifenschweiler, Proc. I.A.E.A. Symp. on Pulsed neutron research, Karlsruhe 1965, Vol. 2, p. 609-622.
[13] On these detectors, see for example: The scintillation spectroscopy of gamma radiation, editor S. M. Shafroth, publ. Gordon & Breach, New York 1967, Vol. I, chapters 1 and 7.
[14] See for example W. K. Hofker, Semiconductor detectors for ionizing radiation, Philips tech. Rev. **27**, 323-336, 1966.
[15] An account of detection processes in a scintillator may be found in Philips tech. Rev. **18**, 1956/57, on p. 269, and **20**, 209-219, 1958/59.
[16] The geothermal gradient in Australia, though not large in comparison with many other locations, amounts to approximately 6 °C per 300 m, for example at Broken Hill and at Cobar, N.S.W., and at Mt. Isa, Queensland. At a few localities it is less than this, but at Rosebery, Tasmania, 8.7 °C per 300 m is reported. See: Handbook of physical constants, memoir 97, Geological Society of America, editor S. P. Clark, Jr., New York 1966.
[17] H. P. Yule, Anal. Chem. **38**, 103, 1966.
H. P. Yule, Nucl. Instr. Meth. **54**, 61, 1967.
[18] W. F. Miller and W. J. Snow, U.S.A. E.C. Report ANL-6318, 1961.

Experimental investigations

Various rigs have been constructed in the Divisional Laboratories of the Commonwealth Scientific and Industrial Research Organisation at Port Melbourne to simulate bore holes. Two general types are required. For simulating low concentrations of metal ores disseminated with some degree of uniformity in plutonic or metamorphic rocks, it is convenient to mix appropriate concentrations of metal or metal concentrate with a cement-sand or cement-sand-aggregate mixture which can be cast in a suitable mould around a central pipe. Although large castings can be made, these are not only cumbersome but somewhat inflexible in use from the experimental point of view, and structures built to any size and shape from large bricks are more convenient to assemble (*figs. 3* and *4*). The composition of the castings or bricks can be varied over wide limits, although the final water content is somewhat higher than in normal rocks. This might cause large differences, since hydrogen is the most effective element for slowing down neutrons to thermal velocities particularly at low initial energies. In a real situation, however, the presence of water in open pores and fissures compensates for much of the deficiency of the laboratory models.

The size of the rigs is adjusted for each type of neutron source and rock matrix, so that there are no significant effects of size on the gamma spectrum and negligible health hazards. Thus the inner portion of the solid matrix permits the slowing down and diffusion of neutrons according to the neutron properties of the rock, while the outer portion serves as a radiation shield.

Water in a bore hole can be simulated satisfactorily by using a polythene or polypropylene tube for the pipe defining the hole, and using similar solid materials around the source.

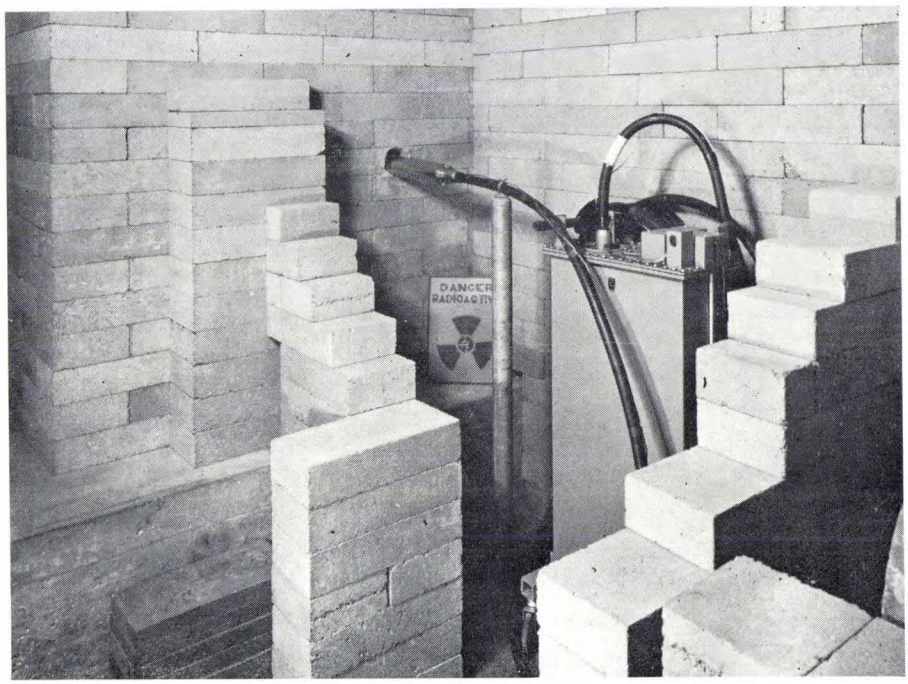


Fig. 3. Part of an experimental rig for simulating a bore hole. The rock surrounding the hole is simulated by bricks of a suitable composition. The neutron tube is seen entering the hole. The high-voltage generator and cable are also visible.

For simulating the higher concentrations of metals in a mineral lode, dilution is undesirable, and maximum element concentration is obtained by filling finely crushed ore concentrate into a polythene drum fitted with a central pipe (see fig. 4). In a few cases it may be desirable to overcome the porosity introduced by this arrangement by working with actual specimens of the lode material, but the difficulty of obtaining these in adequate size precludes their general use. Alternatively, corrections can be applied to experimental data for crushed material to convert the readings to those appropriate for the solid rock.

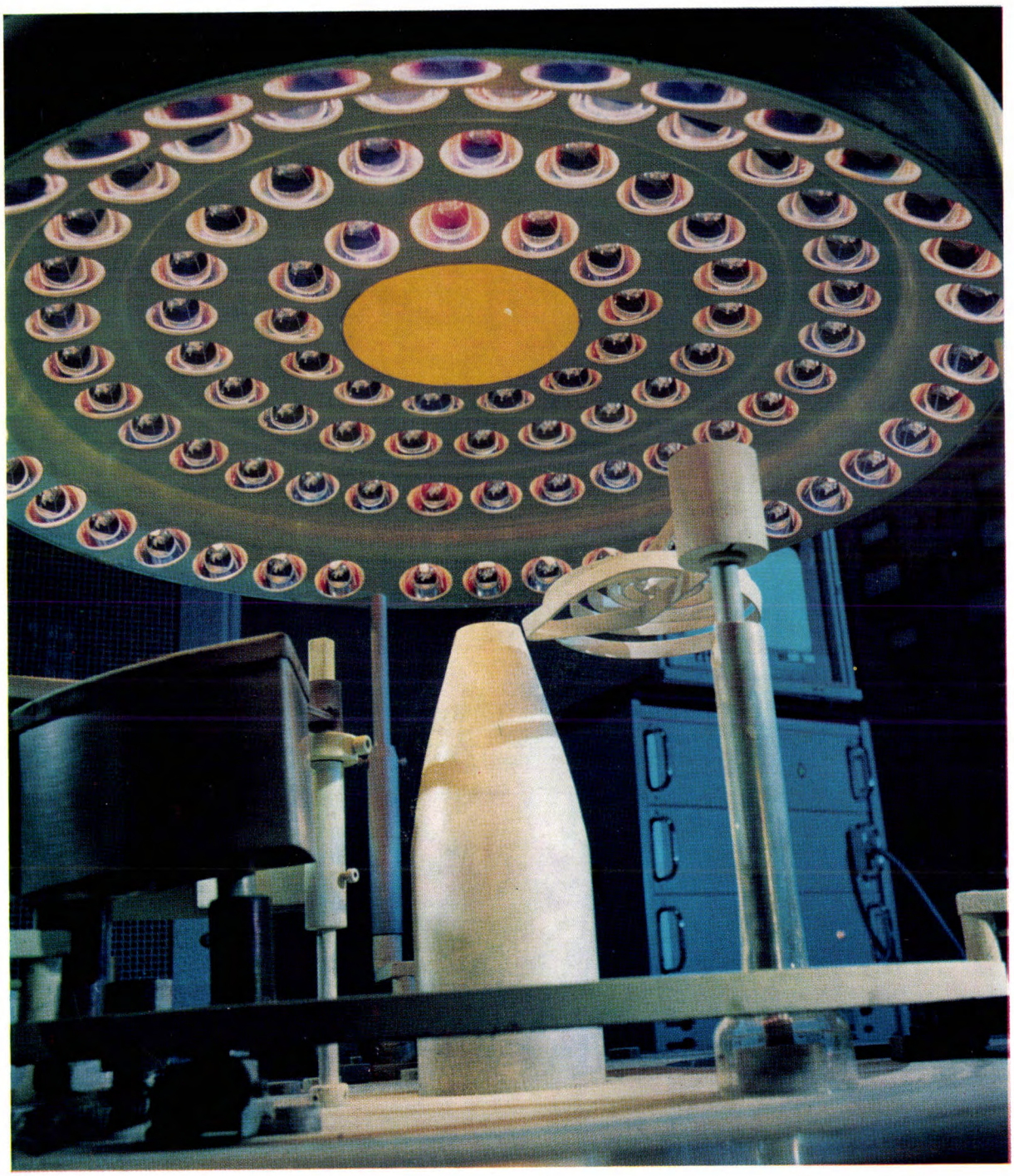
Some preliminary work has been done on construct-

ing a general model of a bore hole to avoid the need for simulating individual situations (fig. 4). Such a model, with a heterogeneous distribution of rodlike material in regions of maximum interest, would offer prospects for more rapid evaluation of the conditions yielding a useful gamma spectrum.

Fig. 4. Experimental rigs in the Port Melbourne laboratory for simulating bore holes. On the left is a horizontal rig 6 metres in length with removable sections to simulate mineralized zones 50 cm wide and spaced 1 m apart. On the right is a rig which enables the introduction of a radioactive nuclide neutron source into a pile of suitable brickwork or into a barrel containing finely ground material.

Summary. Bore-hole logging by nuclear techniques is a very promising alternative for the time-honoured but expensive method of core drilling in the search for minerals. The procedure discussed consists in lowering down the drill hole a source of fast neutrons which excite gamma radiation from the material of the wall of the hole. This radiation is detected by a scintillator crystal with photomultiplier mounted on top of the neutron source and suitably shielded from the direct beam of source radiation. From the pulses transmitted to the surface the energy spectrum of the radiation is derived by a multichannel pulse-height analyser and recorded during the lowering-down operation. After a brief discussion of the problems of spectrometry in a confined cylindrical space under adverse environmental conditions, the article reviews the contribution of the various processes of neutron scattering, radiative neutron capture, and activation, to the total gamma spectrum. The available neutron sources and detectors are also briefly reviewed. The slowing down of neutrons in rocks facilitates resolution of the spectrum since the various spectral components resulting from these processes can be separated in time provided a pulsed neutron source is used. Two types of sealed neutron generator tubes, both marketed by Philips, are suitable for this purpose. The problem of deriving the relevant information from the radiations detected is considered in some detail and experimental means for investigating this problem in the laboratory by simulating possible bore-hole situations are described.

Vacuum evaporation of reflectors for projection lamps



Halogen lamps used in projectors for the smaller sizes of cine film are mounted in an elliptical reflector which has a special reflecting layer that will transmit heat radiation, reflecting only light radiation. This special layer is made by alternately evaporating two substances of different refractive index on to the surface until 20 layers have been formed. The layers must be very uniform and of accurately controlled thickness.

The photograph shows a vacuum evaporation system, set up at the bottom of a bell-jar, which is used for making large numbers of these reflectors at the same time. The glass backings for the reflectors are mounted over holes in a slowly rotating curved metal shield. Two heated trays containing the materials to be evaporated are placed underneath the rotating shield. Each of the trays is alternately screened by a kind of box (on the left in the photograph). Before the evaporation begins a glow discharge is maintained for a short time between the rotating plate and the spiral electrode at the centre of the photograph; the discharge cleans the glass surfaces by particle bombardment.

At the centre of the shield there is a glass plate, which is coated in exactly the same way as the reflectors. This glass plate is illuminated by a beam of light through the tapered tube at the centre, and during the process the light transmitted by the plate is monitored continuously by a photoelectric cell. The light transmitted gives a measure of the thickness of the layer, enabling the operator to regulate the evaporation time for each layer. The total evaporation time is about 20 minutes.

Stoichiometry

W. Albers and C. Haas

II. Point defects and the control of their concentrations

In a solid compound, say AB, the randomly dissolved native atoms and molecules are known as native point defects, and are denoted by an asterisk, e.g. *A and *B. Foreign point defects such as atoms of C dissolved in AB are not considered in this article. The study of the chemical relations between point defects is referred to as defect chemistry [4]. The name "point defect" arose from comparison with the "ideal" crystal, which is perfectly ordered and in which all the atoms are located at specific lattice sites. In the first part of this article[*], which we shall refer to here as I, it was shown that this ideal crystal is only stable at the absolute zero point, and can therefore never actually be obtained. At $T > 0$ °K every crystal contains imperfections, even if it has been prepared under ideal conditions and it is in perfect equilibrium.

All zero-dimensional deviations from the ideal crystal are known as point defects. One-dimensional defects (line defects such as dislocations), two-dimensional defects (lattice-plane defects, such as stacking faults) and three-dimensional defects (volume defects, such as polytypes) do not occur in equilibrium. This can be seen as follows.

Let us assume that one point defect corresponds to an enthalpy of ε (free bonds etc.) and that in a line defect a whole row of atoms carries such free bonds, each of enthalpy ε. Let us also make the corresponding assumptions for a plane defect and a volume defect.

a) *Point defects.* Let n be the number of point defects in a crystal consisting of N_0 atoms. The enthalpy contributed to the system by the point defects is then $H_{\mathrm{pd}} = n\varepsilon$. The number of possible positions for one point defect is of the order of N_0, so that the number of possible arrangements is of the order of $W = N_0^n/n!$. The Gibbs free energy contributed by the point defects, $G_{\mathrm{pd}} = H_{\mathrm{pd}} - TS_{\mathrm{pd}}$, is therefore:

$$H_{\mathrm{pd}} - kT \ln W = n\varepsilon - nkT \ln (N_0/n) - nkT.$$

Using the equilibrium condition $\partial G_{\mathrm{pd}}/\partial n = 0$ we then obtain the following relation for the equilibrium concentration of point defects:

$$n/N_0 = \exp(-\varepsilon/kT).$$

In other words, the concentration is independent of N_0. At $T > 0$ °K the point-defect concentration is therefore not zero, even in an infinitely large crystal.

b) *Line, plane and volume defects.* Let m be the number of line defects in a crystal cube containing N_0 atoms. One line contains $N_0^{1/3}$ atoms, so that the enthalpy contributed by the lines is $H_{1\mathrm{d}} = mN_0^{1/3}\varepsilon$. One line can be positioned in $N_0^{2/3}$ ways, i.e. $W = N_0^{2m/3}/m!$, and hence:

$$G_{1\mathrm{d}} = mN_0^{1/3}\varepsilon - mkT \ln (N_0^{2/3}/m) - mkT.$$

From this we find the number of dislocation lines in equilibrium:

$$m = N_0^{2/3} \exp(-\varepsilon N_0^{1/3}/kT),$$

so that m tends to zero as N_0 approaches infinity. Similarly we can calculate that the numbers of plane defects p and volume defects q in equilibrium are given by:

$$p = N_0^{1/3} \exp(-\varepsilon N_0^{2/3}/kT),$$
$$q = \exp(-\varepsilon N_0/kT),$$

so that in these cases the equilibrium concentration also becomes zero for $N_0 \to \infty$. In equilibrium considerations we therefore need only take account of point defects.

Incorporation types

Dissolved native atoms (*A and *B) which have been dissolved in different ways represent different states and are thus different point defects. *Fig. 8* shows three ways in which native atoms can be incorporated in AB. Fig. 8*a* shows *B interstitially dissolved in AB, i.e. a B atom at a lattice site which is unoccupied in the ideal crystal (B_i). Fig. 8*b* shows a point defect that

Dr. W. Albers is with Philips Research Laboratories, Eindhoven; Prof. Dr. C. Haas, formerly with Philips Research Laboratories, is now Professor of Inorganic Chemistry at the University of Groningen.

[*] W. Albers and C. Haas, Stoichiometry, I. Existence region, Philips tech. Rev. **30**, 82-88, 1969 (No. 3).

[4] C. Wagner and W. Schottky, Z. phys. Chemie **B 11**, 163, 1930/31.
F. A. Kröger, The chemistry of imperfect crystals, North-Holland Publ. Co., Amsterdam 1964.
F. A. Kröger and H. J. Vink, Solid State Phys. **3**, 307, 1956.
H. J. Vink, in: Proc. Int. School of Physics "Enrico Fermi", Course XXII, Semiconductors, editor R. A. Smith, Academic Press, New York 1963, p. 68.
W. Albers, in: Physics and chemistry of II-VI compounds, editors M. Aven and J. S. Prener, North-Holland Publ. Co., Amsterdam 1967, chapter 4.

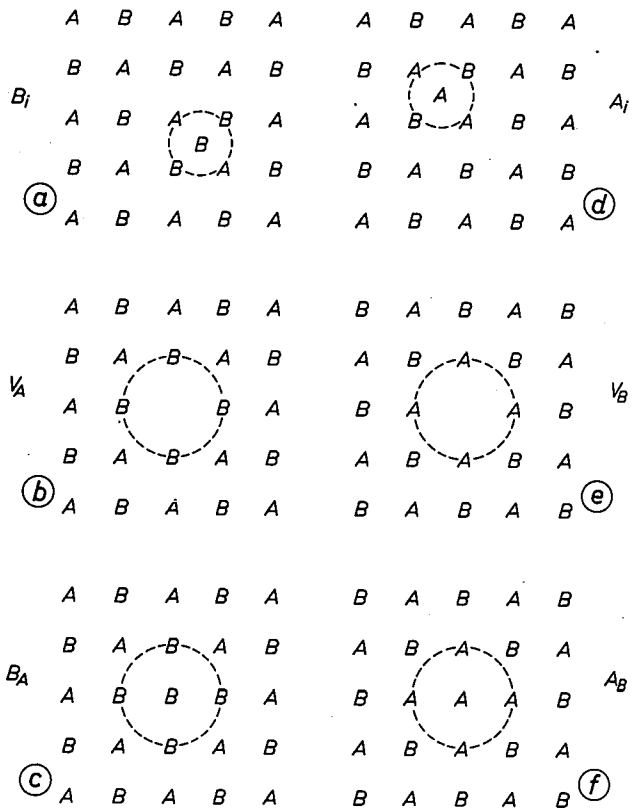


Fig. 8. *a*)-*f*) Schematic diagrams showing six simple point defects (referred to as B_I - V_A - B_A - A_I - V_B - A_B).

also corresponds to an extra B atom, owing to the absence of an A atom (V_A, where V stands for vacancy). The diagram of fig. 8*c* represents an incorporated $*B_2$ molecule. This point defect may be considered to have been formed by substituting a B atom for an A atom in the ideal crystal, i.e. a B atom at an A lattice site (B_A). Obviously, the diagram of fig. 8 is a very simple representation. In actual fact the atoms of the crystal which in the figure are located in the dotted circles will be displaced from their "ideal" sites, so that the point defect extends over a few lattice spacings. It remains, however, a single point defect (zero-dimensional). Nor need the incorporation take place symmetrically. For example, in fig. 8*c* the B atom inside the circle can be so far off centre that it could be described as a (V_AB_i) associate, representing the same dissolved $*B_2$ molecule. It cannot therefore in any way be claimed that the symbols we have mentioned represent the exact microstructure of the point defect.

The same applies of course for *A, which can be incorporated in AB as A_i or V_B (fig. 8*d* and *e*), and $*A_2$ as A_B (fig. 8*f*).

Equilibrium relations between point defects in one phase

Point defects can have a chemical potential assigned to them (see I). In principle all incorporation types occur simultaneously, and a compound AB therefore contains many kinds of native point defects at $T > 0$ °K. In a similar way to that in I we can derive the following relations in equilibrium:

$$\frac{\mu(*A_1)}{\nu_1} = \frac{\mu(*A_2)}{\nu_2} = \ldots = \frac{\mu(*A_n)}{\nu_n}, \quad . \quad (28)$$

and similarly:

$$\frac{\mu(*B_1)}{\nu_1'} = \frac{\mu(*B_2)}{\nu_2'} = \ldots = \frac{\mu(*B_m)}{\nu_m'}, \quad . \quad (29)$$

$$\mu(AB) = \frac{\mu(*A_k)}{\nu_k} + \frac{\mu(*B_l)}{\nu_l'}, \quad . \quad . \quad . \quad . \quad (30)$$

where $*A_k$ is a native point defect consisting of ν_k atoms A, incorporated as type k. For example, if $*A_k = A_i$ or V_B, we have $\nu_k = 1$, and for $*A_k = A_B$ we have $\nu_k = 2$. The same holds for $*B_l$ and ν_l'.

For small concentrations we can write:

$$\mu(*A_k) = \mu^0(*A_k) + RT \ln [*A_k], \quad . \quad (31)$$

$$\mu(*B_l) = \mu^0(*B_l) + RT \ln [*B_l], \quad . \quad (32)$$

where the square brackets denote mole fractions. The other symbols are used in the same way as in I. Substituting from equations (31) and (32) into (30) gives:

$$[*A_k]^{1/\nu_k}[*B_l]^{1/\nu_l'} = \exp \frac{\mu(AB) - \frac{\mu^0(*A_k)}{\nu_k} - \frac{\mu^0(*B_l)}{\nu_l'}}{RT} = K_{kl}, \quad . \quad (33)$$

in which the disorder equilibrium constant K_{kl} at a given temperature (and pressure) is constant to a good approximation, because $\mu(AB)$ remains very nearly unchanged for small variations of $[*A_k]$ and $[*B_l]$. We can write K as $K^0 \exp(-\Delta H/RT)$, where ΔH is the reaction enthalpy and K^0 is a factor that contains the non-configurational part of the reaction entropy. Equation (33) can be considered to have been obtained by applying the law of mass action to the chemical reaction

$$AB_{crystal} \rightleftarrows \frac{*A_k}{\nu_k} + \frac{*B_l}{\nu_l'}. \quad . \quad . \quad . \quad (34)$$

Relations between the equilibrium concentrations of native point defects can therefore be derived directly by applying the law of mass action if a disordered crystal is described in terms of a chemical equilibrium reaction.

From the derivation of equation (33) it is evident that the reaction equation (34) in no way expresses that molecules of AB are to be found in the crystal AB. AB is the stoichiometric formula of the ideal crystal, i.e. the stoichiometric composition. As we noted ear-

lier, the native point defects $*A_k$ and $*B_l$ are usually also very complex in structure. Relations like equation (33) do not depend on the structure of the condensed phase. They are valid both for liquids and for crystals, covalent as well as ionic. The *numerical value* of the equilibrium constant K does of course depend on structural factors.

At the exact stoichiometric composition we can write:

$$\sum_k [*A_k]\, \nu_k = \sum_l [*B_l]\, \nu_l';$$

on the AB(A) side the left-hand term is the larger, and on the AB(B) side the right-hand term is the larger.

With the incorporation types described in fig. 8 we can now use equation (33) to write down the following nine native disorder relations of the binary compound AB:

$$[A_i]\,[B_i] = K_1, \quad (35)$$
$$[A_B]^{1/2}\,[B_i] = K_2, \quad (36)$$
$$[V_B]\,[B_i] = K_3, \quad (37)$$
$$[A_i]\,[B_A]^{1/2} = K_4, \quad (38)$$
$$[A_B]\,[B_A] = K_5, \quad (39)$$
$$[V_B]\,[B_A]^{1/2} = K_6, \quad (40)$$
$$[A_i]\,[V_A] = K_7, \quad (41)$$
$$[A_B]^{1/2}\,[V_A] = K_8, \quad (42)$$
$$[V_B]\,[V_A] = K_9. \quad (43)$$

(The constants K_1 - K_9 are not independent of one another, e.g. $K_2K_4 = K_1K_5^{1/2}$.) The native disorder relations (37) and (41) describe Frenkel disorder, equation (39) describes antistructure disorder, and equation (43) describes Wagner-Schottky disorder. As a rule one type of disorder predominates. The Frenkel type is likely to be found if one of the constituent atoms is large and the other small, as in silver halides, for example. The alkali halides and PbS mainly have Wagner-Schottky disorder, while antistructure disorder is to be expected in compounds with atoms (ions) possessing roughly the same electronegativity and atomic or ionic radius, such as CdSb, ZnSb, Mg_2Sn and Bi_2Te_3.

Effectively charged point defects

Up to now we have discussed thermodynamic relations between native point defects that are effectively neutral. A point defect in a substance is effectively neutral if the substance does not become charged when only this point defect is added. Apart from effectively neutral point defects a substance also contains effectively charged point defects and free electrons e and holes h, in accordance with equilibrium reactions such as:

$$*A \rightleftarrows *A^+ + e, \quad (44)$$
$$*B \rightleftarrows *B^{2-} + 2h. \quad (45)$$

In equations (44) and (45) *A is a donor, such as an electropositive metal atom, and *B is an acceptor, such as an electronegative metalloid.

Charged point defects such as $*A^+$ can also be assigned a chemical potential $\mu(*A^+)$. The same applies to free charge carriers, so that for the equilibria (44) and (45) we may write:

$$\mu(*A) = \mu(*A^+) + \mu(e), \quad (46)$$
$$\mu(*B) = \mu(*B^{2-}) + 2\mu(h). \quad (47)$$

The chemical potential of electrons is also referred to as the Fermi energy.

For small concentrations the concentration dependence of the chemical potential is given by:

$$\mu(*A^+) = \mu^0(*A^+) + RT \ln [*A^+], \quad (48)$$
$$\mu(*B^{2-}) = \mu^0(*B^{2-}) + RT \ln [*B^{2-}], \quad (49)$$
$$\mu(e) = \mu^0(e) + RT \ln n, \quad (50)$$
$$\mu(h) = \mu^0(h) + RT \ln p, \quad (51)$$

where n and p represent the concentrations of electrons and holes, respectively. These must be small in comparison with the effective density of states N_c in the condition band and N_v in the valence band; the following expressions (for parabolic bands) apply:

$$N_c = 3.5\times10^{19} \left(\frac{T}{300}\right)^{3/2} \left\{\frac{m^*(e)}{m}\right\}^{3/2} \text{cm}^{-3}, \quad (52)$$

$$N_v = 3.5\times10^{19} \left(\frac{T}{300}\right)^{3/2} \left\{\frac{m^*(h)}{m}\right\}^{3/2} \text{cm}^{-3}. \quad (53)$$

Here T is the absolute temperature, m^* the effective mass and m the rest mass of the free electron. The equations (52) and (53) therefore indicate that (50) and (51) are valid at room temperature for charge-carrier concentrations of the order of 10^{19}-10^{20} cm^{-3} if $m^*/m \geqq 1$ (i.e. about 0.1 % of the total number of atoms). If the temperature is low or $m^*/m < 1$, the range of application for these equations is very small (smaller than that of equations 31, 32, 48 and 49). Exact but more complicated expressions also exist, however, for larger charge-carrier concentrations (Fermi-Dirac statistics).

If effectively charged point defects are present in arbitrary numbers, the phase will have a net charge. The electrostatic interaction between the charges gives an extra contribution to the energy of the order of

$$E \approx \varrho^2 V^2/R,$$

where ϱ is the charge density, and $V = 4\pi R^3/3$ is the volume of the phase (taken as a sphere of radius R).

If conventional thermodynamical considerations are to be valid, then the extensive thermodynamic quantities, such as energy, must be proportional to the quantity of matter. This means that the charge density

ϱ for an infinitely large phase ($V \to \infty$) must be zero; in other words the phase must be electrically neutral.

The neutrality condition in the case where only effectively single-valued charged atomic native point defects occur in AB is:

$$[*A^+] + p = [*B^-] + n \ , \quad . \ . \ . \quad (54)$$

if A and B form respectively positive and negative point defects.

Given equations (31), (32) and (48)-(51), we can apply the law of mass action to reaction equations such as $*A \rightleftarrows *A^+ + e$ and $*B \rightleftarrows *B^- + h$, which leads to concentration relations such as:

$$\frac{[*A^+]n}{[*A]} = K_{10} \ , \quad . \ . \ . \ . \ . \ . \ . \quad (55)$$

$$\frac{[*B^-]p}{[*B]} = K_{11} \ . \quad . \ . \ . \ . \ . \ . \ . \quad (56)$$

From the equilibrium

$$AB_{crystal} \rightleftarrows *A^+ + *B^- \quad . \ . \ . \quad (57)$$

it can be shown that

$$[*A^+]\,[*B^-] = K_{12} \quad . \ . \ . \ . \quad (58)$$

(compare equation (58) with the equilibrium equation for water $[H^+]\,[OH^-] = K(H_2O)$). Combination of equations (33), (55), (56) and (58) gives the well-known equilibrium relation from semiconductor physics [5]:

$$np = K_i \ . \quad . \ . \ . \ . \ . \ . \quad (59)$$

The equilibrium constants K can again be written as $K = K^0 \exp(-\Delta H/RT)$, where ΔH is the reaction enthalpy and K^0 is a factor.

In some cases the reaction enthalpy can be calculated. For example, the ΔH of the reaction $NaCl_{crystal} \rightleftarrows *Na^+ + *Cl^-$, in which the point defects are incorporated as V_{Cl}^+ and V_{Na}^-, has been found to have a value of 2.0 ± 0.1 eV [6]. It has been found experimentally from the following cycle that $\Delta H = 2.12$ eV, which agrees well with the calculated value:

$NaCl_{crystal}$	$\rightleftarrows V_{Na}^- + Na_{gas}^+$	-4.80 eV
$NaCl_{crystal}$	$\rightleftarrows V_{Cl}^+ + Cl_{gas}^-$	-5.14 eV
$Na_{gas}^+ + Cl_{gas}^-$	$\rightleftarrows NaCl_{crystal}$	$+7.82$ eV
$NaCl_{crystal}$	$\rightleftarrows V_{Cl}^+ + V_{Na}^-$	-2.12 eV

The factor K^0 cannot be calculated, however, so that in practice the equilibrium constants have to be determined experimentally.

From equations (54) and (58) it follows that the semiconducting substance AB is an N-type conductor if $[*A^+] > [*B^-]$ and a P-type conductor if $[*B^-] > [*A^+]$. This is not to say that the semiconductor AB with surplus A is always an N-type conductor and the one with a surplus of B is always P-type. This depends on the incorporation type. If, for example, surplus B occurs mainly in the form of dissolved $*B_2$ molecules, then these may be donors, in accordance with $*B_2 \rightleftarrows *B_2^{y+} + ye$ (y is the number of charge units), so that AB with surplus B is then an N-type conductor. This will be the case for example if $*B_2$ is incorporated as B_A, and B has more valence electrons than A, which is entirely comparable with the fact that arsenic incorporated in germanium at a Ge lattice site (As_{Ge}), and indium incorporated in cadmium telluride at a Cd lattice site (In_{Cd}) are donors. The converse applies for $*A_2$ in AB.

Control of the chemical potential of native point defects and charge carriers with the aid of the coexisting gaseous phase

If, at a certain temperature T and pressure P, the solid phase AB, containing the point defects $*A_k$ and $*B_l$, is in equilibrium with a gas phase containing the molecules A_v and B_w, then:

$$\mu(*A_k) = \frac{\nu_k}{v}\,\mu(A_v)_{gas} \ , \quad . \ . \ . \ . \ . \ . \quad (60)$$

$$\mu(*B_l) = \frac{\nu_l'}{w}\,\mu(B_w)_{gas} \ . \quad . \ . \ . \ . \ . \ . \quad (61)$$

For ideal gases (i.e. at a not unduly high pressure) the equations are:

$$\mu(A_v)_{gas} = \mu^0(A_v)_{gas} + RT \ln P(A_v) \ , \ . \quad (62)$$

$$\mu(B_w)_{gas} = \mu^0(B_w)_{gas} + RT \ln P(B_w) \ , \ . \quad (63)$$

where $P(A_v)$ and $P(B_w)$ are the partial vapour pressures of A_v and B_w in the coexisting gas phase. Substituting from (31), (32), (62) and (63) into (60) and (61) we obtain the general equations:

$$[*A_k] = K_{13}\,\{P(A_v)\}^{\nu_k/v} \ , \quad . \ . \ . \ . \quad (64)$$

$$[*B_l] = K_{14}\,\{P(B_w)\}^{\nu_l'/w} \ . \quad . \ . \ . \ . \quad (65)$$

Combining these with equations such as (55), (56) and (58) gives equations such as:

$$[*B^-]p = K_{15}\,\{P(B_2)\}^{1/2} \ , \quad . \ . \ . \ . \ . \quad (66)$$

$$[*B^-]p = \frac{K_{16}}{P(A)} \ , \quad . \ . \ . \ . \ . \ . \ . \quad (67)$$

$$[*A^+]n = K_{17}P(A) \ , \quad . \ . \ . \ . \ . \ . \ . \quad (68)$$

$$[*B^{2-}]p^2 = \frac{K_{18}}{P(A)} \ , \text{ etc.} \quad . \ . \ . \ . \ . \ . \ . \quad (69)$$

From the above equations it follows that the chemical potentials and the concentrations of native point defects and charge carriers in the condensed phase (e.g.

crystals) can be determined by measuring the corresponding chemical potentials or partial pressures in the coexisting gas phase. The partial pressures may vary considerably with the exact composition of the solid phase. For example, in many binary compounds AB, the partial equilibrium pressure $P(B_2)$ at a given temperature is many powers of ten higher on the B side of the existence region of AB than on the A side, even when the existence region is very narrow (see *Table II*). This means in principle that only a very rough approximation to the composition of the coexisting vapour phase is required to achieve a very exact composition for the solid compound within the extremely narrow existence region, thus producing accurately defined physical properties. For example, by combining equations (66)-(69) and equation (59) we see that it is possible to choose within certain limits (i.e. corresponding to those of the existence region) the type and the magnitude of the electrical conductivity of a semiconducting compound.

This will be further illustrated with a simple example in which it is assumed that the only atomic native point defects occurring in semiconducting AB are $*A^+$ and $*B^-$; in other words, the ionization energies of *A and *B are assumed to be small compared with kT. The relations for this case are:

$$[*A^+]n = K_{17}P(A), \quad \ldots\ldots \quad (68)$$

$$np = K_i, \quad \ldots\ldots \quad (59)$$

$$[*A^+]\,[*B^-] = K_{12}, \quad \ldots\ldots \quad (58)$$

$$[*A^+] + p = [*B^-] + n. \quad \ldots\ldots \quad (54)$$

If the equilibrium constants K_{17}, K_i and K_{12} are known, we can calculate $[*A^+]$, $[*B^-]$, n and p with the aid of these four equations as a function of the coexisting partial vapour pressure $P(A)$, and plot the result in a graph. This is a somewhat complicated procedure. It can be simplified by dividing the graph into three parts (*fig. 9*), so that reduced neutrality conditions can be used to a good approximation. In area *I*, with very low $P(A)$, where $[*A^+]$ and n are small (see equation 68), the reduced electroneutrality relation becomes $p = [*B^-]$. Similar considerations apply to area *III*, with high $P(A)$: $n = [*A^+]$. For area *II* in between we either have $n = p$ or $[*A^+] = [*B^-]$, depending on

Table II. The existence region of a number of binary compounds [7]. Columns 5 and 6 show the order of magnitude of the partial equilibrium pressures of O_2, S_2 and Te_2 in the coexisting gas phase at the metal-rich and metalloid-rich boundaries of the existence region, $P(B_2)^{AB(A)}$ and $P(B_2)^{AB(B)}$ respectively. The last column gives the ratio of these pressures, Q.

Compound	T in °C	Existence region in at.% of the metalloid component	Width of existence region in at.%	$P(B_2)^{AB(A)}$ in torr	$P(B_2)^{AB(B)}$ in torr	Q
FeO	1230	51.2 -54.0	2.8	10^{-8}	10^{-5}	10^3
Fe_3O_4	1130	57.1 -57.4	0.3	10^{-7}	10^{-1}	10^6
SnS	730	50.00 -50.02	0.02	10^{-6}	10^2	10^8
CdTe	750	49.9999-50.0002	0.0003	10^{-6}	10^3	10^9
Cu_2S	600	33.3 -36.5	3.2	10^{-10}	10^3	10^{13}

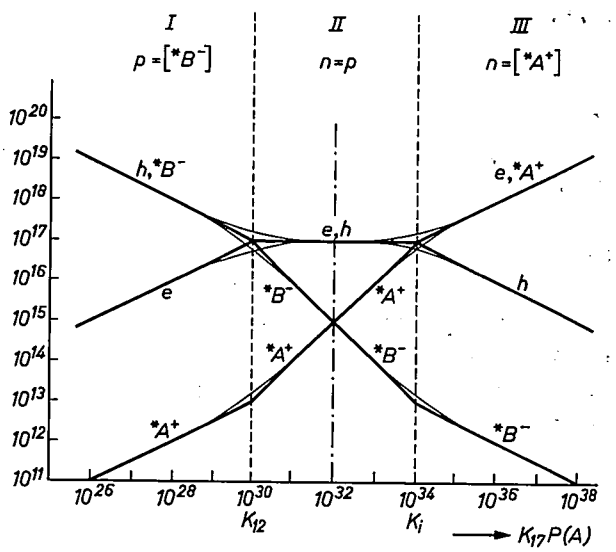


Fig. 9. Calculated concentrations (in numbers/cm³) for electrons e, holes h and two native point defects $[*A^+]$ and $[*B^-]$ as a function of the partial pressure $P(A)$ (at constant temperature). The calculation is based on a number of assumptions explained in the text. The calculation is valid *inside* the narrow existence region of a semiconductor (the chain-dotted line gives the stoichiometric composition). For simplicity it is assumed that the reduced neutrality condition $p = [*B^-]$ holds in area *I*, and that $n = p$ in area *II* and $n = [*A^+]$ in area *III*. The thin lines indicate the corrections to be applied if the complete neutrality condition $[*A^+] + p = [*B^-] + n$ is taken into account. Representative values were taken for this calculation. It illustrates (see also fig. 10) that large changes in the partial pressures are required for small changes in the concentrations of the point defects and charge carriers, making accurate control of these concentrations possible.

[5] See F. H. Stieltjes and L. J. Tummers, Philips tech. Rev. **17**, 233, 1955/56.

[6] F. G. Fumi and M. P. Tosi, Disc. Faraday Soc. **23**, 92, 1957.

[7] The data shown in the table have been taken from the following sources:
for FeO and Fe_3O_4: L. S. Darken and R. W. Gurry, J. Amer. Chem. Soc. **68**, 798, 1946;
for SnS: H. Rau, J. Phys. Chem. Solids **27**, 761, 1966;
for CdTe: D. de Nobel, Philips Res. Repts. **14**, 361, 430, 1959;
for Cu_2S: H. Rau, J. Phys. Chem. Solids **28**, 903, 1967.

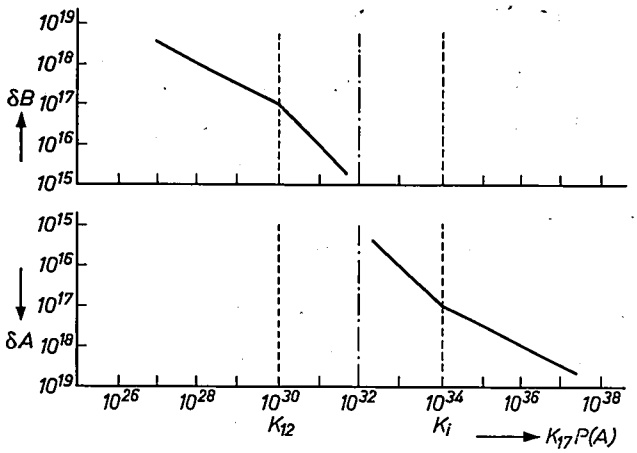


Fig. 10. Deviations δA and δB from the stoichiometric composition as a function of the partial pressure P(A) for a semiconducting compound AB, characterized by fig. 9. The deviations are given in the number of atoms/cm³. The figure shows that P(A) must increase by a factor of 10^8 in order to change AB with a surplus of 10^{18} B atoms (10^{-2} at.% B) into AB with a surplus of 10^{-2} at.% A.

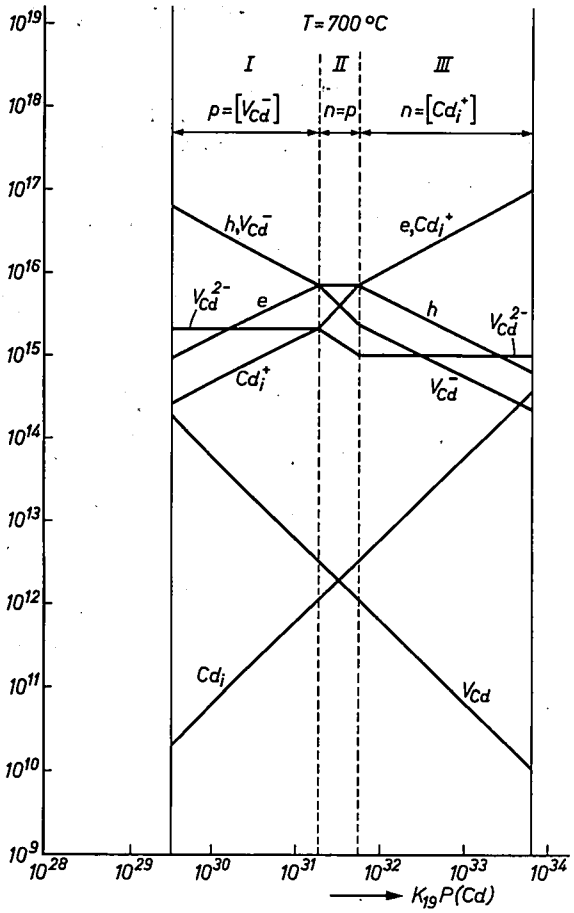


Fig. 11. Figure corresponding to fig. 9 for the concentrations of electrons, holes and native point defects of CdTe at 700 °C. The vertical lines indicate the boundaries of the existence region. (The calculation will be given in part III.)

whether K_i is larger or smaller than K_{12}. Taking the arbitrary but fairly typical values of 10^{34} cm^{-6} and 10^{30} cm^{-6} for K_i and K_{12} (which means that the compound at the exact stoichiometric composition contains 10^{15} native point defects $*A^+$ and 10^{15} native point defects $*B^-$, that is about 10^{-5}%), then area *II* is characterized by $p = n$. The vertical axis of fig. 9 indicates the concentrations of native point defects and of charge carriers, and the horizontal axis indicates $K_{17}P$(A). The thick lines represent the concentration curves calculated with the reduced neutrality conditions; the thin lines that round off the corners give the corrections to be applied if equation (54) is taken fully into account.

The deviation from the stoichiometric composition δA (with surplus A) and δB (with surplus B) is given by: $\delta A = [*A^+] - [*B^-] = -\delta B$. This deviation may be derived directly from fig. 9, and is shown in *fig. 10* as a function of P(A). Fig. 10 illustrates the case where the partial pressure P(A) in the gas phase coexisting with solid AB has to increase by a factor of 10^8 in order to change AB with a surplus of 10^{18} B atoms cm^{-3}, i.e. about 10^{-2} at.% (atomic per cent) into AB with 10^{-2} at.% of surplus A. The existence regions of figs. 9 and 10 are limited (not shown in the figure) by a low P(A) value, where AB ceases to be stable at the given temperature and changes to a phase richer in B, for example B(A), and by a high value at which AB changes to the neighbouring phase richer in A, for example A(B). Typical limiting values for some binary compounds are given in Table II. It can be seen from this that the ratio of these limiting values can be very large, even for compounds with a very narrow existence region.

Fig. 11 shows the concentrations of a number of point defects and charge carriers occurring within the existence region of the semiconductor CdTe at 700 °C as a function of the partial vapour pressure of Cd in the coexisting gas phase.

Summary. The partial vapour pressures in the coexisting gas phase can be used to give accurate control of the concentrations of point defects and charge carriers within the existence region of the solid compound. This is clarified by means of a simple theoretical model of a semiconductor. The result of measurements and calculations for the semiconductor CdTe is given as a practical example.

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

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.

C. S. Aitchison, C. D. Corbey & B. H. Newton: Self-pumped Gunn-effect parametric amplifier.
Electronics Letters **5**, 36-37, 1969 (No. 2). *M*

G. Arlt & W. Hermann: Internal friction by electron hole pair recombination in tellurium.
Solid State Comm. **7**, 75-77, 1969 (No. 1). *A*

R. Astor: Emetteur pour équipement mobile de reportage de télévision.
Acta electronica **11**, 437-450, 1968 (No. 4). *L*

Ph. van Bastelaer: The design of band-pass filters with piezoelectric resonators.
Rev. HF **7**, 193-206, 1968 (No. 7). *B*

J. Basterfield: Domain structure and the influence of growth defects in single crystals of yttrium iron garnet.
J. appl. Phys. **39**, 5521-5526, 1968 (No. 12). *M*

G. Blasse & A. Bril: Phosphors based on lanthanide oxysulphates (Ln_2SO_6).
Philips Res. Repts. **23**, 461-468, 1968 (No. 6). *E*

G. Blasse & A. Bril: Energy transfer between Eu^{2+} ions in nonequivalent sites in strontium-silicate-phosphate.
Physics Letters **28A**, 572-573, 1969 (No. 8). *E*

G. Blasse, A. Bril & J. de Vries: Luminescence of alkaline-earth borate-phosphates activated with divalent europium.
J. inorg. nucl. Chem. **31**, 568-570, 1969 (No. 2). *E*

R. Bleekrode: Absorption spectroscopy of Fe(I) in low-pressure oxyacetylene flames.
Appl. Spectroscopy **22**, 536-539, 1968 (No. 5, Part 1). *E*

R. Bleekrode: Intensity alternations in the second positive bands ($C^3\Pi_u$-$B^3\Pi_g$) of $^{14}N_2$ and $^{15}N_2$.
Physica **41**, 24-26, 1969 (No. 1). *E*

R.-D. Böhnke & H. Zimmer: Technologie bei der Herstellung von supraleitenden Bleiresonatoren.
J. less-common Met. **17**, 235-241, 1969 (No. 2). *H*

G. A. Bootsma & F. Meyer: Ellipsometry in the submonolayer region.
Surface Sci. **14**, 52-76, 1969 (No. 1). *E*

J. C. Brice: Analysis of the temperature distribution in pulled crystals.
J. Crystal Growth **2**, 395-401, 1968 (No. 6). *M*

R. Brun: Caméra pour équipement portable de télévision.
Acta electronica **11**, 361-368, 1968 (No. 4). *L*

R. Brun: Générateur de signaux de synchronisation pour la caméra de prise de vues.
Acta electronica **11**, 369-397, 1968 (No. 4). *L*

R. Brun: Dispositifs d'intersynchronisation pour caméra portable de télévision.
Acta electronica **11**, 399-429, 1968 (No. 4). *L*

K. Bulthuis & R. Tree: Annealing behaviour of *p*-type layers formed by ion implantation of gallium in silicon.
Physics Letters **28A**, 558-560, 1969 (No. 8). *M*

K. H. J. Buschow & A. S. van der Goot: The intermetallic compounds in the gadolinium-cobalt system.
J. less-common Met. **17**, 249-255, 1969 (No. 3). *E*

K. H. J. Buschow & W. A. J. J. Velge (Philips Philite and Metalware Manufacturing Division, Eindhoven): Permanent magnetic materials of rare earth cobalt compounds.
Z. angew. Physik **26**, 157-160, 1969 (No. 2). *E*

F. M. A. Carpay: Vanadium chalcogenides and the nickel-arsenide structure.
Thesis, Utrecht 1968. *E*

H. B. G. Casimir: On Maxwell's equations and the theory of optical rotation.
Special book, Prof. Kastler's 60th birthday, 1968, p. 185-190. *E*

J. Cayzac: Equipement portatif de prise de vues de télévision.
Acta electronica **11**, 351-360, 1968 (No. 4). *L*

M. G. Collet: Depositing silicon nitride layers at low temperature using a photochemical reaction.
J. Electrochem. Soc. **116**, 110-111, 1969 (No. 1). *E*

L. Crousel & J. Neirynck: Approximation of the ideal filter by truncated expansions in Chebyshev polynomials.
Proc. int. Symp. on network theory, Belgrade 1968, p. 320-329. *B*

B. J. Curtis: The preparation of shear-mode-evaporated cadmium sulfide ultrasonic transducers by silver diffusion.
J. appl. Phys. **40**, 433-434, 1969 (No. 1). *M*

H. J. van Daal & K. H. J. Buschow: Influence of *s-f* exchange interaction on electrical conduction in rare-earth dialuminides.
Solid State Comm. **7**, 217-221, 1969 (No. 1). *E*

F. Desvignes & J. Ohnet: Caractérisation et mesure des propriétés des capteurs de rayonnement.
Techniques Philips 1968, No. 6, 2-19. *L*

G. Diemer & A. de Kool (Natuur en Techniek): Universiteit en industrie. De betekenis van de fysicus in een onderneming.
Natuur en Techniek **37**, 58-65, 1969 (No. 2). *E*

N. Dorrepaal & J. A. W. van Laar (Philips Radio, Television and Record-playing Equipment Division, Eindhoven): Cold-check and physical properties of polyester lacquer coatings.
IXth Fatipec Congress, Brussels 1968, sec. 2, p. 116-124.

Y. Duflos: Dispositif d'antenne d'émission à pointage automatique.
Acta electronica **11**, 457-477, 1968 (No. 4). *L*

G. Engelsma: The influence of light of different spectral regions on the synthesis of phenolic compounds in gherkin seedlings, in relation to photomorphogenesis, V. The temperature dependence.
Acta bot. neerl. **17**, 499-505, 1968 (No. 6). *E*

F. C. Eversteyn & H. L. Peek: Preparation and stability of enhancement *n*-channel MOS transistors with high electron mobility.
Philips Res. Repts. **24**, 15-33, 1969 (No. 1). *E*

N. V. Franssen: Sur l'amplification des champs acoustiques.
Acustica **20**, 315-323, 1968 (No. 6). *E*

R. Genève: La mobilité dans le reportage de télévision.
Acta electronica **11**, 337-349, 1968 (No. 4). *L*

A. A. van der Giessen: Chemical and physical properties of iron(III)-oxide hydrate.
Thesis, Eindhoven 1968. *E*

A. H. Gomes de Mesquita: Polytypism in silicon carbide.
J. Crystal Growth **3/4**, 747-750, 1968. *E*

D. Gossel: Der steuerbare Gyrator — ein Element zur Synthese nichtlinearer Systeme.
Nachrichtentechn. Z. **21**, 778-789, 1968 (No. 12). *H*

H. C. de Graaff: Approximate calculations on the spreading resistance in multi-emitter structures.
Philips Res. Repts. **24**, 34-52, 1969 (No. 1). *E*

H. C. de Graaff: Two new methods for determining the collector series resistance in bipolar transistors with lightly doped collectors.
Philips Res. Repts. **24**, 70-81, 1969 (No. 1). *E*

G. J. van Gurp: Flux motion and noise in superconductors.
Thesis, Eindhoven 1969. *E*

K. H. Härdtl & H. Rau: PbO vapour pressure in the $Pb(Ti_{1-x}Zr_x)O_3$ system.
Solid State Comm. **7**, 41-45, 1969 (No. 1). *A*

F. W. Harrison & G. K. Lang: Attempts to prepare $LiF.Fe_2O_3$.
J. Phys. Soc. Japan **25**, 1609-1610, 1968 (No. 6). *M*

J. Hasker: Initial-velocity effects in cathode-ray tubes.
Thesis, Eindhoven 1969. *E*

E. E. Havinga: Band structure and superconductivity of non-transition metals.
Proc. 11th int. Conf. on low temperature physics, St. Andrews 1968, Vol. 2, p. 756-759. *E*

E. P. Honig, J. H. Th. Hengst & P. Hirsch-Ayalon (Weizmann Institute of Science, Rehovoth, Israel): Observations on the development and structure of a precipitate-membrane.
Berichte Bunsenges. phys. Chemie **72**, 1231-1242, 1968 (No. 9/10). *E*

K. Hoselitz: The crystals we need.
J. Crystal Growth **3/4**, 5-12, 1968. *M*

B. B. van Iperen: Impedance relations in a diode waveguide mount.
IEEE Trans. **MTT-16**, 961-963, 1968 (No. 11). *E*

R. E. Jesse: The Soret effect and the solidification of binary single- and two-phase systems.
J. Crystal Growth **5**, 132-134, 1969 (No. 2). *E*

G. H. Jonker: Semiconducting properties of lanthanum-cobalt oxide.
Philips Res. Repts. **24**, 1-14, 1969 (No. 1). *E*

T. J. van Kessel: An integrated operational amplifier with novel HF behavior.
IEEE J. Solid-State Circuits **SC-3**, 348-352, 1968 (No. 4). *E*

F. M. Klaassen & J. Prins: Noise of VHF and UHF MOS tetrodes.
Philips Res. Repts. **23**, 478-484, 1968 (No. 6). *E*

J. Liebertz: Einkristallzüchtung von Wismutgermanat ($Bi_4(GeO_4)_3$).
J. Crystal Growth **5**, 150, 1969 (No. 2). *A*

J. M. Noothoven van Goor & H. Zijlstra: Structural and magnetic properties of a Bi-MnBi composite.
J. appl. Phys. **39**, 5471-5474, 1968 (No. 12). *E*

C. van Opdorp: Capacitance-voltage relations of Schottky and *p-n* diodes in the presence of both shallow and deep impurities.
Phys. Stat. sol. **32**, 81-89, 1969 (No. 1). *E*

J. A. Pals & H. C. de Graaff: On the behaviour of the base-collector junction of a transistor at high collector current densities.
Philips Res. Repts. **24**, 53-69, 1969 (No. 1). *E*

R. F. Pearson, A. D. Annis & P. Kompfner: Photomagnetic anneal properties of silicon-doped yttrium iron garnet.
Phys. Rev. Letters **21**, 1805-1807, 1968 (No. 27). *M*

A. Rabenau, H. Rau & G. Rosenstein: Darstellung und Eigenschaften von $AuTe_2J$.
Angew. Chemie **81**, 148, 1969 (No. 4). *A*

H. Rau & A. Rabenau: Hydrothermal growth of some elements.
J. Crystal Growth **3/4**, 417-421, 1968. *A*

C. J. M. Rooymans & W. F. Th. Langenhoff: Hydrothermal growth of single crystals and phase width of tetragonal lead monoxide.
J. Crystal Growth **3/4**, 411-416, 1968. *E*

H.-D. Rüpke: Magnetisch abstimmbare Mikrowellenfilter mit hoher Güte.
Z. angew. Physik **26**, 32-35, 1969 (No. 1). *H*

G. M. Schaeffer & M. H. van Maaren: Critical field and specific heat measurements on spinel-type superconductors.
Proc. 11th int. Conf. on low temperature physics, St. Andrews 1968, Vol. 2, p. 1033-1036. *E*

P. Schagen: Fortschritte auf dem Gebiet der Bildwandler und Bildverstärker.
Umschau in Wiss. u. Technik **69**, 85, 1969 (No. 3). *M*

B. Schneider: Interaction between spin waves and electrons in a hybrid structure of YIG and InSb.
Appl. Phys. Letters **13**, 405-407, 1968 (No. 12). *H*

L. A. Æ. Sluyterman & M. J. M. de Graaf: The activity of papain in the crystalline state.
Biochim. biophys. Acta **171**, 277-287, 1969 (No. 2). *E*

L. A. Æ. Sluyterman & B. G. Wolthers (Laboratorium voor Structuurchemie, Groningen): A tentative mechanism of papain action.
Proc. Kon. Ned. Akad. Wetensch. **B 72**, 14-16, 1969 (No. 1). *E*

P. J. Strijkert & J. S. Sussenbach: Arginine metabolism in Chlamydomonas reinhardi. Evidence for a specific regulatory mechanism of the biosynthesis.
European J. Biochem. **8**, 408-412, 1969 (No. 3). *E*

J. S. Sussenbach & P. J. Strijkert: Arginine metabolism in Chlamydomonas reinhardi. On the regulation of the arginine biosynthesis.
European J. Biochem. **8**, 403-407, 1969 (No. 3). *E*

J. Tillack & Renate Kaiser: Wolfram(IV)-oxidbromid, $WOBr_2$.
Angew. Chemie **81**, 149-150, 1969 (No. 4). *A*

W. Tolksdorf: Growth of yttrium iron garnet single crystals.
J. Crystal Growth **3/4**, 463-466, 1968. *H*

J. O. Voorman: RC one-ports.
Philips Res. Repts. **23**, 485-515, 1968 (No. 6). *E*

Q. H. F. Vrehen, A. Broese van Groenou & J. G. M. de Lau: Relaxation of ferromagnetic precession by excitation of spin-waves in a polycrystalline ferrite.
Solid State Comm. **7**, 117-121, 1969 (No. 1). *E*

W. L. Wanmaker, A. H. Hoekstra, J. L. Ouweltjes & J. G. Verriet (Philips Lighting Division, Eindhoven): Some experiments concerning the solubility of Sb^{3+} in calcium chlorapatite.
Philips Res. Repts. **23**, 469-477, 1968 (No. 6).

C. H. Weijsenfeld: The Hall effect in type II superconductors.
Proc. 11th int. Conf. on low temperature physics, St. Andrews 1968, Vol. 2, p. 947-950. *E*

K. Weiss: Fehlordnung der Kationen in β-Ag_2S.
Z. Naturf. **24a**, 184, 1969 (No. 1). *E*

F. W. Willmott: An informal symposium of the Gas Chromatography Discussion Group.
J. Gas Chromatogr. **6**, 569-571, 1968 (No. 12). *M*

H. J. de Wit: Some numerical calculations on Holstein's small-polaron theory.
Philips Res. Repts. **23**, 449-460, 1968 (No. 6). *E*

W. J. Witteman: High-power single-mode CO_2 laser.
IEEE J. Quantum Electronics **QE-4**, 786-788, 1968 (No. 11). *E*

P. Zalm: Thermionic cathodes.
Adv. in Electronics and Electron Phys. **25**, 211-275, 1968. *E*

A. L. Zijlstra & A. J. Burggraaf (Philips Glass Division, Eindhoven): Fracture phenomena and strength properties of chemically and physically strengthened glass, II. Strength and fracture behaviour of chemically strengthened glass in connection with the stress profile.
J. non-cryst. Solids **1**, 163-185, 1969 (No. 2).

Contents of Electronic Applications **29**, No. 1, 1969:

C. J. Boers: Recent developments in circuits and transistors for television receivers, IX. E.h.t. generation from the line flyback pulse (p. 1-15).
B. J. M. Overgoor: An operational amplifier with differential FET input (p. 16-28).
W. Ebbinge & D. C. de Ruiter: Improved speed control for a permanent magnet motor (p. 29-34).

Contents of Mullard Technical Communications **10**, No. 98, 1969:

R. W. G. Clarke & I. P. Russell: Spectroscopic detector amplifier for 100 Hz to 10 kHz (p. 230-233).
P. Bissmire: Single system i.f. amplifier for U.K. 625-line system (p. 234-243).

PHILIPS TECHNICAL REVIEW

VOLUME 30, 1969, No. 5

A precision lathe with hydrostatic bearings and drive

H. J. J. Kraakman and J. G. C. de Gast

Philips Research Laboratories first took up work on precision lathes many years ago because of a need for turned surfaces of optical quality. At first all the work that was required could be handled by a single lathe with special modifications, but eventually the demand became so heavy that more machines had to be built. This provided a suitable opportunity for introducing an entirely new design, and the first precision lathe built to the new design is now in use in the Laboratories. A unique feature of the machine is the exclusive use of hydrostatic bearings and drive; some of the bearings are made to have infinite stiffness by a special control system. The movements of workpiece and tool are hydraulically controlled to an accuracy previously unattainable.

About thirty years ago the need arose for a powerful, wide-angle lens system that was free from spherical aberration, for projection-television receivers. These requirements were met by the Schmidt optical system, which consists of a spherical mirror with a correction plate. It was found possible to make this correction plate from gelatine, using a circularly symmetric mould which was copied from a jig on a lathe [1]. The required surface quality was obtained by using a diamond-tipped tool. A further improvement in finish was achieved in later years by changing the headstock bearing of the lathe for a precision journal bearing with clearances less than 1 μm [2].

Because of the high dimensional accuracy and good finish that could be obtained with this lathe when a good cutting tool was used, it was found to be invaluable for making objects such as resonant cavities, lenses and mirrors for lasers, parabolic reflectors, pole-pieces for electron microscopes and electrodes for the miniature "Plumbicon" camera tube. Eventually there was more work than one machine could handle and it was decided to design a completely new precision lathe, which would also be more suitable for workshop conditions. This lathe is described in this article.

The main spindle and carriages in this lathe are supported by *hydrostatic bearings*, that is to say bearings in which the oil film is maintained by the supply of oil under pressure. In this type of bearing the thickness of the oil film between the bearing surfaces, and hence the position of the moving part, is independent of their relative speed. Moreover, under certain conditions, the clearances can be quite large (20 to 40 μm), which means that large fluctuations in the temperature of the hydrostatic bearings cannot cause them to seize up. Since there is never any metallic contact between the moving parts of a hydrostatic bearing, there is no "stick-slip" or wear.

Hydrostatic forces are also used for driving the carriages. A *linear hydraulic motor*, with piston and cylinder, was found to be a good choice of drive since it has no periodic element. Periodic systems — such as the conventional lead-screw and nut — have the disadvantage that periodic irregularities (known as "ghosts") are produced in the machined surface [3]. The headstock spindle is driven by a *rotary hydraulic motor*. The speed

Ir. H. J. J. Kraakman is with Philips Research Laboratories, Eindhoven; Ir. J. G. C. de Gast, previously with Philips Research Laboratories, is now with the Philips Electronics Components and Materials Division (Elcoma), Eindhoven.

[1] H. Rinia and P. M. van Alphen, The manufacture of correction plates for Schmidt optical systems, Philips tech. Rev. **9**, 349-356, 1947/48.

[2] The diamond-tipped tool and the high-precision bearing were made by L. M. Leblans; see L. M. Leblans, A high-precision lathe headstock, Philips tech. Rev. **19**, 68-69, 1957/58.

[3] W. R. Horsfield, Ruling engine with hydraulic drive, Appl. Optics **4**, 189-193, 1965. The "ghosts" occur even when the lead-screw is hydrostatically supported in the nut: J. H. Rumbarger and G. Wertwijn, Hydrostatic lead screws, Machine Design **40**, No. 9, 218-224, 1968.

of this kind of motor can be regulated easily, it is small enough to be built into the headstock, and it can be made fairly insensitive to overload.

The use of these hydrostatic components means that oil continuously circulates through and around all the parts of the lathe, so that the lathe quickly assumes the same temperature as the oil. This makes effective temperature control possible. Because of this temperature control and the absence of wear in the mating surfaces of the hydrostatic bearings, and also because the headstock motor is not sensitive to overload, all kinds of turning can be done on the lathe without any gradual deterioration in the highest standard of finish that can be attained; in addition, the time needed to reach this degree of finish is negligibly short.

In this hydrostatic bearing and drive system the control of pressures and rates of flow plays an important part. To obtain work with a high standard of finish it is of course essential to keep the speeds of movement of tool and work accurately constant. We have developed the elements needed for these three functions — bearing, drive and control — to the required degree of accuracy and it can be expected that the new elements will be of considerable interest in many other applications besides the lathe described here. Outstanding among these elements are the *double-film hydrostatic bearings* for the carriages, which can be made to have infinite stiffness by means of a special method of pressure control, the *linear motor*, the *rotary hydraulic motor* and the *control valves*. A method has been found which largely compensates for the unwanted forces that arise because of the oil flow through these valves. In the following these four components will be dealt with at greater length, but first of all we shall give a general description of the design of the lathe as a whole and of the hydraulic system.

Design objectives

The lathe was required for machining work up to 200 mm in length and diameter with a surface finish of optical quality. A dimensional tolerance of 1 μm was given as a target figure for the design, with a maximum out-of-roundness of 0.1 μm in the finished work. These tolerances, which were laid down purely as a target, have been achieved in practice; with a readily machinable material and a good (diamond-tipped) tool it is likely that even closer tolerances could be met. However, to meet dimensional tolerances of this order it is essential that the temperature of the work remains constant: for example, a piece of brass 100 mm long expands by 1.9 μm for a temperature increase of 1 °C. Where dimensional accuracy is important, it will therefore be necessary to use the lathe in a controlled-temperature environment. This, however, is not necessary to achieve a high surface finish, since the control of the temperature of the circulating oil makes the lathe sufficiently independent of the ambient temperature. Achieving a high surface finish was incidentally the most difficult of the objectives, because the closer one gets to a surface of optical quality the more every flaw shows up. A scratch only 0.1 μm deep is clearly visible.

To avoid all causes of surface flaws, the two carriages that move the tool longitudinally and radially are not mounted one above the other as they are in a standard lathe. Instead, the toolholder is mounted on a traversing carriage and the headstock spindle on an identical carriage which moves longitudinally, i.e. parallel to the axis of rotation. The tailstock is mounted on a third carriage in line with the headstock, and the centring cone is fixed (i.e. it does not rotate). The three carriages are mounted on a lathe bed, with which they form a compact and rigid assembly. A plan view of the lathe is shown in *fig. 1*.

The linear hydraulic motors that apply the longitudinal and traversing motion to the carriages for the headstock spindle and for the toolholder are mounted in line with the carriages to ensure well-defined movement and a rigid construction. The motor housing, the cylinder *Cyl* (*fig. 2*), is fixed to the slideway *S* and the piston rod is attached directly to the carriage *C*.

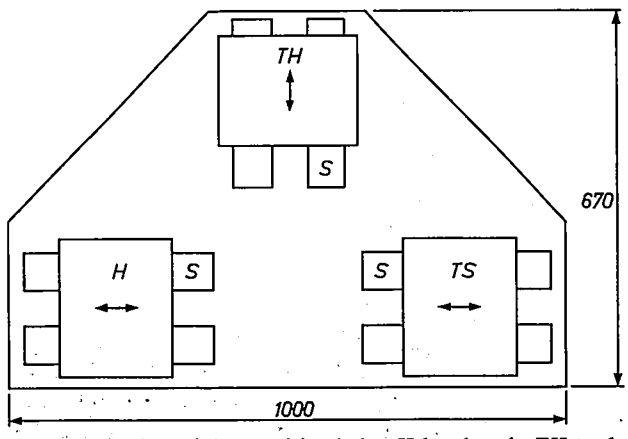


Fig. 1. Plan view of the precision lathe. *H* headstock. *TH* toolholder. *TS* tailstock. *S* slideways of the carriages which carry these three main units. The dimensions of the lathe bed are given in millimetres.

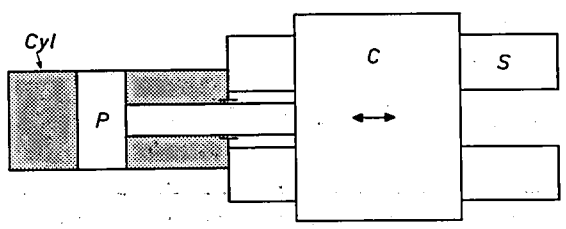


Fig. 2. The drive of a carriage *C* by the oil pressure in cylinder *Cyl*, working on piston *P*. *S* is the slideway.

The rotary hydraulic motor that drives the main spindle is incorporated in the headstock which, in a standard lathe, would be a fixed unit. In our lathe the headstock unit is mounted on the carriage that moves longitudinally.

The photograph in *fig. 3* clearly shows the three main units: headstock, toolholder and tailstock. The front plate of the lathe contains the operating controls for setting positions and speeds. Behind the lathe there are two decimal counters, which show the displacements of the longitudinal and cross carriages to within an accuracy of 0.5 μm. These displacements are measured by means of an optical system [4].

[4] An article on this will shortly be appearing in this journal.

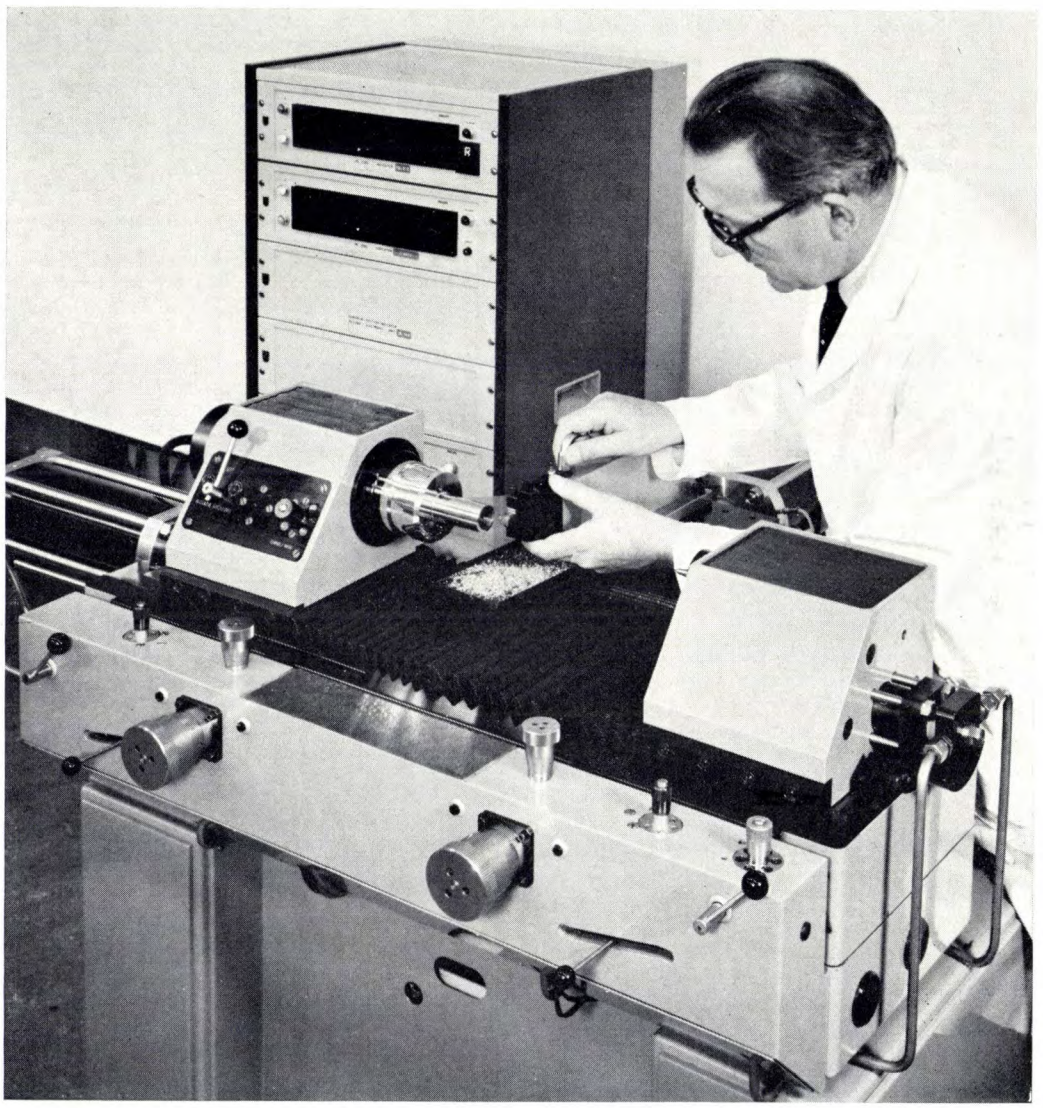


Fig. 3. The precision lathe. Control valves for adjusting the speed of the longitudinal and cross movements are mounted on the front plate. Instruments that indicate the displacement in the longitudinal and radial directions to an accuracy of 0.5 μm can be seen at the rear.

Hydraulic system

A diagram of the hydraulic system of the precision lathe is given in *fig. 4*, with the conventional symbols. The diagram presents the three most important assemblies of the lathe, the headstock unit *H*, the toolholder unit *TH* and the tailstock unit *TS*; *P* is the pump unit. Each unit has its own linear motor (*1*, *2* and *3*). The linear motor *1* of the headstock unit is connected to the variable flow-control valve *9* through the directional valve *7*: the valve *9* regulates the rate of flow at which the cylinder of *1* fills up or empties and thus controls the speed of movement of the piston. The same arrangement is found in the toolholder unit *TH*. The system used for the tailstock unit is different, since here the

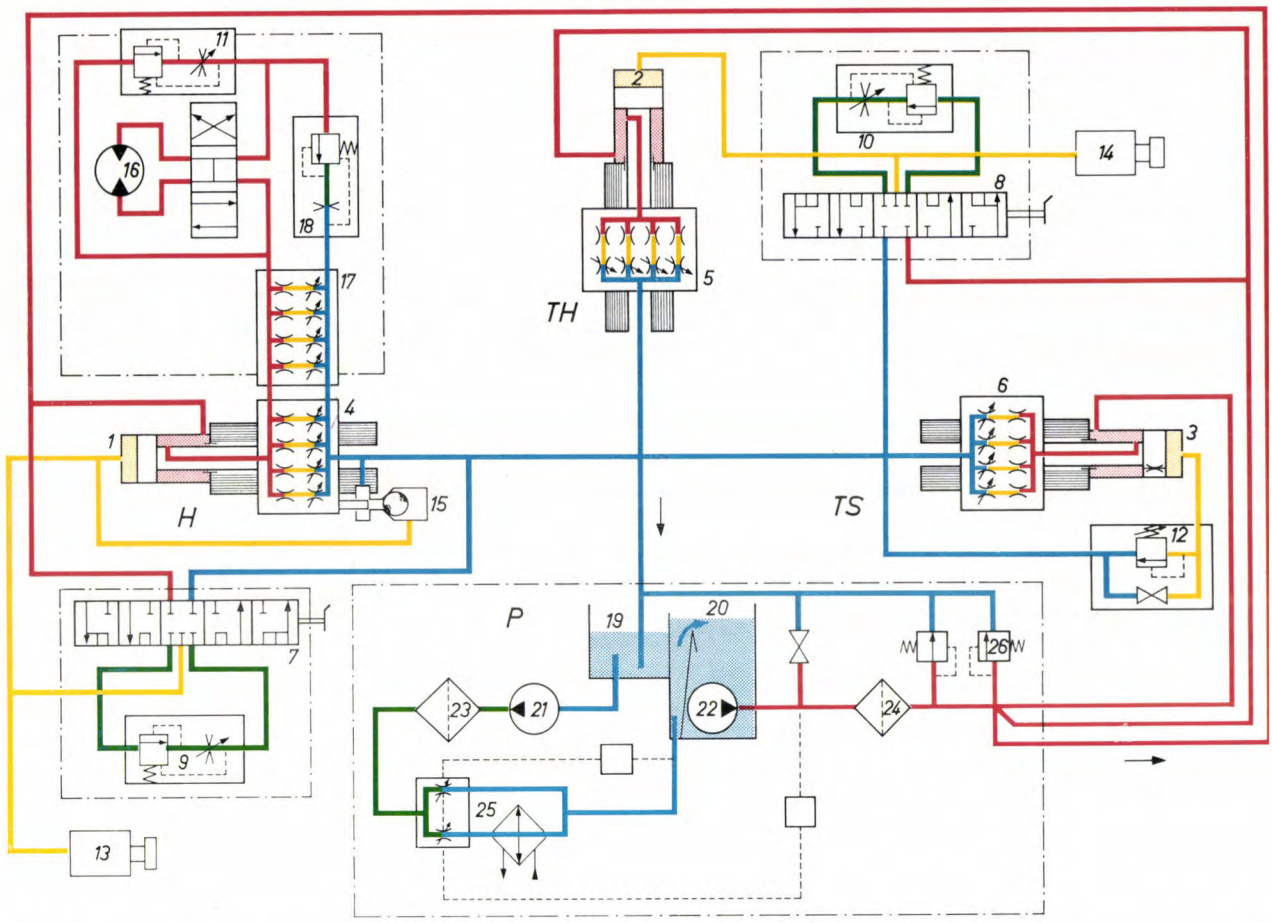


Fig. 4. Diagram of the hydraulic system of the precision lathe. *H* headstock unit. *TH* toolholder unit. *TS* tailstock unit. *P* pump unit. The colours indicate the various oil pressures: *red* = supply pressure p_s, *yellow* = $\frac{1}{2}p_s$, *green* < $\frac{1}{2}p_s$ but higher than atmospheric pressure, *blue* = atmospheric pressure.
1,2,3 linear hydraulic motors. *4,5,6* carriages with hydrostatic bearings. *7,8* directional valves. *9,10,11* adjustable flow-control valves. *12* adjustable pressure-control valve for the tailstock carriage. *13,14* fine adjustment for the position of the linear motors. *15* backstop. *16* rotary hydraulic motor. *17* hydrostatic bearings of the rotary hydraulic motor. *18* preset flow-control valve. *19* backing reservoir. *20* main reservoir. *21* auxiliary pump. *22* main pump. *23,24* oil filters. *25* temperature-control system. *26* supply-pressure control valve.

linear motor serves to maintain a constant pressure against the workpiece.

The headstock unit *H* also includes the rotary hydraulic motor *16*. This has a parallel branch with a variable flow control *11*. The sum of the flows through both branches is kept constant by a fixed flow-control valve *18*. The motor speed is controlled by varying the flow through *11*; since the sum of the flows through *16* and *11* remains constant, the flow through *16* then varies in the opposite direction, changing the motor speed accordingly. An advantage of a total flow independent of the motor speed is that any pressure drop in the connections always remains the same, so that the pressure level at all points is independent of the speed of the motor.

The pump unit *P* supplies the whole hydraulic system with clean oil at the required temperature and at a pressure of 40 bars [5]. This pressure is rather lower than is usual in hydraulic systems. The advantage of this is that the inevitable pressure fluctuations due to the pump are then smaller and easier to smooth out. The amplitude of these pressure fluctuations depends not only on the pressure level but also on the presence of air in the oil. This air has an easy opportunity to escape in the main reservoir *20*; the oil returned to the backing reservoir *19* is pumped by means of an auxiliary pump *21* via a filter *23* and a temperature-control system with heat exchanger *25* to the main reservoir *20*, where it emerges just below the surface of the oil through a diffuser opening out in the form of a funnel, in which turbulent oil comes to rest. The main pump *22* supplies oil at the rate of 20 litres per minute to the three main units; a pressure-control valve *26* at the branching point of the three supply lines keeps the supply pressure constant.

The hydrostatic bearings

In a hydrostatic bearing the parts *I* and *II* (see *fig. 5*) are separated by a fluid film whose pressure p_r enables it to support the load *F*. Since the bearing is not sealed,

fluid can escape, the outflow resistance R_2 being dependent upon the thickness of the film (the gap height) h, the length l and the width b. The fluid (usually oil) is supplied from an external source through a series resistance R_1, and at the centre of the bearing the clearance is usually deepened to form a recess [6].

The fluid film can be built up by the bearing itself through the pumping action due to the relative movement of the bearing surfaces. Bearings of this type, which include the Leblans precision bearing mentioned earlier [2] and the spiral-groove bearing described earlier in this journal [7], are called hydrodynamic bearings. In a hydrodynamic bearing the fluid film, and hence the position of the mating surfaces, depends on the relative speed. For precision machines this is undesirable, particularly if the speed is sometimes zero, which is often the case for machine-tool carriages.

It is seen that a hydrostatic bearing always consists of two resistances R_1 and R_2 in series, which are supplied from an external source with a fluid (oil) under pressure. In the steady state the fluid flow through the resistances R_1 and R_2 has the same value Φ and the pressure p_r between the resistances depends on R_2 in accordance with the relation:

$$p_r = p_s R_2/(R_1 + R_2), \quad \ldots\ldots \quad (1)$$

where p_s is the pressure of the source.

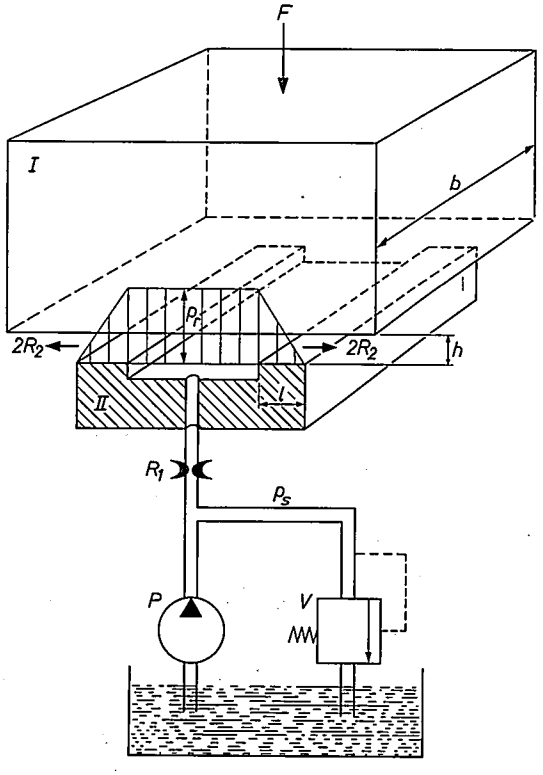


Fig. 5. Diagram of a hydrostatic bearing. I and II the two parts of the bearing. F load. P oil pump. V pressure-control valve. R_1 flow restrictor. R_2 total outflow resistance, formed by two outlet gaps in parallel with width b, length l and height h. p_s supply pressure. p_r pressure in the bearing recess; the pressure decrease across the bearing gaps is shown in the figure.

The flow through the gap is laminar; R_2 consists of the parallel arrangement of two identical bearing gaps each with a resistance $2R_2$, which follows from the equation for fluid flow through a laminar resistance:

$$\tfrac{1}{2}\Phi = \frac{p_r}{2R_2} = \frac{bh^3}{12\eta l}\, p_r. \quad \ldots\ldots \quad (2)$$

In this equation, which we shall encounter several times in this article, η is the viscosity coefficient of the fluid. The laminar flow resistance R_2 is inversely proportional to the third power of the gap height h. The way in which the recess pressure p_r varies with the gap height h can most simply be seen if we write:

$$R_2 = R_{20}\,(h_0/h)^3,$$

where R_{20} and h_0 are the values of R_2 and h for zero load, and then substitute this in (1). We then obtain:

$$p_r = \frac{1}{1 + (R_1/R_{20})\,(h/h_0)^3}\, p_s. \quad \ldots \quad (3)$$

We see from this expression that when the gap height h is reduced because of an increasing load F, the recess pressure p_r in the bearing increases; the higher pressure, acting on the effective bearing surface A, delivers a force which balances the increased load.

In the design of a hydrostatic bearing the dimensions are determined by the required load-carrying capacity and the required stiffness of the bearings. The load-carrying capacity of the hydrostatic bearing refers to the maximum permissible load F_{max} (i.e. the load under which the bearing surfaces are only just clear of one another). Under this load the bearing gap height h is approximately zero, so that the outlet resistance R_2 tends to infinity and the recess pressure p_r becomes equal to the supply pressure p_s, giving $F_{max} = Ap_s$.

The bearing stiffness K is equal to the load change ΔF divided by the resultant displacement $-\Delta h$ of the moving part I towards II:

$$K = -\frac{dF}{dh}.$$

With $F = Ap_r$ we can calculate K from (3). Using an approximation that is permissible if the deviation Δh is small we find:

$$K = \xi \frac{Ap_s}{h_0}. \quad \ldots\ldots \quad (4)$$

[5] 1 bar = 10^5 N/m^2 = 1.020 at. All pressure levels specified here refer to the excess pressure above atmospheric pressure.

[6] H. C. Rippel, Design of hydrostatic bearings, Machine Design 35, 1963 (10 parts). See also H. Opitz, Aufbau und konstruktive Auslegung hydrostatischer Lager, Bericht über die VDW-Arbeitstagung am 4. und 5. Februar 1965, Techn. Hochschule Aachen.

[7] E. A. Muijderman, New forms of bearing: the gas and the spiral groove bearing, Philips tech. Rev. 25, 253-274, 1963/64; E. A. Muijderman, Spiral groove bearings, thesis, Delft 1964.

Here ξ is a dimensionless constant mainly determined by the inlet resistance or restrictor R_1. The relation (4) indicates the requirement to be met if a stiff bearing is to be obtained with a fixed restrictor: the bearing surface and the supply pressure should be large, the gap height should be small.

The same expression, but with a proportionality constant twice as large, also holds for a hydrostatic bearing of the double-film type (*fig. 6*). In the double-film hydrostatic bearing the principle illustrated in fig. 5 is applied to both sides of the moving part *I*. The symmetrical construction thus obtained considerably improves the running and the temperature dependence of the bearing.

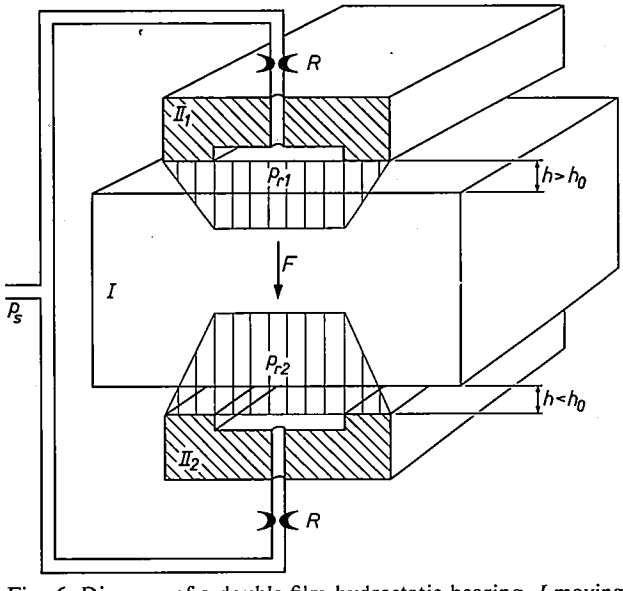


Fig. 6. Diagram of a double-film hydrostatic bearing. *I* moving part. II_1 and II_2 stationary parts. *F* load. *R* flow restrictors. *h* bearing gap height. h_0 value of *h* when $F = 0$. p_s supply pressure. p_{r1} and p_{r2} pressures in the bearing recesses.

The way in which the type of restrictor employed affects the bearing stiffness can clearly be seen from the graph in *fig. 7*. The dashed line in this graph represents the condition which the rate of flow Φ through the bearing should meet for infinite stiffness, i.e. in a bearing where the gap height *h* is independent of the load. If in such a case a force is applied to the part *I*, the flow rate to the recess with the higher pressure must increase and the rate to the other recess must decrease in order to obtain the necessary pressure difference. The solid curves indicate the variation of the flow rate Φ with the recess pressure p_r for four types of flow restrictor. Three of these are commonly used in hydrostatic bearings [6]: the laminar-flow resistance (*L*), the turbulent-flow resistance (*T*) and a constant-flow device (*C*). Using these restrictors it is not possible, as can be

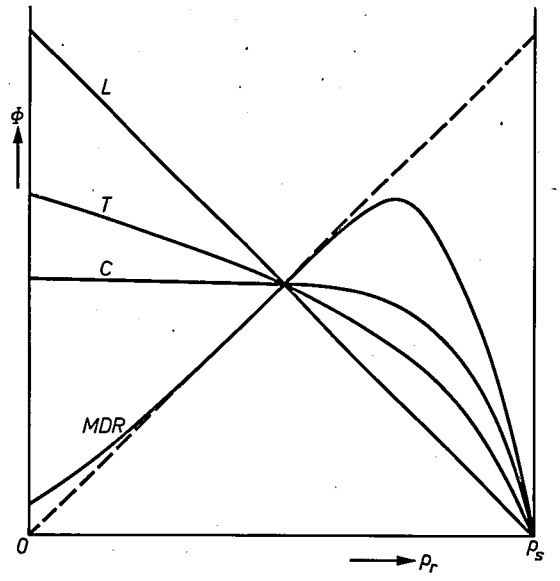


Fig. 7. The oil flow Φ through a hydrostatic bearing as a function of the pressure p_r in the bearing recess for different types of flow restrictor: *L* laminar-flow resistance, *T* turbulent-flow resistance, *C* constant-flow device, *MDR* membrane double restrictor. The dashed line indicates how the oil flow should vary with recess pressure in order to give the bearing infinite stiffness. Only the M.D.R. meets this requirement within a specific range.

seen from fig. 7, to make an infinitely stiff bearing. The fourth curve relates to the variable flow restrictor, known as the "membrane double restrictor" (M.D.R.), developed at Philips Research Laboratories [8]. With this device it is possible to meet the requirements for infinite stiffness within a given range.

Unlike the other types of restrictor mentioned, the membrane double restrictor is affected by load variations, with the result that a feedback control system is obtained. *Fig. 8* shows a diagram of the device together

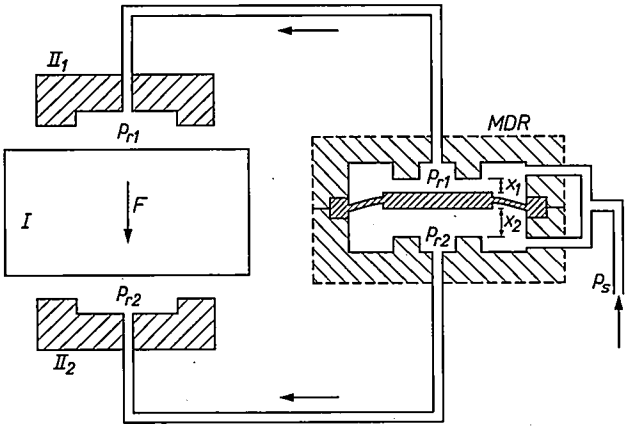


Fig. 8. Diagram of a double-film hydrostatic bearing with a membrane double restrictor (*MDR*) which controls the oil flow through the bearing. *I* moving part. II_1 and II_2 fixed parts of the bearing. p_s supply pressure. With increasing load *F* the pressure p_{r1} rises and the pressure p_{r2} drops. This causes the membrane of the M.D.R. to flex upwards, admitting more oil into the lower bearing recess.

with a bearing. When a force F is applied to the moving part I, a pressure difference $p_{r2} - p_{r1}$ must be created in the two bearing recesses in order to balance this force. The pressure difference $p_{r2} - p_{r1}$ also acts on the membrane, causing it to flex towards the side where the pressure is lower. Then the gap height x_2 between the membrane and the housing, at the side where the recess pressure is higher, increases. As the flow through the gap is laminar, the effect of the increased gap height is greater than that of the reduced pressure difference $p_s - p_{r2}$ (equation 1) and there is therefore an increase in the flow to the recess with the higher pressure.

An important feature of the membrane double restrictor is that one of the parameters is the supply pressure p_s; this enables the stiffness in an assembled bearing to be modified — from a positive value, through infinity, to a negative value (*fig. 9*). This facility enables elastic deformations in other parts of the bearing to be compensated.

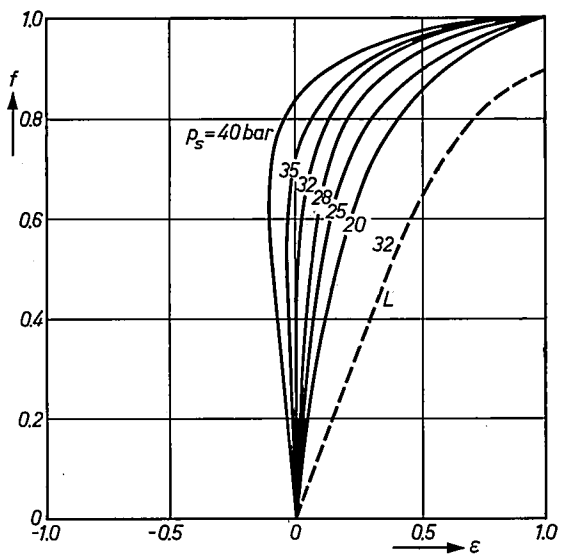


Fig. 9. The load of a double-film hydrostatic bearing with M.D.R. control as a function of the corresponding displacement of the moving part supported in the bearing, for different supply pressures p_s. The curves relate to normalized quantities; the vertical axis shows the load divided by the load-carrying capacity of the bearing ($f = F/Ap_s$) and the horizontal axis shows the variation of the bearing gap height divided by the gap height under no load ($\varepsilon = \Delta h/h_0$). The stiffness of the bearing is equal to the slope of the curve. The M.D.R. is dimensioned in such a way that at $p_s = 32$ bars the bearing stiffness is infinite for moderate loads. At supply pressures greater than 32 bars the bearing stiffness is actually negative over part of its range. For comparison the dashed curve L represents the case in which the flow restriction is a laminar-flow resistance.

Since the bearing and membrane double restrictor form a feedback system, there is the possibility of instability. In such a system this would take the form of spontaneous oscillation at a particular frequency. Even when the system is stable, it may be so close to instability that a transient oscillation occurs after every disturbance. Both the oscillation effects and the existence of dominant resonant frequencies can be avoided by choosing the bearing dimensions in such a way that there is sufficient damping in the system. This can be seen from *fig. 10* where the amplitude and phase angle of the deflection of the bearing under a sinusoidally

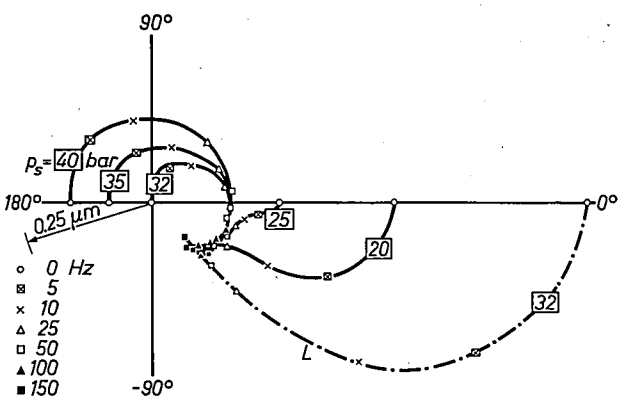


Fig. 10. Polar diagram of the amplitude and phase of the deflection of a double-film hydrostatic bearing with M.D.R. control when subjected to a sinusoidally varying load. Each curve connects the measured values for different frequencies at a given supply pressure p_s in the bearing. At $p_s = 32$ bars the bearing has infinite static stiffness. The shape of the curve shows that there are no marked resonances at particular frequencies, so that the system is dynamically stable. At the highest frequencies the supply pressure hardly affects the curve, the damping in the bearing then being dominant.

The chain-dotted line L represents the dynamic behaviour of the double-film hydrostatic bearing when a laminar-flow resistance is used as the flow restriction instead of the M.D.R.

varying load are shown for different frequencies and different supply-pressures. No marked resonance then occurs at any frequency [9].

The membrane double restrictor not only gives a very high bearing stiffness; it also has the important advantage that it gives sufficient stiffness with a large gap height, low supply pressure and small bearing dimensions. These are just the very conditions that equation (4) shows to be undesirable; however, they are often desirable for other reasons [10].

[8] 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, Pergamon Press, Oxford 1967.

[9] J. G. C. de Gast, Dynamic behaviour of a double-film hydrostatic bearing with variable flow restrictor, to be published in Trans. ASME, J. Lubr. Technol.

[10] Because of this, hydrostatic bearings can also be used for fast rotating shafts (6000 rev/min) with low oil consumption and friction losses. Stiffness, load-carrying capacity and damping are better in this case than with an air bearing; see H. L. Wunsch, Air bearing applications to machine tools and measuring instruments, Trans. ASME, J. Lubr. Technol. 90F, 680-686, 1968 (No. 4).

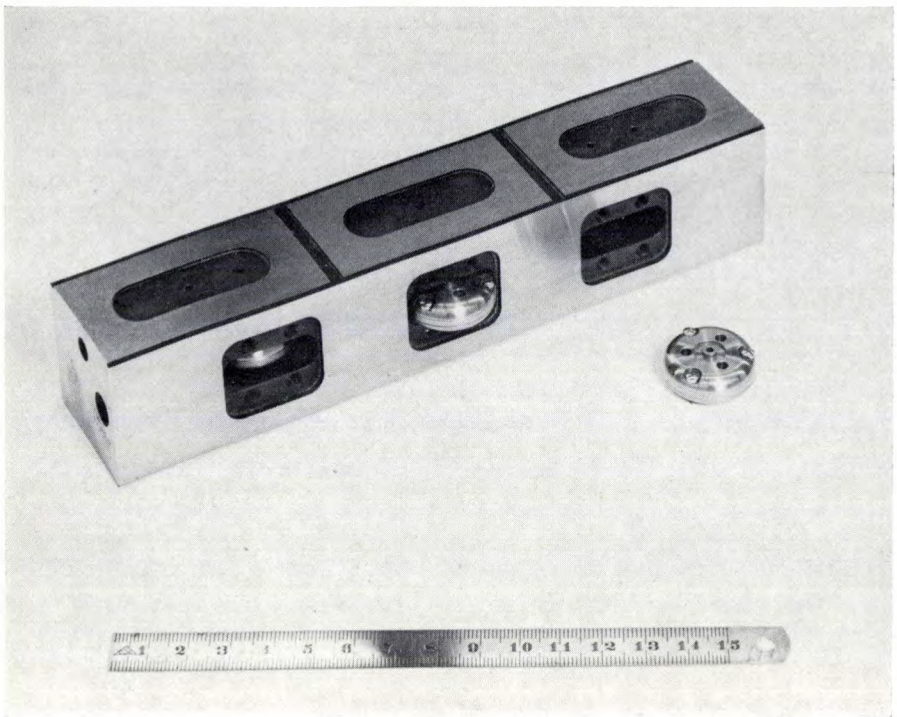


Fig. 11. A bearing block of the carriage with three pairs of hydrostatic-bearing recesses. Each pair has an M.D.R., which is located in a groove of the bearing block.

Fig. 11 shows how the membrane double restrictor is incorporated in a slot cut into one of the bearing blocks. Each pair of bearing recesses has its own membrane double restrictor.

Carriage with slideway

When double-film hydrostatic bearings are used, there are always large preloading forces present (these may reach 10^4 N in the carriage bearings, for example), which can cause severe structural deformation if the forces give rise to bending moments. In the design of the slide system, shown in cross-section in *fig. 12*, the hydrostatic forces are applied in such a way that both the carriage and the slideway are subjected only to compressive, tensile and shearing forces that together form a closed force diagram in the plane of the drawing [11].

The carriage has three identical bearing blocks (as illustrated in fig. 11), which are mounted on the carriage frame. Since the blocks are supported in bearings at both sides, each block is hydraulically balanced so that the carriage frame is relieved of hydrostatic preloading forces. The slideway shown in fig. 12 was machined from one piece and all the guide faces were precision-ground.

The slideway can also be built up from rectangular bars (*fig. 13*). This construction is less rigid than that of fig. 12. The deformations caused by the hydrostatic forces acting on a slideway built up in this way have been measured and are shown in *fig. 14*. An important advantage, however, is that the bars can be

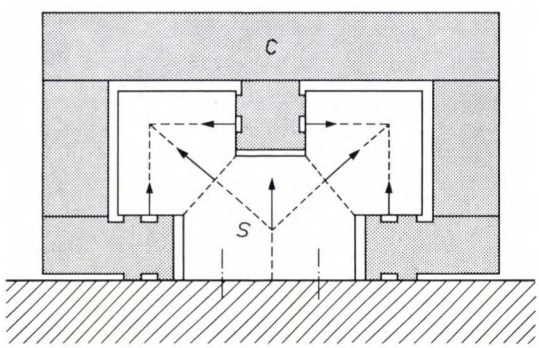


Fig. 12. Cross-section of slideway *S* and carriage *C*. The slideway is made in one piece and constructed in such a way that the various forces acting on it result in a tensile force and thus do not deflect it from a straight line.

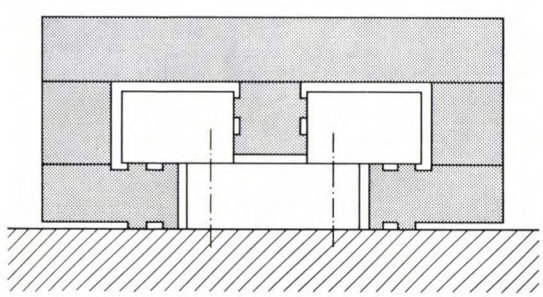


Fig. 13. Cross-section of a slideway not made from one piece but built up from three rectangular bars. With this construction small deviations can be corrected by flexing the bars before they are clamped in position.

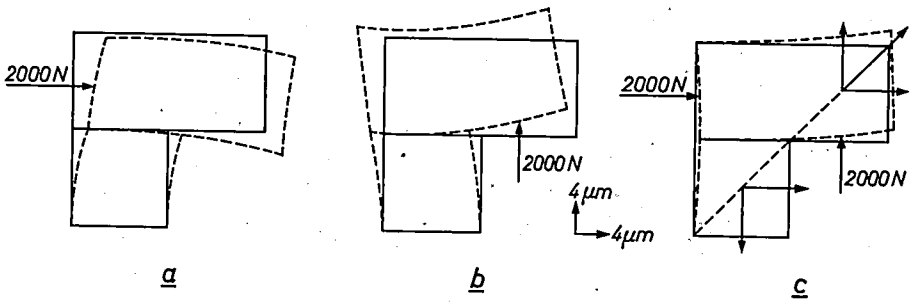


Fig. 14. Measured deformation of a slideway of the type shown in fig. 13, *a*) when only a horizontal force is operative, *b*) when only a vertical force is operative, *c*) when both are operative together. In case (*c*) there is only a small residual deformation.

assembled in such a way that deviations over a length greater than half the bearing-block length can be corrected by flexing and clamping. Slideways of this type were used for the guides in the step-and-repeat camera recently described in this journal [12]; for a 5-cm displacement of the carriage the deflection from the straight is less than 0.03 μm and the angular rotation of the carriage is less than 0.1 seconds.

The linear motor

The linear motor used is a combination of piston and cylinder (*fig. 15*) with a piston to piston-rod surface ratio of 2:1. When the piston is stationary and the supply pressure p_s is applied to the annular space between the cylinder and the piston rod, the pressure in the oil-filled space to the left of the piston is p_1. Taking the equilibrium of forces at the piston, then for the given surface ratio and zero external forces, p_1 is given by:

$$p_1 = \tfrac{1}{2}p_s. \qquad \ldots\ldots\ldots \quad (5)$$

If the space to the left of the piston is now connected via a resistance to a supply of oil at a pressure p_s, the piston is given a velocity v to the right; if the space to the left of the piston is connected via R to oil at atmospheric pressure, then the piston is given an equal velocity v to the left. In both cases the oil flow rate $\Phi = (\pi/4)D^2v$ (D being the inside diameter of the cylinder) is given by:

$$\Phi = \frac{p_s - p_s/2}{R}.$$

The velocity is thus given by:

$$v = \frac{2p_s}{\pi D^2 R}. \qquad \ldots\ldots\ldots \quad (6)$$

For a good surface finish it is necessary that v should remain constant to within 0.1 %. It can immediately be seen from (6) that the essential condition for this is that the supply pressure p_s should be constant to within 0.1 %. The load on the motor should also be constant; this has been neglected in the derivation of (6), but can be taken into account by a term to be subtracted from p_s. Since it is extremely difficult to keep the supply pressure and particularly the load as constant as this, it was decided to ensure the uniformity of v directly by keeping the flow rate Φ at an accurately constant value. This was done by replacing R by a flow-control valve (whose operation is described in the section on control valves).

The uniformity of the velocity is also adversely affected by the pressure in the cylinder which can cause the inside diameter D of the cylinder to have a slightly different value at different places. In order to make the deformation of the cylinder wall as uniform as possible, it is not rigidly attached to the cylinder head but mounted freely; this can be seen in the detailed cross-section of the linear motor shown in *fig. 16*.

Another difficulty that can give rise to non-unifor-

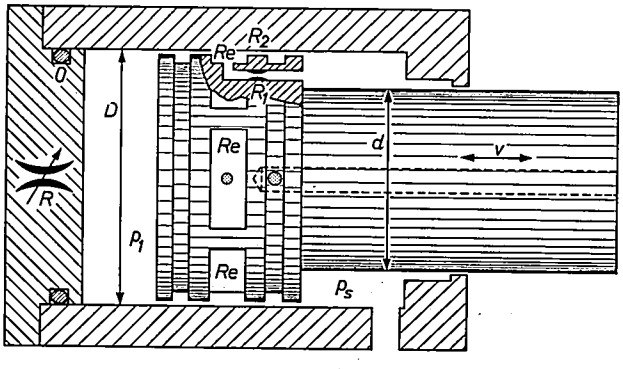


Fig. 15. Simplified cross-section of a linear motor. The space to the left of the piston communicates through a laminar-flow resistance R with the supply pressure p_s (piston moves to the right) or with atmospheric pressure (piston moves to the left). The piston surface $\frac{1}{4}\pi(D^2 - d^2)$, on which p_s acts, is half the size of the surface $\frac{1}{4}\pi D^2$, on which p_1 acts, so that when the hydraulic forces are balanced, $p_1 = \frac{1}{2}p_s$. The hydrostatic-bearing action of the piston in the cylinder is obtained by means of four recesses Re with restrictors R_1 incorporated in the piston and the flow resistance R_2 between piston and cylinder wall.

[11] J. G. C. de Gast, Berekening en constructie van hydrostatische lagers, Polytechn. T. Werktuigbouw **23**, 141-150, 187-196, 1968 (Nos. 4, 5).

[12] F. T. Klostermann, A step-and-repeat camera for making photomasks for integrated circuits, Philips tech. Rev. **30**, 57-70, 1969 (No. 3).

mity in the piston velocity is the presence of leakage between the piston and the cylinder wall. However, by including a control valve (*7* in fig. 16) in the piston the oil pressure in the annular grooves and bearing recesses in the piston surface can be adjusted in such a way that hydrostatic bearing action and a perfect seal are obtained at the same time.

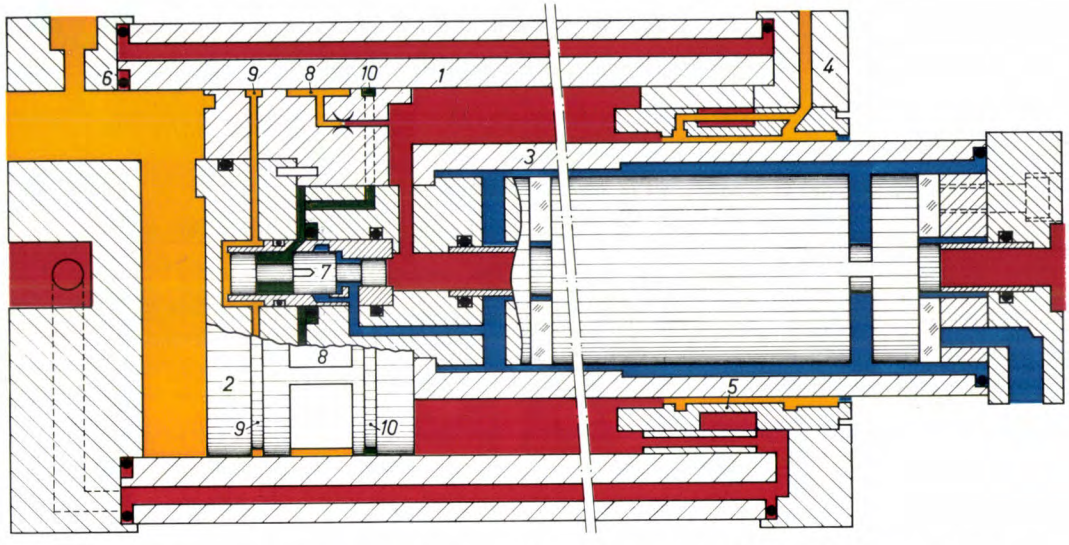


Fig. 16. Complete cross-sectional drawing of the linear motor in the lathe. The colours correspond to various oil pressures: *red* = supply pressure p_s; *yellow* = $\frac{1}{2}p_s$; *green* < $\frac{1}{2}p_s$; *blue* = atmospheric pressure.
1 cylinder. *2* piston. *3* piston rod. *4* cylinder head. *5* thin-walled extension in which the piston rod can be clamped to hold it accurately stationary. *6* O-ring seal on the head surface of the cylinder jacket. *7* control valve. *8* bearing recesses. *9* annular groove under average bearing-recess pressure. *10* outlet slot. The control valve *7* incorporated in the piston *2* keeps the pressure in the bearing recesses *8* of the piston equal to that on the left of the piston, i.e. equal to $\frac{1}{2}p_s$.

The permissible *deviations in the cylinder diameter* are very small. The effect on the velocity of a deviation ΔD in the diameter of the cylinder may be deduced from (6) and, neglecting terms of higher order, may be written as:

$$\frac{\Delta v}{v_0} = -\frac{2\Delta D}{D_0},$$

where v_0 and D_0 are the normal values of v and D. If Δv is required to be less than $10^{-3}v_0$, then the error due to a deviation of diameter should be about an order of magnitude smaller, since there are other sources of error. We then have:

$$2\Delta D < 10^{-4}\, D_0 .$$

For a cylinder diameter of 80 mm the deviations ΔD must therefore be less than 4 μm. Now ΔD has reached 10 μm for a pressure difference of 40 bars and a wall thickness of 6 mm. If the cylinder were constructed as shown in fig. 15, this deviation would occur at the centre of the cylinder but not at the two ends. This is because one end is reinforced by the cylinder head and the O-ring seals off the other end from the pressure. In the design shown in fig. 16 these effects are avoided by mounting the cylinder freely and not rigidly connected to the cylinder head *4* and by applying the seal *6* to the end face of the cylinder tube.

Deformation of the cylinder is also caused by the pressure difference across the piston. The length L subject to this slightly tapering deformation depends upon the cylinder diameter D and the wall thickness t and is given by:

$$L \approx 4.5\sqrt{Dt}.$$

The pressure difference across the piston is now concentrated in a narrow zone of the very thin oil film between the piston and the cylinder wall. We shall see that this zone is located near the high-pressure end (in fig. 16 the right-hand end) of the piston. The operation of the linear motor, on the other hand, is based on the displacement of oil by the left-hand end of the piston. The cylinder deformation here due to the pressure difference should be imperceptible, which implies that the piston length should at least be equal to the deformation length L.

The *sealing of the piston*, which has the pressure difference $p_s - p_1$ across it, calls for special measures, since the running of the piston must not be unfavourably affected by the seal. It is of course essential that the piston and piston rod should be accurately centred in their guides for proper running of the piston, otherwise, owing to the high surrounding pressure, these moving parts could seize up against the cylinder wall or cylinder head as soon as they came into contact with them. In hydraulic engineering this effect is known as "hydro-lock" [13]. The necessary centring is obtained by supporting both the piston and the piston rod in hydrostatic bearings. The four bearing recesses (*Re* in fig. 15) provided for this in the piston circumference are connected via four identical resistances R_1 to the annular high-pressure space between the piston rod and the cylinder. The oil from these bearing recesses does not flow into the low-pressure space on the left of the piston [14], but into an annular groove situated between the recesses and the high-pressure space and open to atmospheric pressure via the hollow piston rod.

The piston seal thus becomes a seal between the space with the pressure p_1 and the bearing-recess pressure p_r. We take $p_r = \frac{1}{2}p_s$, since a hydrostatic bearing works best with this ratio be

tween recess pressure p_r and supply pressure p_s; since also $p_1 = \frac{1}{2}p_s$ (see equation 5), the pressure difference to be sealed is small.

With a simple rubber seal, such as an O-ring, this small pressure difference, which is independent of the pressure level, can be adequately sealed without adversely affecting the movement of the piston. What will be affected is the positioning of the piston, because of the hysteresis of the material, while the static friction is also a function of time. An element giving such a complete seal is not necessary, however, provided p_1 and p_r do not differ by more than 0.01 bar, at which level the leakage of oil through the gap is negligible. This has been achieved by means of a pressure-control valve incorporated in the piston (see fig. 16) which builds up a sufficiently high pressure in the annular outlet groove *10* (*green*) to keep the average recess pressure (present in the annular groove *9* on the left of the recesses in fig. 16) equal to the pressure $\frac{1}{2}p_s$ (*yellow*).

Keeping the piston stationary for a long period to an accuracy of better than 0.1 μm requires special measures. This is a feature of the lathe that is vital, and has been achieved by hydraulically clamping the piston rod. The clamping action is provided by a thin-walled cylindrical extension on the inside of the cylinder head (*5* in fig. 16). The pressure is always high on the outer side of this cylindrical extension and a decrease in pressure between the extension and the piston rod causes an elastic deformation and hence the clamping action. The low pressure is adjusted by bringing the two adjacent annular grooves up to atmospheric pressure by means of a magnetically operated valve. The pressure-control valve in the piston, which keeps the pressure in the bearing recesses of the piston at $\frac{1}{2}p_s$, ensures that in the clamped state the pressure p_1 in the space behind the piston does not rise or fall as a result of slight oil leaks.

It can also be seen from fig. 16 that the cylinder has a double wall: this construction has been used to give protection to the inner cylinder wall and to permit oil of controlled temperature to be circulated around the oil space behind the piston. The piston rod is hollow and is provided with two backlash-free universal joint couplings. This enables the cylinder to be fixed to the slideway and the drive shaft to the carriage; a misalignment of $\pm 2°$ is permissible. The piston also has an internal high-pressure passage, instead of flexible tubing, for supplying oil to the carriage, etc. Sufficient oil circulates through the linear motor to bring the complete unit to the controlled temperature of the oil under all operating conditions.

Headstock spindle and motor

The accuracy of rotation of the headstock spindle when hydrostatic bearings are used is determined by the circularity of the spindle itself, and can even be better than this, the shape of the bearing bush being of secondary importance. For example, an experimental shaft machined on a headstock spindle with an out-of-roundness of 0.7 μm was found to have an out-of-roundness of 0.1 μm. It was decided to give the headstock spindle a cylindrical shape without protrusions, since this shape can be made accurately circular to within 0.1 μm by lapping.

The distance d between the two journal bearings (see *fig. 17*) was chosen so as to minimize the tilt of the

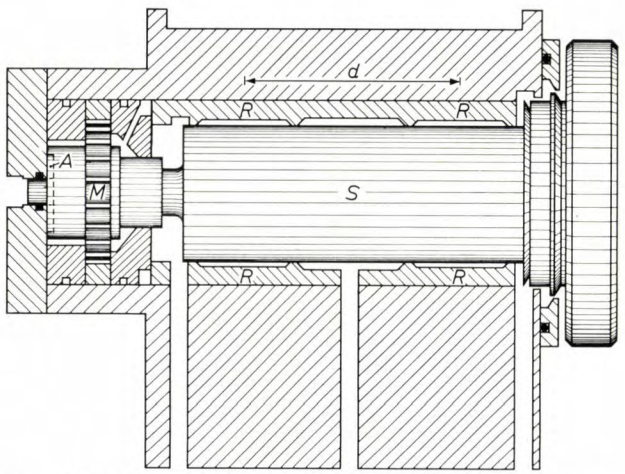


Fig. 17. Cross-sectional drawing of headstock spindle *S* and vane motor *M*. *A* thrust bearing. *R* journal bearings.

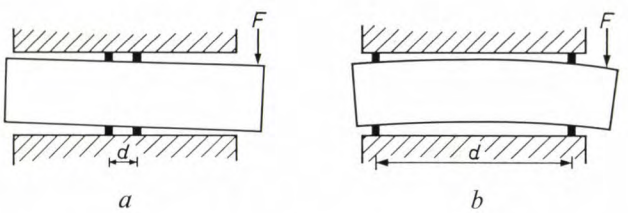


Fig. 18. A force *F* acting on the projecting part of the headstock spindle presses the spindle sideways. A relatively large deviation may be caused by tilting of the whole spindle if the distance *d* between the bearings is small (*a*), and by flexing of the spindle if the distance between the bearings is large (*b*). Between these there is an optimum distance *d* which gives the least deviation.

headstock spindle in the front bearing due to a radial force acting from outside the bearings (see *fig. 18*). This tilt is to be avoided because it changes the direction of the axis of rotation of the workpiece clamped in the chuck. The optimum distance d depends on the modulus of elasticity E, on the linear moment of inertia I of the cross-sectional area of the headstock spindle and on the bearing stiffness K; the relation is:

$$d = \sqrt[3]{c\frac{EI}{K}},$$

where the constant c is about 16 if all quantities are in SI units [15].

The thrust bearing (*A* in fig. 17) is fitted at the end of the headstock spindle *S* and is supported by a flat plate bolted to the headstock. This is a hydrostatic single-film

[13] J. Boyd, The influence of fluid forces on the sticking and the lateral vibration of pistons, Trans. ASME, J. appl. Mech. **31**, 397-401, 1964.

[14] See A. A. Raimondi and J. Boyd, Fluid centering of pistons, Trans. ASME, J. appl. Mech. **31**, 390-396, 1964.

[15] If the tilt of the main spindle is not minimized but the radial deviation of the chuck instead, another expression must be used (see reference [11]).

bearing pre-loaded by the supply pressure. The design is such that there is no interaction between the journal and the thrust bearings, and the front journal bearing lies as close as practicable to the chuck. This is the best arrangement for accurate running and for the stiffness of the headstock spindle. One disadvantage is that the thrust bearing is a long way from the chuck, which means that the axial position of the work clamped in the chuck is dependent on the temperature of the headstock spindle. Fluctuations in this temperature are very small, however, on account of the method of regulating the speed of the headstock motor.

We have seen on page 120 that this is done by varying the flow through a parallel branch and at the same time keeping the sum of the flows through motor and parallel branch constant. Since the pressure is also kept constant, this constant total flow supplies to the headstock a hydraulic power (pressure × volume flow) which is independent of the motor speed. This hydraulic power is dissipated partly in the hydrostatic bearing of the motor and the spindle by friction and partly by throttling in the flow-control valve in the parallel branch. In addition, energy is used up in machining, but for a light finishing operation (e.g. a cut of $20 \times 12\ \mu m$) the power required is negligible. The heat generated in the hydrostatic bearings increases with the motor speed; since however the total dissipated power is independent of speed, the heat developed in the flow-control valve decreases equally. The oil flows from the two parallel branches are now mixed between the two headstock bearings, and the mixed flow, which thus supplies a quantity of heat independent of the speed, circulates around the headstock spindle, which therefore remains at the same temperature at different speeds.

The headstock motor is of the vane type. Since a uniform angular velocity can be obtained with this type of motor, it can be attached directly to the headstock spindle without the need for a coupling to smooth out deviations from uniformity. Moreover, no speed-selection mechanism is necessary for the transmission since the speed of the motor can be continuously regulated within wide limits, between 300 and 1200 rev/min in our lathe. In addition it is hydraulically balanced radially and axially and is small.

A vane-type motor consists of a rotor with a number of vanes, and a stator (*fig. 19*). The stator contains several "chambers", two in our case, with a stroke h; the chamber walls are cylindrical and concentric with the inner bore over an arc of Θ radians. The inner bore is also cylindrical over an arc of Θ radians, and there is a smooth transition from chamber to inner bore. The vanes are kept in contact with the transition part of the wall by a spring. In the chambers the vanes are subjected on the one side to the supply pressure p_s and on the other to a lower pressure p_m. The pressure difference acting on the projecting part of the vane causes a force to act on the vane. This force gives rise to a couple that drives the rotor. Oil is supplied and removed on the transition curves and there is always at least one vane between feed and outlet.

Given an angular displacement φ the vane displaces a volume V of oil given by:

$$V = hd\,\frac{D+h}{2}\,\varphi .$$

Owing to the symmetrical design of the motor, the same volume is displaced by the vane in the chamber at the opposite side, so that the volume displaced for each radian that the motor rotates is:

$$V_r = hd\,(D+h),$$

which is referred to as the swept volume per radian. Since h is constant over the arc Θ of the chamber, the volume displaced per radian is independent of the position of the rotor, so that this type of motor has in theory a uniform angular velocity ω. This is determined by the oil feed Φ and the oil leakage Φ_1:

$$\omega = \frac{\Phi - \Phi_1}{V_r}.$$

In spite of slight deviations the oil leakage Φ_1 may readily be defined in a wide operating range as a linear function of the pressure difference $p_s - p_m$. The change in Φ_1 can therefore be compensated in the precision lathe by using a membrane load compensator like the one developed some time ago at these Laboratories for a linear servomotor [16].

As soon as a vane has reached the end of a chamber (see fig. 19) the next vane has to take over the sealing action. Consequently

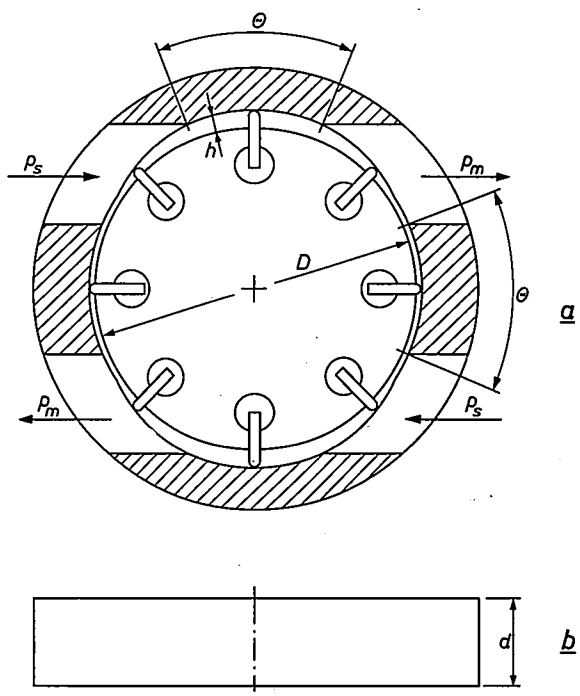


Fig. 19. *a*) Cross-section and *b*) plan view of a vane motor. p_s pressure at input side and p_m pressure at output side of the motor. h height of stroke.

the number of vanes must be at least $2\pi/\Theta$. With this number of vanes, however, there is no overlapping at all at the change-over and there is an irregularity at each change-over. It is therefore better to use more vanes so that there is some overlap. Even then, some change-over irregularity remains: since it is impossible to avoid a small amount of play in the grooves for the vanes, each vane "flips" across as it takes up the pressure difference; moreover, the leakage is halved during each short interval when the two vanes are in series. In addition, successive vanes take up pressure simultaneously in the two chambers, because the rotor must remain radially balanced. The resultant irregularities are visible on a trial workpiece after a plane perpendicular to the axis has been machined. These irregularities can be removed by using twice the number of vanes, so that the seal is always effected by two vanes in series. Even then there is still some unevenness in the leakage flow since there is a short overlap period in which three vanes are in series. This difficulty can be overcome by cutting 20 μm deep "relief" grooves in the overlap part of the chamber, to permit sufficient additional leakage for the flow during the overlap period to be equal to the leakage rate when there are two vanes in series.

The motor used in the lathe has two chambers with 18 vanes and an overlap of 10°. The diameter of the rotor is 60 mm, its width is 7 mm and the stroke is 2 mm. At the maximum speed of 1200 rev/min a power of 250 W is available at the chuck. The power is kept as low as this because the present lathe is used only for finishing, and even this small amount of power is seldom required. However, if necessary, a higher power could be used with a headstock of these dimensions.

The control valves

In the foregoing we have several times referred to control valves whose purpose was to keep a pressure or a flow to a prescribed value. Flow control can always be reduced to the control of a pressure difference across a fixed flow resistance, so that in fact all control requirements amount to controlling a pressure difference. In the precision lathe this must be done with an accuracy as good as 0.1%, whereas the best that conventional control valves can give is about 1 to 2%. It was therefore necessary to design new control valves for the precision lathe, paying special attention to all the unwanted forces that arise as a result of fluid flows or friction in the valves. The required accuracy was achieved by compensating the forces associated with the fluid flow and by making further use of hydrostatic bearing action.

A hydraulic control valve operates by varying the opening of a flow orifice under the action of varying pressure levels. In the precision lathe described here each control valve consists of a moving piston which controls the opening of the outlet ports in a cylinder.

Two cases are to be distinguished for a constant pressure. In the first case a pressure-control valve keeps the pressure constant at the input to the valve; the oil flow through the valve is then a bypass flow. This case is illustrated in *fig. 20*; the regulated pressure p_c at one end of the piston balances the spring force F acting on the piston. A schematic diagram representing this case is also shown in fig. 20. Φ is the main flow, Φ_2 the bypass flow. The dashed line indicates

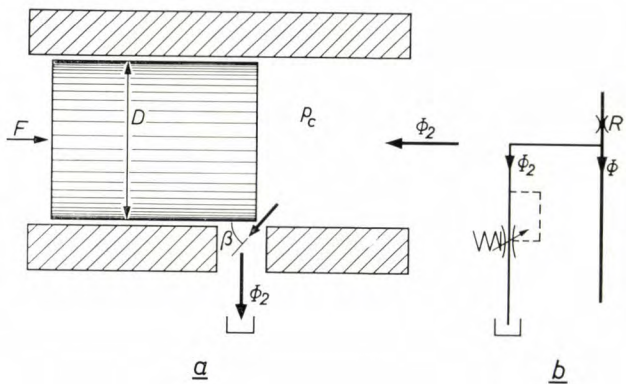


Fig. 20. *a*) Simplified cross-section of a control valve that keeps the pressure at the *input* to the valve to a constant value p_c. The force due to p_c acting on the piston balances the force F, usually derived from a spring. The oil flow Φ_2 through the outlet port is a bypass flow, required for the control action. The outflow direction makes an angle β with the axis of the valve.
b) Diagram of the hydraulic circuit containing the control valve. The dashed line indicates that the pressure at the input to the valve affects the magnitude of the flow orifice. Φ main flow. R internal resistance of the pressure source.

that the pressure at the *input* to the valve affects the magnitude of the flow orifice. The second case is that of the reduction of a higher pressure to a constant value p_c, independently of the flow. Pressure-control valves for this purpose (reducing valves) carry the main flow, the pressure being controlled on the *output* side of the valve (see *fig. 21*).

Compensation of flow force

In both types of control valve there is an equilibrium of forces which is given by the equation:

$$F = \tfrac{1}{4}\pi D^2 p_c,$$

where D is the diameter of the piston. The force F must be independent of the position of the piston and of the flow rate through the valve to ensure that the controlled pressure p_c remains constant. The force F that actually appears in the valve, however, contains in addition to the component derived from a spring a component

[16] T. J. Viersma, Investigations into the accuracy of hydraulic servomotors, thesis, Delft 1961. The deviations from the linear relationship between volume flow and pressure which applies for laminar flow arise because the shape of the leakage gap and the temperature of the oil, and hence its viscosity, are pressure-dependent.

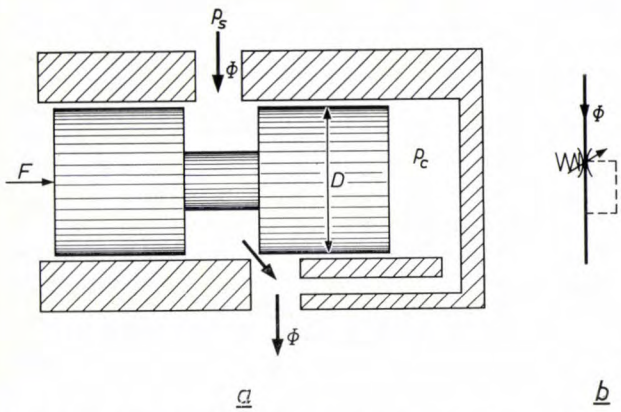


Fig. 21. *a*) Simplified cross-section of a control valve that keeps the pressure on the *output* side of the valve at a constant value p_c. The force due to p_c acting on the piston balances the force F, usually derived from a spring. The oil supplied under pressure p_s exerts no force on the piston, the main flow Φ goes through the valve. *b*) Diagram of the control valve. The dashed line indicates that the pressure at the *output* of the valve affects the magnitude of the flow orifice.

that varies with the valve orifice, the flow force F_i. This force, which may be regarded as a reaction force applied to the piston by the outflowing fluid, can amount to 20% of the applied spring force F_s, and thus seriously interferes with the operation of the control valve. We have now built a control valve in which the flow force is compensated in the following way.

The flow force is proportional to the fluid mass flowing out per unit time, $\varrho\Phi$ (ϱ is the density of the fluid), and to the axial component of the outflow velocity v. The relation is:

$$F_i = \varrho\Phi v \cos\beta, \quad . \quad . \quad . \quad . \quad . \quad . \quad (7)$$

where β is the angle between the outflow direction and the axis of piston and cylinder (see fig. 20). It is possible to apply a compensating counter-force, proportional to Φ (see below), but there is the complication that F_i is also proportional to v. It turns out, however, that the effect of this proportionality is opposite to that of another undesirable effect that we have not yet mentioned. This is the variation of the spring force F_s with the position of the piston, and by suitably choosing the spring stiffness the two effects can be made to approximately cancel out in the operating region of the valve.

The variation of the spring force F_s with a displacement x of the piston is given by:

$$F_s = F_{s0} + K_s x, \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (8)$$

where K_s is the spring stiffness and F_{s0} the spring force acting at the point chosen as $x = 0$. If we take $x = 0$ for the position at which the piston just closes the outlet ports, then x can be expressed in terms of the outflow velocity v and the outlet flow Φ.

Let xb be the area of the outflow opening, b being the width of the outlet ports; then the flow rate Φ is given by:

$$\Phi = \alpha x b v. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (9)$$

Here α is a flow coefficient; in a wide range α is approximately equal to 0.7 [17].

Equations (7), (8) and (9) show that the total force F acting on the piston is given by:

$$F = F_{s0} + \Phi\left(\frac{K_s}{\alpha b v} + \varrho v \cos\beta\right).$$

We see that the factor between brackets contains one term proportional to v and one term inversely proportional to v. An expression of this kind is fairly constant in a particular range of the (independent) variable v. If the spring stiffness K_s is given a suitable value, this range can be situated about the operating point of the control valve. If the outflow velocity is v_0 at this operating point, then with proper dimensioning the force F is given by:

$$F = F_{s0} + \Phi\cos\beta\left(\frac{v_0^2}{v} + v\right). \quad . \quad . \quad . \quad . \quad (10)$$

The variation of the factor in brackets is shown in *fig. 22*. It can be seen from this that the unwanted force proportional to Φ has been made approximately independent of the outflow velocity in the vicinity of the operating point.

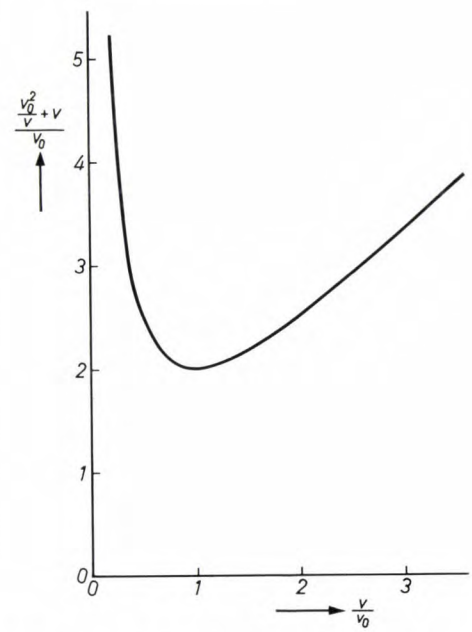


Fig. 22. Variation of the expression $(v_0^2/v + v)$ with v. The minimum is obtained when $v = v_0$; in a range about this point the value of this expression is virtually constant.

To compensate the flow force the oil flow Φ, after leaving the valve, is passed through a laminar resistance R (see *fig. 23*, a more detailed cross-sectional drawing of the pressure-control valve, whose principle is illustrated in fig. 20). The pressure drop across a laminar resistance is proportional to the quantity of oil flowing through it (see equation (2) which relates to laminar flow in a hydrostatic bearing) and the pressure p

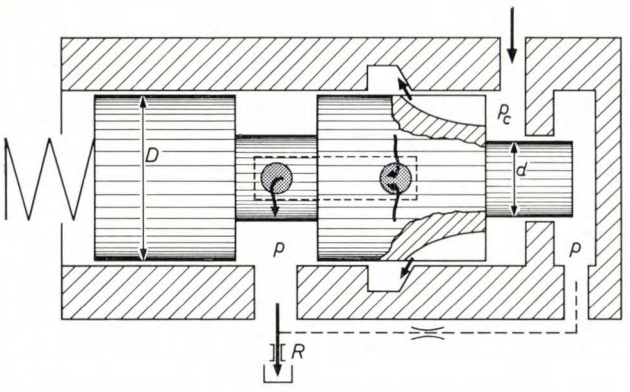


Fig. 23. A more detailed cross-sectional drawing of the control valve in fig. 20. A fixed laminar resistance R is included in the bypass flow. The resultant pressure p acts on an additional piston surface, thus providing compensation for the flow force.

built up by the resistance R is thus proportional to Φ. This compensating pressure p acts on an extra piston face with reduced diameter d and exerts a force on the piston which is opposite to the flow force and cancels it for every value of Φ if R is correctly chosen [18]. The only reservation here relates to the effect of temperature on the viscosity of the oil and hence on the magnitude of the resistance R; this may sometimes make it impossible to achieve perfect compensation.

Design of the valves

The speed at which the oil leaves the outlet ports is fairly high (e.g. 65 m/s at a pressure difference of 20 bars). To allow the fast-flowing oil to settle down again, an annular groove is cut in the cylinder wall (see fig. 23). If this groove were rectangular in cross-section, the oil flow would create an undesired pattern in it [19], and the inlet angle β would continuously fluctuate. The wall along which the oil enters the groove is therefore placed at an angle of 60° as indicated in fig. 23; the oil flow thus follows the wall [20], and the inlet angle is stabilized. To guide the oil flow along the piston there are two (or four) slots. The oil in the annular groove flows away through a wide opening in the piston.

The pistons of the control valves may seize up because of hydro-lock [21]. To prevent this the pressure difference that always exists across the pistons is used to provide them with a hydrostatic bearings action. An additional effect is that the accuracy of 1 to 2% usually obtainable with control valves has been improved to 0.1%; a pressure-control valve responds reproducibly to a variation as small as 10^{-3} bar at a pressure level of 10 bars.

Fig. 24 gives an idea of the dimensions of the control valves, showing cylinder and piston for two versions with piston diameters of 12 and 6 mm.

Servo-operated pressure-control valve

The pressure-control valve shown in fig. 23 is used for keeping the supply pressure of the lathe constant. Owing to the system of speed control employed, the oil flow through the lathe is always about the same, and consequently the bypass flow through the valve also remains relatively constant. As a result this valve is able to keep the supply pressure accurately constant.

Marked differences in the flow will occur, however, in the control valve whose function is to keep the tailstock pressed against the workpiece with the same force

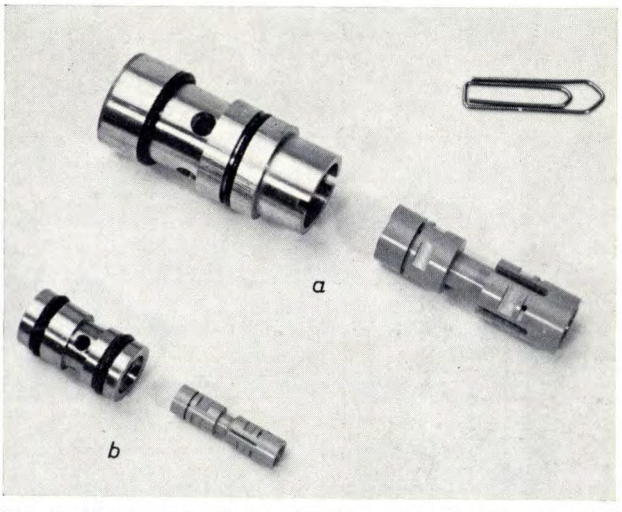


Fig. 24. Piston and cylinder of two control valves incorporated in the precision lathe. *a*) Large version, diameter of piston 12 mm; *b*) small version, diameter 6 mm.

at all times however quickly it is moving (valve *12* in fig. 4). In this case the method described for compensating the flow forces in the control valve is not sufficiently accurate, and to obtain higher accuracy it is necessary to use a pressure-control valve in which an extra controlling force, derived from a pilot valve, is applied in addition to the forces described above.

The circuit of a well-known valve of this type is shown in *fig. 25a*. The pressure-control valve C has to

[17] L. Gross, Bestimmung der Durchflussbeiwerte verschiedener Drosselelemente, besonders bei kleinen Druckunterschieden, Ölhydraulik und Pneumatik **12**, 3-8, 1968 (No. 1).

[18] Subsequent adjustment is possible in this method, unlike the one described by J. F. Blackburn, G. Reethof and J. L. Shearer, Fluid power control, M.I.T. Technology Press, New York 1960: in this method the part of the cylinder that receives the outflowing oil and also the piston surface opposite are specially shaped.

[19] See page 362 of the book mentioned in [18].

[20] This effect is known in fluid logic as the Coanda effect; see P. E. Russ Henke, Fluidics control: yesterday, today and tomorrow, Electromechanical Design, May 1968, page 12.

[21] This effect has been mentioned earlier in the description of the linear motor on p. 125.

keep the pressure p_c to a preset value for widely different flows through C. Since the position of the piston is different for different flows, a spring is unable to deliver a constant reference force. Instead of this a force derived from a constant pressure p_1 is used as the reference force. This control pressure p_1 is kept constant by the pilot valve P. Since the orifice plate A gives a virtually constant flow through P, p_1 is held accurately constant. This is referred to as a pilot-operated control valve.

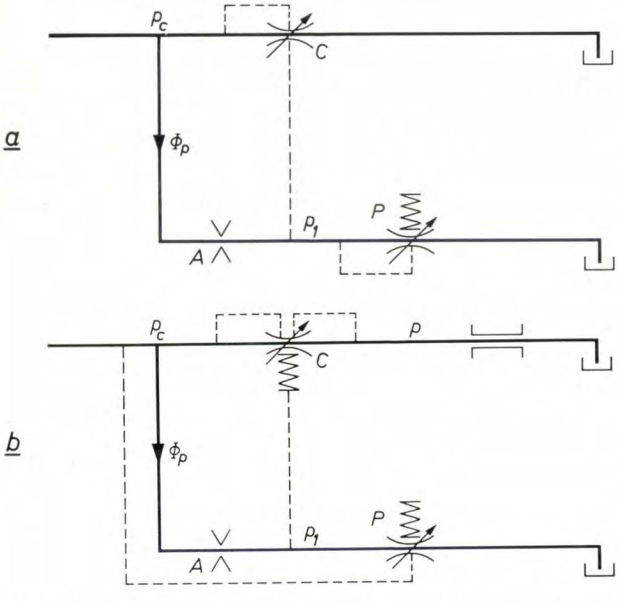


Fig. 25. *a*) Orthodox version of pilot-operated pressure-control valve. Since the flow rate through the control valve C may vary and therefore with it the position of the piston in the valve, a spring cannot supply a constant reference force. This is now supplied by the control pressure p_1 kept at a constant value in the control valve P. Because of the action of the orifice plate A the flow through P changes very little and can thus be used for accurate pressure regulation. Flow forces in C are not corrected. *b*) Servo-operated pressure-control valve as used in the precision lathe. The small pilot valve P here reacts to the pressure p_c to be controlled and delivers a control pressure p_1 that exerts a correcting force on the control valve C. Unlike the situation in (*a*), this has the result that the flow forces in C are also corrected.

Since the reference force in this system must remain accurately constant, the varying flow forces in C occurring with varying flow will show up as variations in p_c, as already described. This difficulty is countered here by using the control system shown in fig. 25*b*. Here the pilot valve p responds directly to the pressure p_c to be regulated, thus generating a control pressure p_1 which acts upon the control valve C in such a way that all deviations are corrected, including those arising from varying flow force and spring force. The resultant feedback system is referred to as a servo-operated control valve.

The pilot valve P in fig. 25*b* is shown in cross-section in *fig. 26*. The control pressure p_1 is regulated in the valve but does not itself exert any force on the piston. The same control valves are also fitted in the pistons of the linear motors for the headstock and toolholder

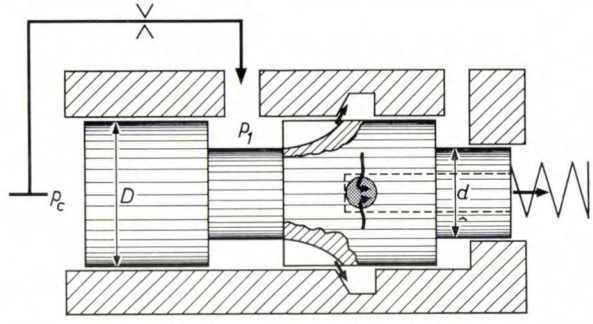


Fig. 26. The pilot valve of the servo-operated pressure-control valve. Deviations in the regulated pressure p_c cause fluctuations in the control pressure p_1, which exert a corrective action on the pressure-control valve. The control pressure p_1 does not itself exert any force on the piston of the pilot valve.

slide systems (see fig. 16). There, however, the reference force is not delivered by a spring but by the supply pressure p_s (*red*) which acts upon a piston surface of half the size and balances the half supply pressure $\frac{1}{2}p_s$ (*yellow*) acting upon the whole piston surface. The valve keeps the pressure in the bearing recesses *8* at half the supply pressure and thus provides a seal for the piston (see page 126).

A force can also be exerted on the control valve through the pressure indicated in *blue* in fig. 16. This scheme can be used for applying a correction to the regulated pressure if required in connection with manufacturing tolerances, and it can also be used to set up a small intentional pressure difference across the sealing gap. This pressure difference will give a very small laminar flow through this gap, thus giving the piston a constant low velocity. This principle has been applied in a more elaborate form in a laminar flow-control valve which is used at Philips Research Laboratories in combination with a linear motor for pulling crystals from a melt by the Czochralski method. This system readily permits piston velocities of 1 μm/s or less [22].

Flow-control valve

This laminar flow-control valve is only suitable for very small flows. From about 0.25 cm³/s a turbulent-flow valve is used, consisting of a pressure-reducing valve with a resistance or orifice plate in series with it, through which the flow is turbulent. Although it would

[22] This method is therefore an alternative to the one described by H. von Weingraber, Vorschubeinrichtung für extrem langsame Geschwindigkeiten, Annals C.I.R.P. **11**, 90-95.

be possible to build a laminar-flow control valve for these higher flows, the use of a turbulent flow has the advantage of reduced temperature dependence. This may be seen from the Bernoulli expression for turbulent flow through an orifice of area A, where the volume flow is given by:

$$\Phi = \alpha A \sqrt{\frac{2(p_1 - p_2)}{\varrho}}.$$

Here the coefficient α is a constant related to the constriction of the flow; it usually has a value of about 0.7 [17]. The factor $p_1 - p_2$ represents the pressure difference across the orifice and ϱ is the density of the fluid; the density varies very little with temperature, unlike the viscosity coefficient η, which plays a role in laminar flow (see equation 2).

All the flow-control valves in the precision lathe — used for controlling the speed of the headstock motor and the speeds of the linear motors — are therefore of the turbulent-flow type. The orifice-plate aperture of area A is incorporated near the valve (*fig. 27*). The pressure difference $p_1 - p_2$ across the orifice plate exerts a pressure on the piston which balances a spring force. Since the flow rate Φ through the valve remains constant, it is not necessary here to compensate for the flow force; however, the spring stiffness must be correctly chosen since the flow force is again proportional to the velocity.

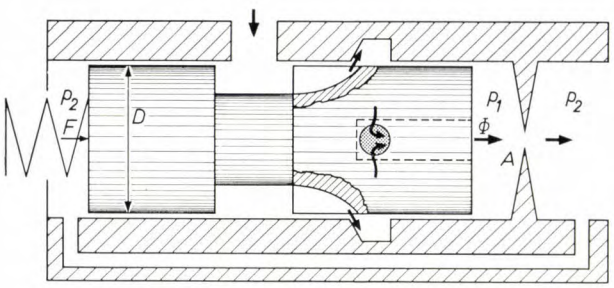


Fig. 27. Flow-control valve. The volume flow Φ produces a pressure difference $p_1 - p_2$ across the orifice plate of aperture A. This difference acts on the piston, which tries to keep it equal to a preset value determined by the spring force F.

G. H. Veldhuizen of the mechanical design group in this Laboratory had an important share in the development of the precision lathe described here. The external design of the lathe is entirely his work.

Summary. A precision lathe has been designed and built at Philips Research Laboratories for use there. All moving parts of this lathe run in hydrostatic bearings. The lathe has three carriages, each driven by a linear motor, consisting of a piston and cylinder. The double-film hydrostatic bearings of the carriages possess infinite stiffness through feedback of the bearing pressure to a variable flow restrictor, referred to as a membrane double restrictor. The main spindle is driven by a hydrostatic vane motor. The lathe can machine work up to 200 mm in length and diameter with a dimensional tolerance of 1 μm; the out-of-roundness of turned work is less than 0.1 μm and the finish of the machined surfaces is of optical quality. The finish does not depend on the ambient temperature, but to achieve a dimensional tolerance of 1 μm it is necessary to keep the ambient temperature constant, purely because of the thermal expansion of the workpiece. The displacements of workpiece and tool are measured with an optical system to within an accuracy of 0.5 μm, and the measurement is displayed on a digital indicator. The article describes in detail the design and operation of the main parts, in particular the slide systems with their hydrostatic bearings, the linear motors, the vane motor and the contol valves used in many parts of the lathe. The accuracy of the valves has been raised to 0.1%.

The MOS tetrode

T. Okumura

In the various designs of MOS transistor that have so far been produced the feedback capacitance is relatively high. This means that they are less suitable for amplifying signals of very high frequency. The article below describes a related circuit element which can be used for frequencies up to 200 MHz and higher. This new device has been developed at the Takatsuki-Osaka laboratory of the Matsushita Electronics Corporation (MEC) a company owned jointly by the Japanese company Matsushita Electronic Industries (MEI) and the Philips Group of Companies, and active in various fields of common interest.

Introduction

The most advanced of the various field-effect transistors known at present is the metal-oxide-semiconductor (MOS) transistor. MOS transistors are relatively simple to manufacture and they combine the advantages of high input impedance and characteristics that give very little cross-modulation. However, they are not very suitable for amplification at frequencies above about 100 MHz because of their relatively high feedback ("Miller-effect") capacitance.

The same problem is encountered in thermionic devices in the triode valve. The usual solution adopted there has been to replace the triode either by a pentode or by a pair of triodes connected in a "cascode" circuit. In solid-state electronics an analogous arrangement to the cascode has been made up from two MOS transistors, and this gives a feedback capacitance much less than that of a single MOS transistor. At Matsushita Electronics Corporation we have been looking into the possibility of producing a practical circuit element of even better performance by making a cascode of two MOS transistors as a solid circuit [1]. Our investigations have been successful and their practical end-result is a new circuit element, the MOS tetrode [2]. This new device does indeed have the desired very small feedback capacitance, and it also surpasses the ordinary MOS transistor on two other counts: in circuits with automatic gain control (AGC) — readily produced with the MOS tetrode, as we shall see below — there is much less cross-modulation with the tetrode, and there is much less variation of the input capacitance with input voltage. The characteristics are also very stable.

In a valve cascode circuit (*fig. 1*) the operating conditions of two triodes are such that the grid current is very small, and the anode current in both valves is about the same. If the grid G_2 is decoupled to prevent the appearance of an a.c. voltage, then the circuit behaves like a triode of very high internal impedance (anode impedance) R_t and very high amplification factor μ_t; the transconductance g_{mt} is about the same as that of one of the triodes. In this case, which corresponds to the situation in the MOS tetrode, we have [3]:

$$R_t = (\mu_2 + 1)R_1 + R_2 \approx R_1\mu_2, \quad \ldots\ldots\ldots \quad (1)$$

$$\mu_t = \mu_1\mu_2 + \mu_1 \approx \mu_1\mu_2 \quad \ldots\ldots\ldots\ldots \quad (2)$$

and

$$g_{mt} = g_{m1}\frac{\mu_2 + 1}{(\mu_2 + 1) + R_2/R_1} \approx g_{m1}, \quad \ldots\ldots \quad (3)$$

from which these statements can readily be verified.

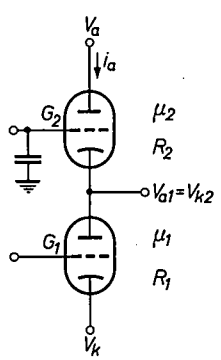


Fig. 1. A valve cascode circuit using triodes. The circuit can be adjusted in such a way that the characteristic curves are analogous to those of a single triode. The internal impedance and the current amplification factor are however much higher, and the feedback from anode to control grid is much less. The transconductance is about the same.

Fig. 2 shows a diagram of the cross-section of one of our designs of a MOS tetrode, next to a section of an ordinary MOS (triode) transistor. The diagram shows that the tetrode has two control electrodes (gates) G_1 and G_2. Between the two gates there is an N-type region which is surrounded by the P-type silicon of the substrate. This N-type region is the "island", which serves as the drain electrode for the triode formed by S, G_1

Dr. T. Okumura is with the Research Laboratory of Matsushita Electronics Corporation, Takatsuki, Osaka, Japan.

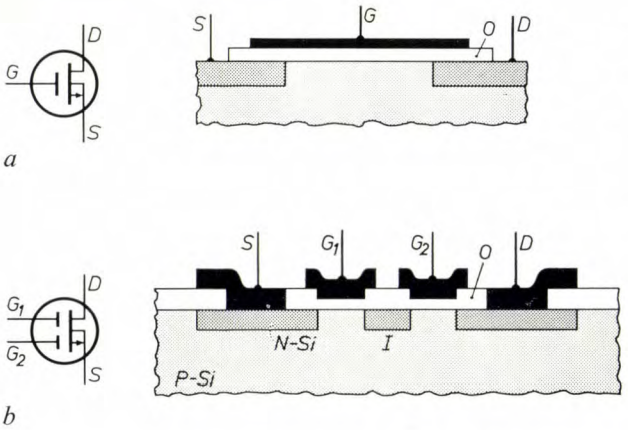


Fig. 2. *a*) Diagrammatic cross-section of a MOS transistor (MOS triode). *S* source. *D* drain. *G* metal gate electrode, insulated from the other parts of the device by a thin oxide layer *O*. *S* and *D* are of the same semiconductor material as the rest of the substrate (shaded) but are of the opposite conduction type. *b*) Diagram of the cross-section of a MOS tetrode on a *P*-type silicon substrate. There is a third *N*-type region, the island *I*, between *S* and *D*. There are now two gate electrodes, G_1 and G_2. The device can be considered to be an integrated cascode circuit of two MOS triodes.

Each of the two circuit elements has a substrate contact which is usually connected to the source. The standard symbols for use in circuit diagrams are shown on the left. The arrow represents the source.

on V_g, and the same is therefore true for the current I_d which flows through it from *S* to *D*.

If curves of the variation of this current with the voltage $V_d - V_s$ are plotted for constant V_g, it is found that the first part of each curve is parabolic and the remainder is very nearly a horizontal straight line (*fig. 3*; V_s is assumed to be zero, as usual). The maximum value $I_{d\,sat}$ of the current is the ordinate value of the straight-line region and occurs consequently in a large interval of V_d values (saturation). The value of $I_{d\,sat}$ is dependent on V_g (fig. 3*b*). In the "saturated" region I_d and V_g are related at constant V_d by

$$I_{d\,sat} = \tfrac{1}{2}\,\beta(V_g - V_{th})^2 \quad . \; . \; . \; . \; . \; . \quad (4)$$

to a very good approximation. Here β is equal to $C\mu/l^2$, where C is the capacity between the gate electrode and the channel, μ the mobility of the majority carriers (the electrons in this case) and l is the length of the channel. V_{th} is the threshold voltage, i.e. the value of V_g for which I_d becomes zero. The variation of the current with V_g is thus quadratic (fig. 3*c*). It can be shown directly from (4) that $g_{m\,sat}$, the value of the transconductance under saturation conditions,

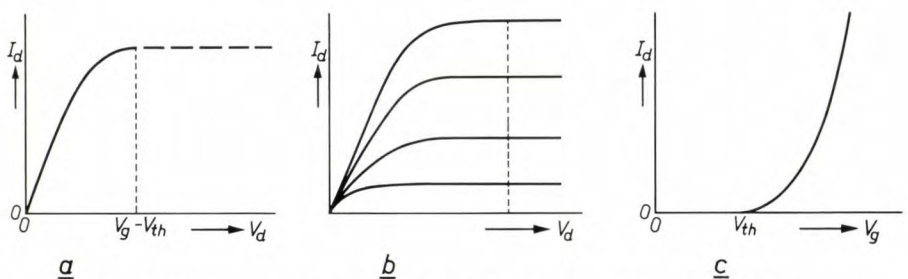


Fig. 3. *a*) The I_d-V_d characteristic for a MOS triode for a fixed value of V_g. The curve can be fairly accurately represented by a combination of a parabolic region and a nearly horizontal straight line. *b*) Set of I_d-V_d characteristics for various values of V_g. *c*) I_d-V_g characteristic for constant V_d (see the dashed curve in *b*). At $V_g < V_{th}$ the current I_d is virtually zero. The parabolic and the horizontal part of the I_d-V_d characteristics meet at $V_d = V_g - V_{th}$.

and *I* and as the source electrode for the triode formed by *I*, G_2 and the drain electrode *D*. The island has no contacts. In most applications the signal to be amplified is applied to G_1 and a capacitor is connected between the electrode G_2 and earth. The substrate has a contact (not shown), which is connected to the source.

Before describing the special electrical and technological features of our new tetrodes, we should perhaps recapitulate some of the more important features of the MOS triode. The fundamental mechanism in the operation of a MOS triode (fig. 2*a*) which has a *P*-type substrate, is the formation at the substrate surface of a thin *N*-type layer when a positive bias V_g is applied to the gate (inversion). The charge density of the mobile majority carrier in this layer, the "channel", depends

is given by:

$$g_{m\,sat} = (\partial I_{d\,sat}/\partial V_g) = \beta(V_g - V_{th}). \quad . \; . \; . \quad (5)$$

The transconductance $g_{m\,sat}$ is thus not independent of V_g.

It is the very nearly quadratic relation between $I_{d\,sat}$ and V_g which accounts for the very small cross-modulation obtained from the MOS transistor [4]. If a

[1] Characteristics and production methods for solid circuits are discussed in: A. Schmitz, Philips tech. Rev. **27**, 192-199, 1966.

[2] During the development of our tetrode some details of similar work were given by N. H. Ditrick and M. M. Mitchell at a meeting of the IEEE Electron Devices Group in October 1965.

[3] See G. Klein and J. J. Zaalberg van Zelst, Precision electronics, Centrex Publ. Co. Eindhoven 1968, Chap. 18.

[4] This is discussed by P. E. Kolk and I. A. Maloff, Electronics **37**, Dec. 14, 1964, p. 71.

radio receiver is tuned to a particular frequency, the modulation of the signal received is affected to some extent by the modulation of other transmitters, particularly those at neighbouring frequencies. The magnitude of this effect is mainly determined by the third-power term — and to a lesser extent by the higher-power odd terms — in the expansion of $I_{d\,sat}$ as a power series in V_g. For the MOS transistor these terms are very small and the cross-modulation is therefore correspondingly small.

MOS transistors for use in computers usually have a substrate of *N*-type silicon, and hence a *P*-type channel. In these transistors V_{th} usually has such a value that the transistor does not conduct when there is no bias on the gate, corresponding to the usual requirements in digital applications. MOS transistors for amplifying high-frequency signals, on the other hand, are preferably made on a *P*-type substrate. The channel then shows *N*-type conductivity, which is more suitable here since the electrons in silicon have a mobility about three times that of the holes. (It is true that the mobility in the layer obtained by inversion is smaller than it is in the bulk material of similar type, but the relative decrease is about the same for both kinds of charge carrier [5].)

The operation of a MOS tetrode

Let us now calculate how the current in a MOS tetrode depends on the various potentials and characteristic quantities. We shall proceed as if the tetrode was a cascode circuit of two triodes: in fact this assumption is not quite justified since a tetrode has one substrate contact, while the substrate of each triode of a cascode circuit is connected to its own source electrode. We shall also assume that the two triodes are identical, so that they each have the same β, V_{th}, etc.

If an equation analogous to (4) is set down for the two triodes, then since both transistors carry the same current it can be shown that

$$V_{g2} - V_{isl} = V_{g1}, \quad \ldots\ldots\ldots \quad (6a)$$

or

$$V_{isl} = V_{g2} - V_{g1}, \quad \ldots\ldots \quad (6b)$$

and

$$I_{d\,sat} = \tfrac{1}{2}\beta(V_{g1} - V_{th})^2. \quad \ldots\ldots \quad (7)$$

This equation for the saturation current of the tetrode is identical with the one that holds for the first triode (see eq. 4). The diagrams of *fig. 4* show how the triodes adjust themselves for the different cases (cf. eq. 6b).

When V_{g1} is greater than $V_{isl} + V_{th}$, the current in the first triode is no longer saturated, and is given by:

$$I_{d1} = \frac{\beta}{2}\{(V_{g1} - V_{th})^2 - (V_{g1} - V_{th} - V_{isl})^2\}. \quad (8)$$

From this relation and the similar one for I_{d2} it follows that

$$V_{isl} = \tfrac{1}{2}[V_{g1} + V_{g2} - 2V_{th} - \{(V_{g1} + V_{g2} - 2V_{th})^2 - 2(V_{g2} - V_{th})^2\}^{\frac{1}{2}}], \quad (9)$$

and

$$I_d = \frac{\beta}{8}[V_{g2} - V_{g1} + \{(V_{g1} + V_{g2} - 2V_{th})^2 - 2(V_{g2} - V_{th})^2\}^{\frac{1}{2}}]^2. \quad \ldots \quad (10)$$

Let us now look at a number of special cases. When V_{g1} is equal to V_{th}, the first triode is cut off and the potential V_{isl} of the island becomes equal to $V_{g2} - V_{th}$; this means that the second triode is just driven into cut-off. If V_{g1} is allowed to increase indefinitely, then, as (9) shows, V_{isl} goes to zero (i.e. V_{isl} becomes equal to V_s) and I_d goes to the limiting value $\frac{1}{2}\beta(V_{g2} - V_{th})^2$.

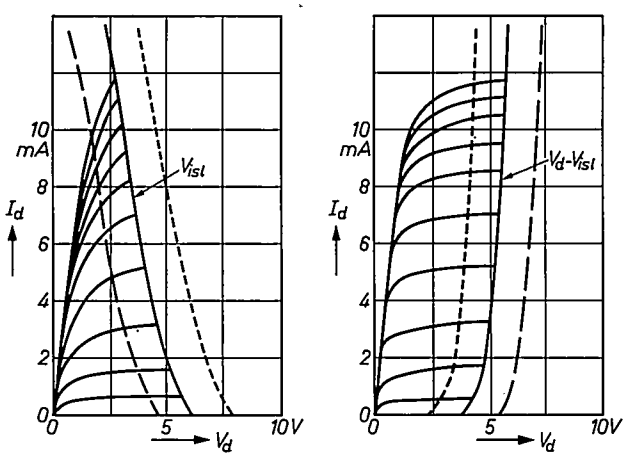


Fig. 4. Operating curves for a pair of MOS triodes in a cascode circuit, those for the first triode on the left and those for the second on the right. The drain potential of the first triode, which is also the source potential for the second, is labelled V_{isl}, since it corresponds to the potential of the "island" of a tetrode. The curves for V_{isl} and V_d—V_{isl} apply for constant V_{g2}. At a lower value of V_{g2} this curve takes up the position indicated by the dashed line, and at higher V_{g2} it takes up the position shown by the dotted line.

This limiting of I_d has advantages on the score of reliability, since a high voltage applied at the gate cannot set up a current large enough to damage the tetrode.

Finally, we should say a word or two about the most important quantity of all in the MOS tetrode — the feedback capacitance. In the MOS triode the feedback action arises because the gate electrode overlaps the drain electrode a little (*fig. 5a*). In the MOS tetrode the main cause of feedback is that any a.c. voltage

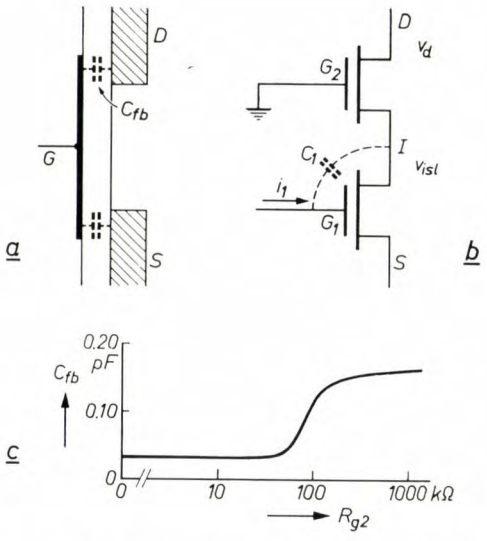


Fig. 5. *a*) The small stray capacitances in a MOS tetrode between G and D and between G and S arise mainly because G slightly overlaps the two other electrodes. The stray capacitance C_{fb} between G and D is responsible for the feedback action. *b*) Diagram referring to the calculation of the feedback capacitance of a MOS tetrode. *c*) The feedback capacitance C_{fb} as a function of the resistance in the lead to the second gate electrode. If this resistance is greater than a few tens of kΩ, the advantages that can be gained with the tetrode are partly lost.

at D sets up a corresponding (but much smaller) a.c. voltage at I, and this in turn is coupled back to G_1 (fig. 5*b*). The direct effect of the potential of G_1 on that of D is very small, since G_2 functions as a screening electrode, provided that the resistance in the lead of G_2 is not too high (fig. 5*c*).

The feedback via the island can be calculated as follows (fig. 5*b*). If the feedback capacitance of the tetrode is represented by C_{fb} and the a.c. current induced in G_1 by an a.c. voltage v_1 superimposed on V_d is represented by i_1, then:

$$i_1 = -j\omega C_{fb} v_d = -j\omega C_1 v_1, \quad \ldots\ldots \quad (11)$$

or

$$C_{fb} = C_1 v_1 / v_d.$$

Since the same current flows in both parts of the tetrode, v_1/v_d is equal to R_1 divided by the total resistance of the cascode circuit, which is equal to $R_1(1 + \mu_2) + R_2$. Consequently:

$$C_{fb} = C_1 \frac{R_1}{R_1(1 + \mu_2) + R_2} \approx C_1/\mu_2. \quad \ldots \quad (13)$$

This shows that in the MOS tetrode the feedback capacitance is about μ_2 times smaller than that of the triode.

If one attempts to verify experimentally that V_{isl} becomes equal to $V_{g2} - V_{th}$ when V_{g1} is made equal to V_{th}, as the relation (6) predicts, a fairly large discrepancy is found. With triodes whose characteristics were like those of fig. 3, and for which $V_{th} = -2.7$ V, we found a value of 6.2 V for V_{isl} at $V_{g2} = 5$ V, as compared with a calculated value of 7.7 V. A much smaller discrepancy was found when we used triodes made on a much less strongly doped silicon substrate (20 Ωcm). This agrees with the calculations of Van Nielen and Memelink [6], from which it can be shown that there will be no discrepancy at all in transistors made on an intrinsic subtrate.

Some suitable configurations

Fig. 6 shows a diagram of the cross-section of two of our experimental MOS tetrodes, which fulfilled expectations in every respect. The configuration shown in fig. 6*a* is very simple and does not really need further explanation. In the configuration of fig. 6*b* it can be seen that the oxide layer does not have the same thickness everywhere. Because of the step in thickness the two triodes have different values of C (the capacitance is of course inversely proportional to the thickness of the oxide layer). In the tetrode of fig. 6*b* C is greater for the second triode, and it can be shown

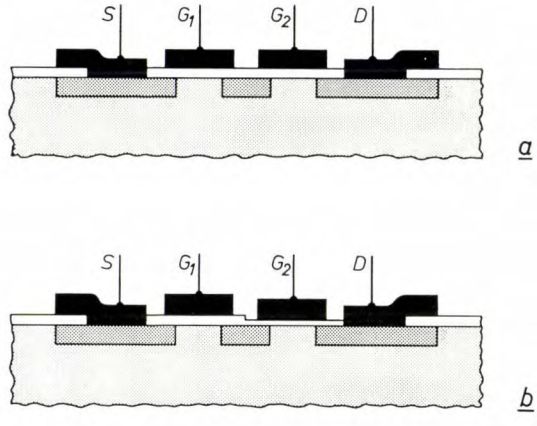


Fig. 6. Diagram of the cross-section of two types of experimental MOS tetrode. In the lower diagram the oxide layer of the second tetrode is thinner than that of the first one, which gives a smaller feedback capacitance and also makes it easier to apply automatic gain control (AGC) via G_2.

that this has the effect of making its current amplification factor μ greater. It follows that for this configuration C_{fb} will be even smaller than it is for that of fig. 6*a*. This arrangement also has the advantage that a smaller voltage is needed at G_2 when the tetrode is used for AGC. As in other tetrodes the cross-modulation in reverse AGC with V_{g2} (see below) is less than in reverse AGC with V_{g1}, so that this configuration combines a number of attractive features.

A variation of the tetrode drawn in fig. 6*a* is now

[5] O. Leistiko, A.S. Grove and C. T. Sah, IEEE Trans. **ED-12**, 248, 1965.

[6] J. A. van Nielen and O. W. Memelink, The influence of the substrate upon the DC characteristics of silicon MOS transistors, Philips Res. Repts. **22**, 55-71, 1967.

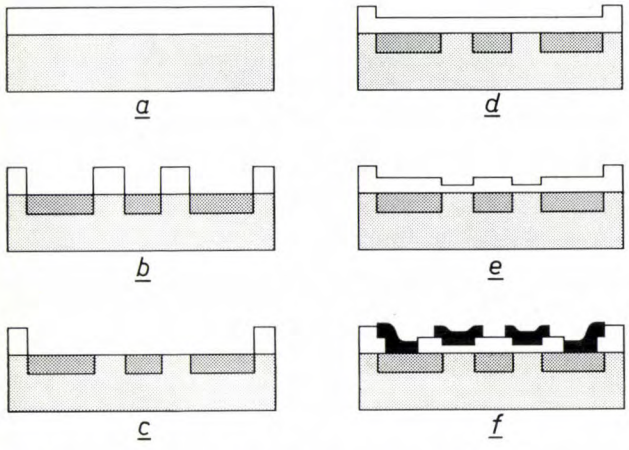


Fig. 7. Stages in the production of the Matsushita 3 SK 32 MOS tetrode. *a*) Oxidation of the silicon substrate. *b*) Oxide is etched away in a particular pattern and phosphorus is diffused into the substrate to obtain source, island and drain. *c*) Removal of the oxide layer. *d*) Second oxidation; the oxide between *S*, *I* and *D* is etched away. *e*) Third oxidation. *f*) Oxide is removed from parts of *S* and *D* and aluminium is deposited.

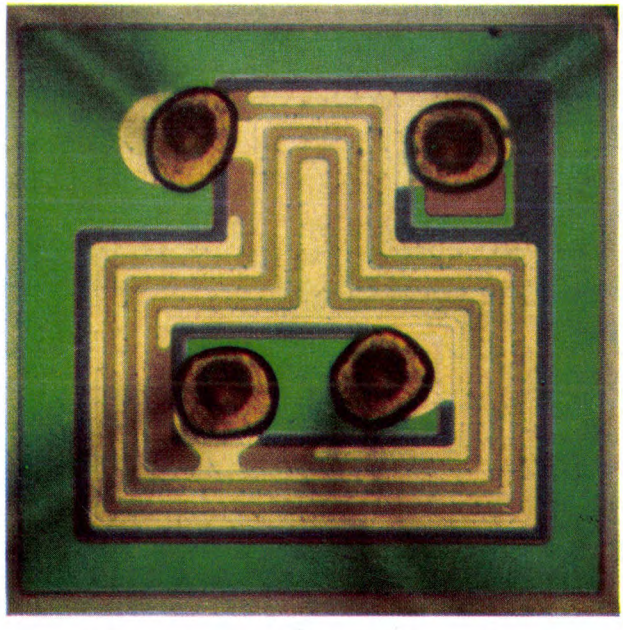


Fig. 8. Microphotograph of a silicon wafer on which a 3 SK 32 MOS tetrode has been produced. Starting at the outside, the yellow lines are the source, the first gate, the second gate and the drain. The four dots are the connections. The four lead wires are blurred but can just be made out.

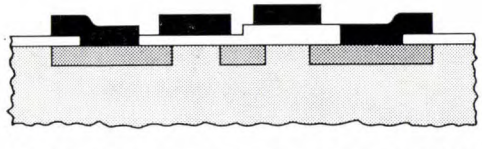


Fig. 9. Experimental MOS tetrode in which the second triode is the one with the thicker oxide layer.

in production (3 SK 32 MOS tetrode). Its characteristic features will be described below; a diagrammatic cross-section and the method of making it are shown in *fig. 7*. The cross-section has been made different from the one of fig. 6*a* to reduce as far as possible the effect of the overlap of the two gate electrodes on *S*, *I* and *D*. *Fig. 8* is a microphotograph of the surface of a silicon wafer on which a tetrode like that of fig. 7*f* has been produced.

Experiments have also been carried out with two other configurations besides the ones shown in fig. 6. The first is like the configuration of fig. 6*a*, but with the island left out. We found that tetrodes of this type were not very successful; the potential of the oxide between G_1 and G_2 changes when V_{g1} or V_{g2} changes, but with a very high time constant. The conductivity of the channel between G_1 and G_2 is therefore not uniquely determined by V_{g1} and V_{g2}.

The fourth configuration that we have investigated [7] is drawn in *fig. 9*. Here the oxide layer of the *second* triode is thicker than that of the first one. This has the result that V_{th} is higher for the second triode than for the first, which means that the island potential is fairly high even at $V_{g2} = 0$; consequently at $V_{g2} = 0$ the current I_d is so large that a fairly high transconductance is obtained.

The Matsushita very high frequency MOS tetrode type 3 SK 32

The 3 SK 32 MOS tetrode, whose configuration has just been described (fig. 7 and 8), is intended for use at VHF (30 to 300 MHz), and is hermetically sealed in a standard TO-72 package. The I_d-V_d characteristics of a sample of this type are reproduced in *fig. 10*. The value of V_{th} is the same for both triodes

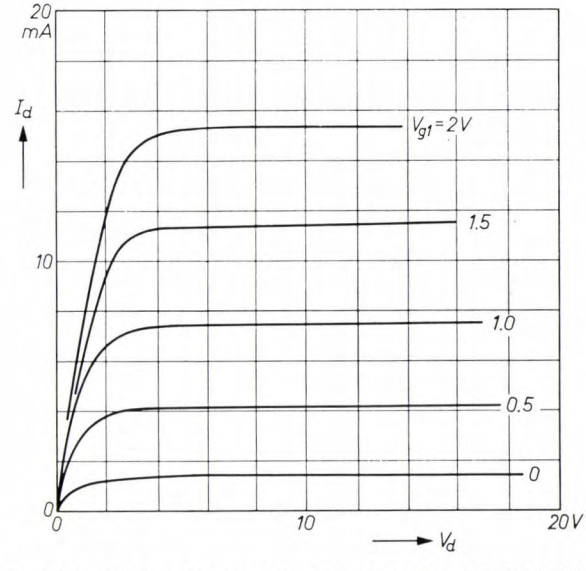


Fig. 10. Set of I_d-V_d characteristics for the 3 SK 32 MOS tetrode for $V_{g2} = 5$ V. At $V_{g1} = 0$ the current I_d is about 1.5 mA.

and is in the region of −0.6 V. At $V_{g1} = 0$ the saturation value of I_d is about 1 mA, and this means that a bias on G_1 is desirable in most applications (enhancement mode operation).

When V_d is set equal to zero and V_{g1} and V_{g2} are made so strongly negative that there is no current in the tetrode ($V_{g1} = V_{g2} = -10$ V), the small-signal input capacitance and output capacitance are 4.5 pF and 3 pF respectively at a frequency of 455 kHz. The variation of the feedback capacitance C_{fb} as a function of I_d is shown in *fig. 11*. The value of C_{fb} is about 0.02 pF, except at very low I_d. In general C_{fb} is never greater than about 0.035 pF.

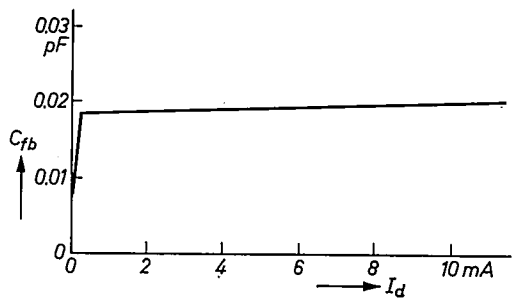


Fig. 11. The feedback capacitance C_{fb} for the 3 SK 32 tetrode as a function of I_d, at a frequency of 445 kHz, with $V_d = 10$ V and $V_{g2} = 5$ V.

Fig. 12 shows how the transconductance varies with V_{g1}, with V_{g2} as a parameter. It can be seen that all the curves have a maximum, which means that g_m can either increase or decrease with increasing V_{g1}. A reduction in V_{g2} always gives a reduction of g_m. This means that there are in principle three ways of obtaining automatic gain control: two using V_{g1} and one using V_{g2}. The most attractive method of the three is the one using V_{g1} — which is chosen to give an operating point in the downward-going part of the curve — since the magnitude of the variation of g_m with V_{g1} can be anywhere within certain limits; all that has to be done here is to set V_{g2} to a suitable value. The fact that a fairly high V_{g1} has to be used in this method of control so that I_d also has a rather high value, is a disadvantage.

Fig. 13 shows how the power gain G and the noise figure F of the 3 SK 32 tetrode depend upon I_d at the frequency 200 MHz and for three values of V_{g2}. For still higher values of V_{g2} the measured curves almost coincide with those for $V_{g2} = 5$ V. It can be seen that for normal conditions and optimum matching the power gain at 200 MHz is 20 dB or more.

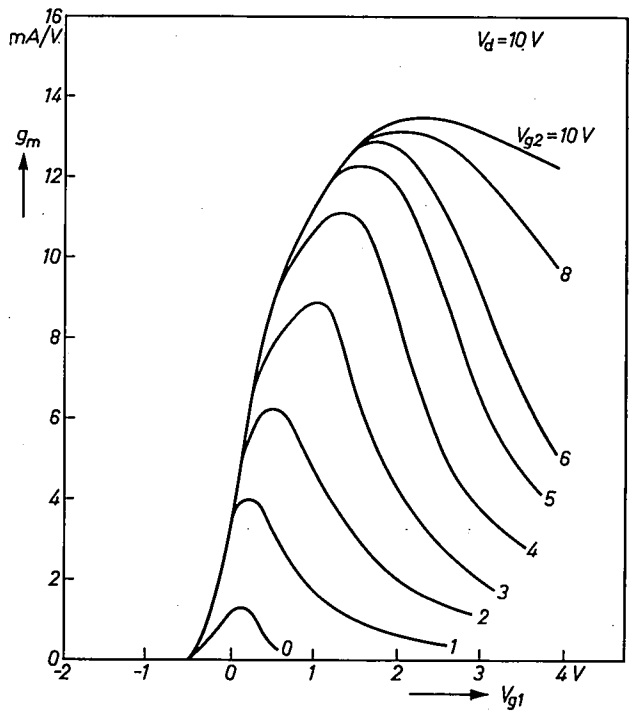


Fig. 12. The transconductance g_m of the 3 SK 32 tetrode as a function of V_{g1} for nine different values of V_{g2}, with $V_d = 10$ V.

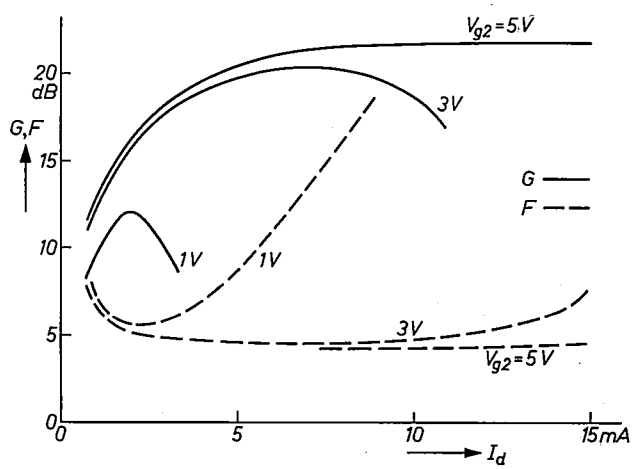


Fig. 13. The power gain G (solid lines) and the noise figure F (dashed lines) of the 3 SK 32 tetrode, measured at 200 MHz. Both quantities were measured under optimum conditions and with a bandwidth of 4 MHz.

The cross-modulation characteristics for circuits using two methods for obtaining AGC are shown in *fig. 14*. The curves were obtained with an untuned circuit which had two signals applied to it, one with a carrier frequency of 200 MHz and the other with a

[7] See also Electronics 39, May 16, 1966, p. 212-213.

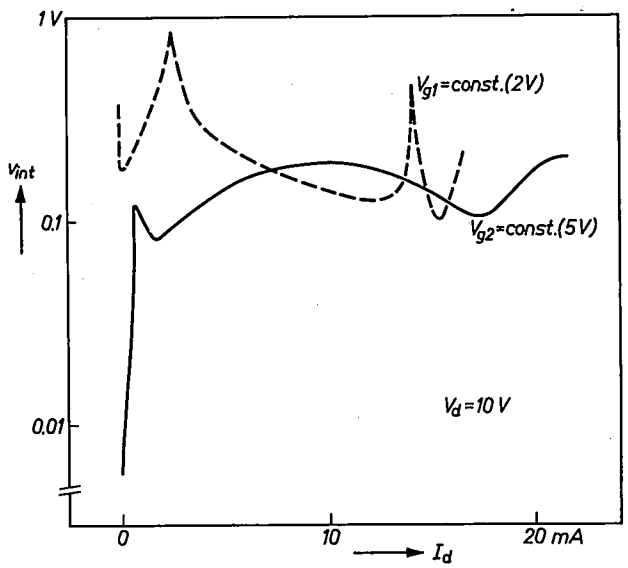


Fig. 14. The interfering signal v_{int} (frequency = 212 MHz), which gives a cross-modulation of 1% in a signal at a frequency of 200 MHz in the 3 SK 32 tetrode, shown as a function of I_d for $V_d = 10$ V. The solid line applies for constant V_{g2} — i.e. the normal situation with the tetrode driven at G_1 — and the dashed line refers to constant V_{g1}.

carrier frequency of 212 MHz. It can be seen that for normal values of I_d a cross-modulation of 1% occurs when the interfering signal is 0.1-1 V, and that the case with fixed V_{g1} is the most favourable one.

In MOS triodes and tetrodes the input capacitance C_i is not entirely independent of the applied voltages. Since this capacitance can contribute to the tuning of a circuit, it is desirable that the variation of C_i should be as small as possible. *Fig. 15* shows how C_i varies with I_d; in one case I_d is controlled by V_{g1} and in the other it is controlled by V_{g2}. It can be seen that the most favourable situation is the one in which V_{g1} is held constant. In this respect AGC via G_2 is therefore preferable.

Since the mobility of the charge carriers and the potential V_{th} both depend on temperature, the current I_d and the transconductance g_m are also to some extent temperature-dependent at constant V_d, V_{g1} and V_{g2}. Conditions can be found for which the two effects cancel one another out. If $\alpha\mu$ is substituted for β, it can be readily shown from (4) and (5) that the two temperature coefficients are given by:

$$\frac{\partial I_d}{\partial T} = \frac{I_d}{\mu}\frac{\partial \mu}{\partial T} - \sqrt{2\,\alpha\mu I_d}\,\frac{\partial V_{th}}{\partial T}, \quad . \quad (14)$$

$$\frac{\partial g_m}{\partial T} = \frac{g_m}{\mu}\frac{\partial \mu}{\partial T} - \alpha\mu\,\frac{\partial V_{th}}{\partial T}. \quad . \quad . \quad (15)$$

Fig. 16 shows how I_d varies with the channel temperature at $V_d = 10$ V and $V_{g2} = 5$ V. At $V_{g1} = 0$ V (or $I_d = 1.5$ mA) the effect of the temperature is almost zero, for $V_{g1} < 0$ the temperature coefficient is positive but very small, and for $V_{g1} > 0$ it is negative. This negative temperature coefficient is another feature which is useful for reliability. (Bipolar transistors do not have this attractive feature, and the current in these is also more strongly temperature-dependent than the current in MOS transistors.)

As in other MOS transistors the V_{th} of our tetrode

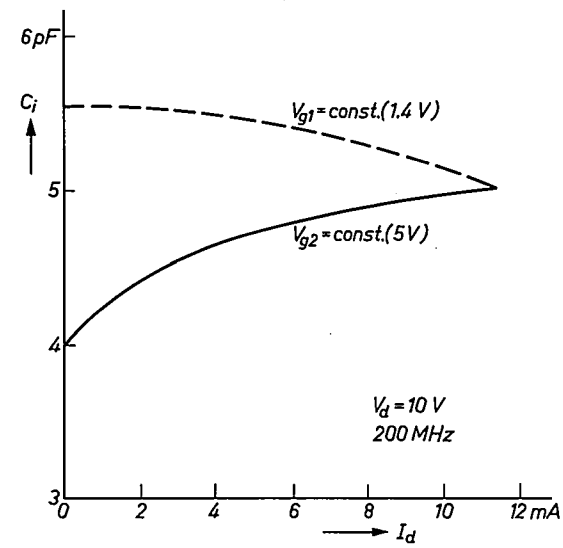


Fig. 15. The input capacitance C_i of the 3 SK 32 tetrode at a frequency of 200 MHz as a function of the current I_d at constant V_{g2} (5 V) and constant V_{g1} (1.4 V), with $V_d = 10$ V. The variation of C_i with I_d is smallest at constant V_{g1}.

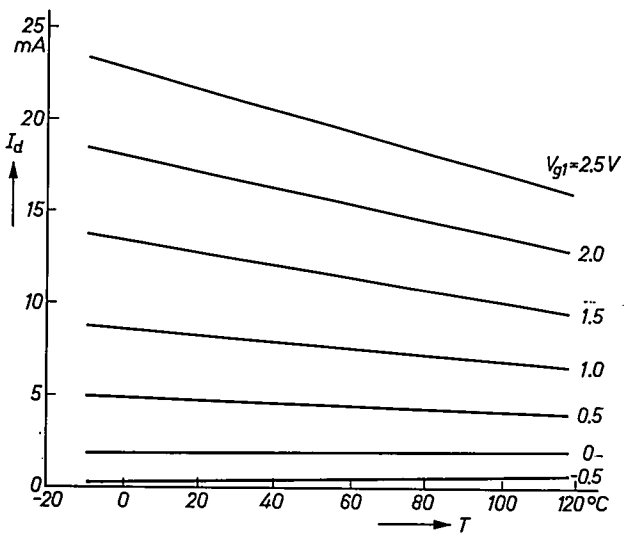


Fig. 16. Variation of I_d with channel temperature at constant V_d and V_{g2} for a number of values of V_{g1}. When V_{g1} is small, I_d is almost independent of the temperature.

is not absolutely constant. *Fig. 17* shows values of V_{th} recorded for four samples over a period of 1000 hours. At first there was a relatively rapid variation, although it was not very large (about 0.05 V in 150 hours), and after that V_{th} remained virtually constant for all four tetrodes.

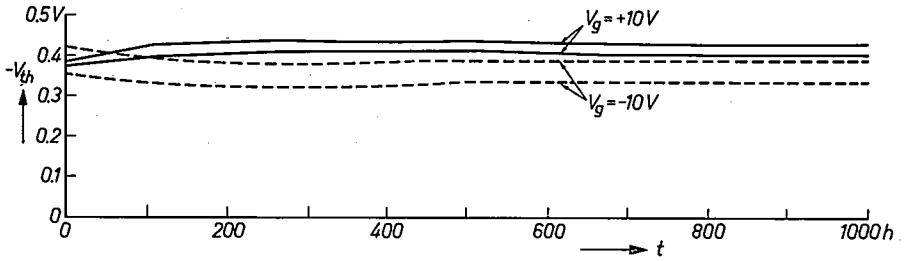


Fig. 17. A record of the variation of the values of V_{th} measured over a period of 1000 hours for four samples of the 3 SK 32 tetrode at a temperature of 150 °C. During the experiments S was connected to D and G_1 to G_2. The voltage between the two pairs of electrodes was +10 V for two of the tetrodes, and —10 V for the other two.

Summary. The MOS tetrode, an integrated cascode circuit of two MOS triodes, has been developed to obtain a MOS device with such a low feedback capacitance that operation in the VHF band would be possible. The slope is about the same as for a MOS triode, and the current amplification factor and the internal impedance are very high. In normal use the gate G_2 of the second triode is kept at a fixed potential (say zero) and the tetrode is driven at G_1. In the 3 SK 32 VHF tetrode developed by Matsushita Electronics Corporation the threshold voltage V_{th} is about —0.6 V; at $V_{g1} = 0$ the current I_d is about 1.5 mA, so that a bias is needed for certain applications. The power gain is 20 dB or more at 200 MHz. The transconductance (about 10 mA/V) has a maximum for a particular value of V_{g1} and also depends on V_{g2}, so that there are three possible ways of obtaining AGC. The feedback capacitance is about 0.02 pF. The cross modulation is very low. At low V_{g1} the current I_d is very nearly independent of the temperature. During the first 150 hours of life V_{th} changes a little, but after this changes hardly at all.

Stoichiometry

W. Albers and C. Haas

III. The determination of existence regions

As we stated earlier in part I, semiconductors and insulators usually have narrow existence regions and compounds of metallic type usually have broad existence regions [*]. This distinction can be made clearer by considering the electron energies. Let us take as an example the solution of native atoms *A in a compound AB. The dissolved native atoms are often ionized, for example as donors, and this can be represented by $*A \rightleftarrows *A^+ + e$. When the ionization is complete (i.e. the ionization energy $E_d < kT$), the energy of solution of A in AB consists of two terms, a term $E(*A^+)$ for the incorporation of $*A^+$, and a term $E(e)$ representing the energy of the electron: $E(A) = E(*A^+) + E(e)$. As the total energy of solution $E(A)$ increases, the existence region becomes narrower (see figs. 4 and 5 in I). In general, a high value of $E(e)$ will therefore give a narrow existence region.

We can obtain some idea of the electron energy $E(e)$ by considering the relative positions of the energy levels of the different electron states. The states that are of significance in semiconducting or insulating compounds of atoms with filled inner shells (CdTe, GaAs, Al_2O_3, H_2O etc.) are shown in *fig. 12b*. Overlapping s and p orbitals form a broad empty conduction band (the width ΔE_c is a few tens of electron volts) and a broad filled valence band, separated by a forbidden energy zone ΔE_g with a width of the order of 1 eV. If A is dissolved as the ionized form $*A^+ + e$, electrons enter the conduction band. In semiconductors with a large band gap ΔE_g, $E(e)$ will then be high, and in general a narrow existence region is to be expected. Moreover, if greater quantities of *A are added, the Fermi level in the conduction band will increase rapidly, so that $E(e)$ also increases rapidly with the concentration of the dissolved atoms.

It is possible for the native atoms *A dissolved in AB to behave as donors, yet with only some of them ionized, since $E_d > kT$. In this case the electrons are in donor levels that lie between the valence band and the conduction band, and $E(e)$ should be taken as the energy of the actual donor level rather than that of the conduction band. If the donor level is high, the electron energy $E(e)$ will also make a significant contribution to the total energy of solution in this case, and thus give narrow existence regions.

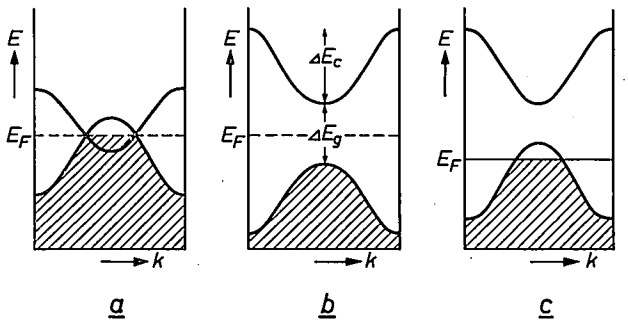


Fig. 12. Electron energy diagrams for metals (*a* and *c*) and for a semiconductor (*b*). *k* wave vector. E_F Fermi level. ΔE_c width of conduction band. ΔE_g forbidden energy zone. The horizontal solid line in *c* represents a localized level or a narrow energy band (whose width ΔE_c, for example 0.1 eV, is not indicated).

In compounds with metallic conduction, on the other hand, the valence and conduction bands overlap (see fig. 12*a*) or there is already a partly filled band in the stoichiometric composition, so that $\Delta E_g = 0$. In this case an energy ΔE_g or $\Delta E_g - E_d$ does not have to be provided before an extra electron can be included and $E(e)$ is small. The existence regions can then be broad, partly because the additions of large quantities of *A will only cause the energy level to increase slowly because of the high density of states at the Fermi level.

This ties up with the striking disparity between compounds such as GaSb and MnSb. The chalcogenides (compounds of S, Se, Te) of the elements of the second group in the periodic system, and the pnictides (compounds of P, As, Sb, Bi) of the elements of the third group are semiconductors and have narrow existence regions (of the type shown in fig. 12*b*). The corresponding chalcogenides and pnictides of the transition metals, however, give metallic conduction and have broad existence regions. This difference in widths can

Dr. W. Albers is with Philips Research Laboratories, Eindhoven; Prof. Dr. C. Haas, formerly with these Laboratories, is now Professor of Inorganic Chemistry at the University of Groningen.

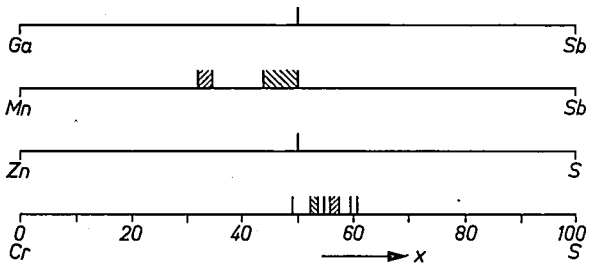


Fig. 13. Chemical spectra, i.e. cross-sections through the T-x diagram at constant T ($\approx$ 500 °C), of the systems Ga-Sb, Mn-Sb, Zn-S and Cr-S. The composition parameter x is expressed as atomic percentages of S or Sb. It can be seen that the Mn compounds of Sb have broader existence regions than the Ga compound. The same holds for the Cr compounds of S as compared with those of Zn.

clearly be seen in *fig. 13*, which shows the chemical spectra of these systems, i.e. a cross-section through the T-x phase diagram at constant T ($\approx$ 500 °C).

In transition-metal compounds two types of electron states are of importance. The s and p electrons form broad bands, whereas the inner orbitals (e.g. the 3d orbitals of the 3d transition metals) form localized levels or very narrow bands (small overlap), including empty levels, since the inner shell of the transition metal is only partly filled. Fig. 12*c* shows one of the eight possible basic types [8], in which the broad valence band is bisected by the lowest band, which is very narrow and was originally unfilled (ΔE_c, for example, of the order of 0.1 eV), so that now both bands are only partly filled. In this case $\Delta E_g = 0$ and ΔE_c is very small, and therefore broad existence regions are possible.

Narrow existence regions

The treatment given in part II enables narrow existence regions to be calculated. This will be briefly discussed with the semiconductor CdTe as an example [9]. In what follows "existence region" (projection) refers to the region of concentrations that is enclosed by the projection on the T-x plane of the curve that indicates the compositions of the co-existing solid compound in the three-phase equilibria in the three-dimensional binary P-T-x phase diagram (see fig. 19) [10]. We can calculate the existence region of CdTe if we know the equilibrium constants and the coexisting partial vapour pressure $P(\mathrm{Cd})^b$ of Cd (or that of Te) along the whole boundary of the existence region [11].

Some idea of the ways in which dissolved Cd and Te are incorporated in CdTe can be gathered from experimental data (*fig. 14*; fig. 11 in part II). CdTe (Cd) turns out to be an N-type semiconductor, and CdTe(Te) a P-type semiconductor. Self-diffusion measurements with radioactive cadmium in samples of CdTe with various deviations from the stoichiometric composition give an indication that the disorder is of the Frenkel type. At high pressures P(Cd) the electron concentration is found to be proportional to $P(\mathrm{Cd})^{1/2}$. At high P(Cd), therefore, Cd_i^+ and e will be dominant, and at low P(Cd): V_{Cd}^- and h. Other native point defects found are effectively neutral Cd_i and V_{Cd} and the effectively double-charged cadmium vacancy V_{Cd}^{2-}.

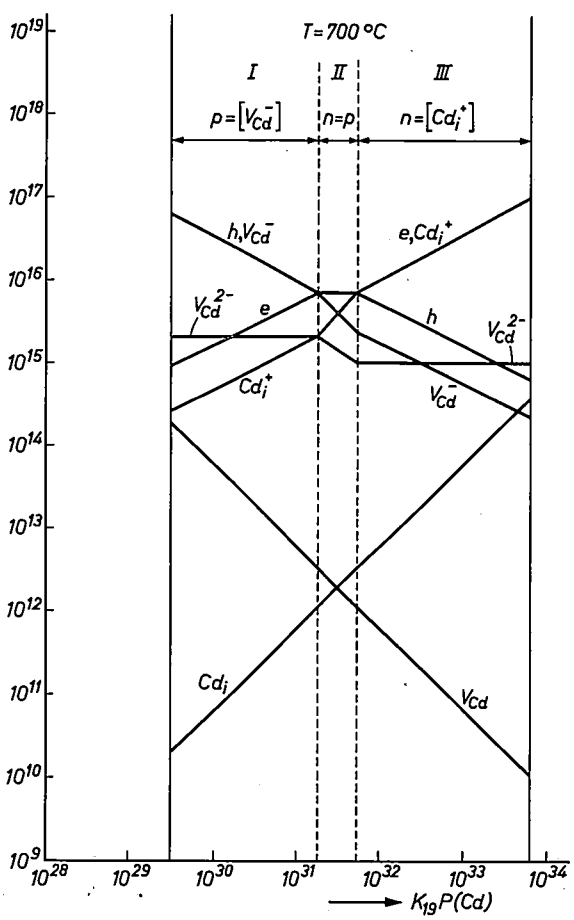


Fig. 14. Concentrations of electrons, holes and native point defects in CdTe as a function of P(Cd) at 700 °C. The vertical lines indicate the boundaries of the existence region. (This same figure was used for fig. 11 of part II.)

[*] Parts I and II of this article: Stoichiometry, I. Existence region, and Stoichiometry, II. Point defects and the control of their concentrations, have appeared in Philips tech. Rev. **30**, 82-88 and 107-112, 1969 (Nos. 3 and 4).

[8] W. Albers and C. Haas, Physics Letters **8**, 300, 1964.

[9] D. de Nobel, Philips Res. Repts. **14**, 361 and 430, 1959.

[10] The existence region (projection) given in fig. 16 should be distinguished from the existence region (cross-section) given in fig. 7 in part I, which presents a section at constant total pressure P through the three-dimensional P-T-x figure.

[11] In part II the symbols $P(\mathrm{A})^{\mathrm{AB(A)}}$ and $P(\mathrm{A})^{\mathrm{AB(B)}}$ were used, but when the entire existence region is under consideration, the AB(A) boundary and the AB(B) boundary merge smoothly into one another and we can use a single symbol $P(\mathrm{Cd})^b$, where b stands for boundary.

The appropriate series of equilibrium relations is:

$$\left.\begin{aligned} P(\mathrm{Cd})\,P(\mathrm{Te_2})^{1/2} &= K(\mathrm{CdTe}) \\ np &= K_1 \\ [\mathrm{Cd_i^+}]n &= K_{19}P(\mathrm{Cd}) \\ \frac{[\mathrm{Cd_i^+}]n}{[\mathrm{Cd_i}]} &= K_{20} \\ \frac{[\mathrm{V_{Cd}^-}]p}{[\mathrm{V_{Cd}}]} &= K_{21} \\ [\mathrm{Cd_i^+}]\,[\mathrm{V_{Cd}^-}] &= K_{22} \\ \frac{[\mathrm{V_{Cd}^{2-}}]p}{[\mathrm{V_{Cd}^-}]} &= K_{23} \end{aligned}\right\} \quad (70)$$

and the electroneutrality condition for the solid phase is:

$$[\mathrm{V_{Cd}^-}] + 2[\mathrm{V_{Cd}^{2-}}] + n = [\mathrm{Cd_i^+}] + p.$$

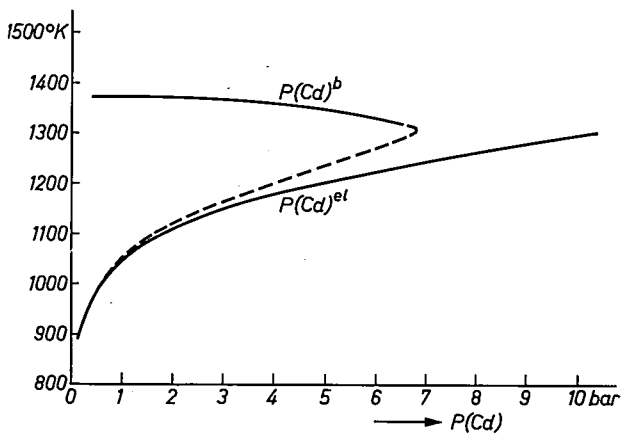


Fig. 15. The partial pressure $P(\mathrm{Cd})^b$ for CdTe saturated with Cd as a function of temperature. $P(\mathrm{Cd})^{el}$ is the saturation vapour pressure of elementary cadmium (after Lorenz [12]). The dashed part of the curve is extrapolated.

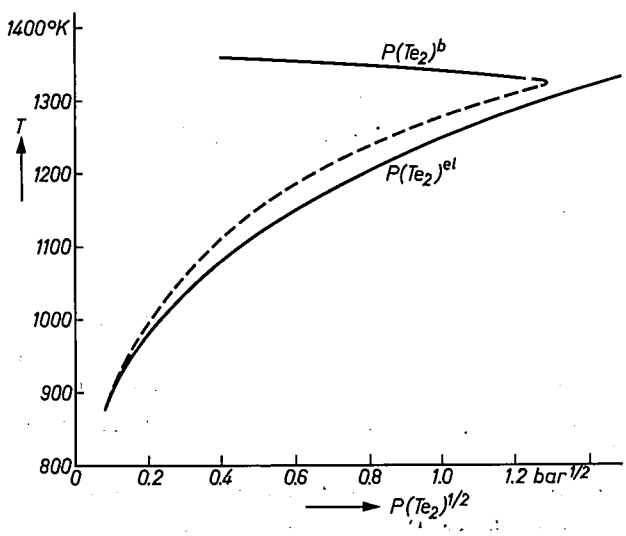


Fig. 16. The partial pressure $P(\mathrm{Te_2})^b$ for Te-saturated CdTe as a function of temperature. $P(\mathrm{Te_2})^{el}$ is the saturation vapour pressure of elementary tellurium (after Lorenz [12]).

These eight equations include $P(\mathrm{Cd})$, $P(\mathrm{Te_2})$, the concentrations of five atomic native point defects and those of holes and electrons. The concentrations of the point defects at the boundary of the existence region can thus be calculated as a function of temperature, and hence the entire existence region of CdTe, if $P(\mathrm{Cd})$ or $P(\mathrm{Te_2})$ and all the equilibrium constants K are known as a function of temperature. *Figs. 15* and *16* give $P(\mathrm{Cd})^b$ and $P(\mathrm{Te_2})^b$ as a function of T as experimentally determined [12]. The equilibrium constant $K(\mathrm{CdTe})$ can be calculated from $RT \ln K(\mathrm{CdTe}) = -\Delta G^0(T)$, where $\Delta G^0(T)$ is the Gibbs free energy of formation at T as given by the reaction

$$\mathrm{Cd(gas)} + \tfrac{1}{2}\,\mathrm{Te_2(gas)} \rightleftarrows \mathrm{CdTe(crystal)}.$$

From known thermodynamic data we find:

$$\log K(\mathrm{CdTe}) = 10.66 - 15440/T. \quad . . \quad (71)$$

From equation (71) and the data from figures 15 and 16 we now know $P(\mathrm{Cd})^b$ along the entire boundary of the existence region (see *fig. 17*). $P(\mathrm{Cd})^b$ goes to zero at $T = 0$ °K, and has kinks at 594 °K and 723 °K where first-order phase transitions from the liquid to the solid state take place, i.e. eutectics in the phase diagram of the system Cd-Te. The equilibrium constant K_1 (see *fig. 18*) has been calculated as a function of T from $K_1 =$

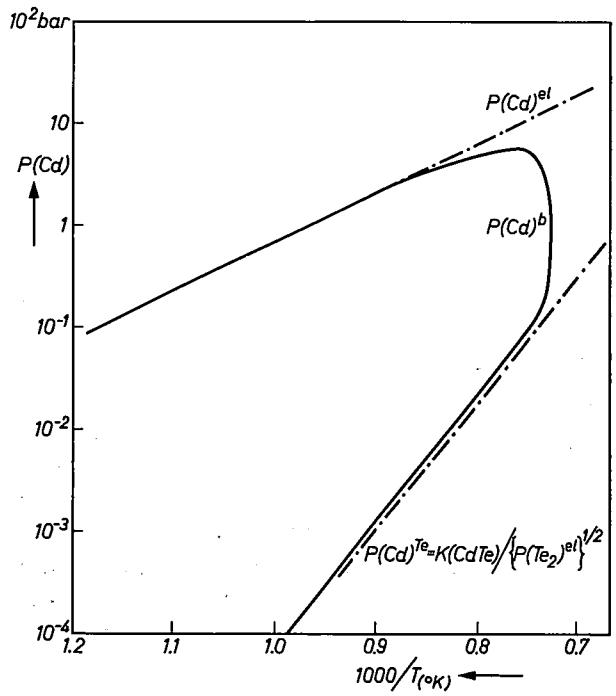


Fig. 17. The partial pressure of cadmium $P(\mathrm{Cd})^b$ as a function of T for CdTe at high temperature. The chain-dotted lines represent the saturation vapour pressure of elementary cadmium and elementary tellurium; the Te vapour pressure is converted to a cadmium vapour pressure with the aid of the formula given in the figure.

$N_c N_v \exp(-\Delta E_g/kT)$, where ΔE_g, the forbidden energy zone for CdTe, was found from optical measurements (1.5 eV at 300 °K). N_c and N_v can be calculated from (52) and (53) of part II if m^*(e) and m^*(h) are known. These have been determined from measurements of the thermoelectric e.m.f. and the Hall effect and are 0.14 m and 0.35 m respectively. K_{20}, K_{21} and K_{23} as a function of T are known from the appropriate ionization energies, which have been determined from the temperature dependence of the Hall effect. For example $K_{20} = 2N_c \exp(-0.02\ eV/kT)$, the factor 2 appearing because two electron spin orientations are possible in the point defect Cd_i^+ in the ground state, whereas only one configuration is possible for the two electrons present in the effectively neutral point defect Cd_i. The calculation of $K_{19} = K_{19}^0 \exp(-\Delta H/RT)$ goes as follows. At high pressures of P(Cd) the reduced electroneutrality relation $[Cd_i^+] = n$ applies (see the right-hand side of fig. 14). Substituting this into equation (70) gives $K_{19} = n^2/P(Cd)$. By determining the electron concentration (for example from Hall effect measurements) at two different temperatures and at a high known cadmium pressure, we can now obtain K_{19}^0 and ΔH, and hence K_{19} as a function of T. In principle, $K_{24}(T)$ can be determined in the same way from the relation $[V_{Cd}^-]p = K_{24}P(Te_2)^{1/2}$, which then gives $K_{22} = \{K_{24}K_{19}K(CdTe)\}/K_i$. The isotherms (concentration against partial pressure) of fig. 14 have been drawn from these data, which have also been used to produce *fig. 19*, which represents the central part of the T-x phase diagram (projection) of the system Cd-Te, but without the gas curve. Fig. 19 shows the complete existence region of CdTe between 0 °K and T_m (the maximum melting point) giving the concentration of Te in at.% on the horizontal axis in the usual way. CdTe of the exact stoichiometric composition begins to melt at a temperature which is 30 °K lower than T_m. Above 1000 °K the existence region is very definitely on the Te side, while below 1000 °K it is on the Cd side. The maximum melting point is on the Te side with about 2×10^{17} extra Te atoms per cm³; this implies a deviation from the stoichiometric composition of the order of 10^{-3} at.%. This composition is about halfway between the extreme boundary values of the existence region, being 49.99985 at.% Te at 1170 °K and 50.00085 at.% Te at 1350 °K.

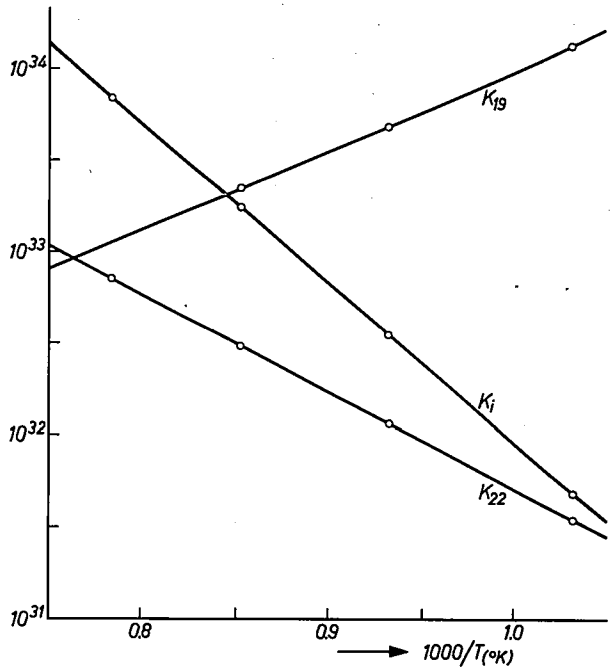


Fig. 18. A number of equilibrium constants as a function of temperature.

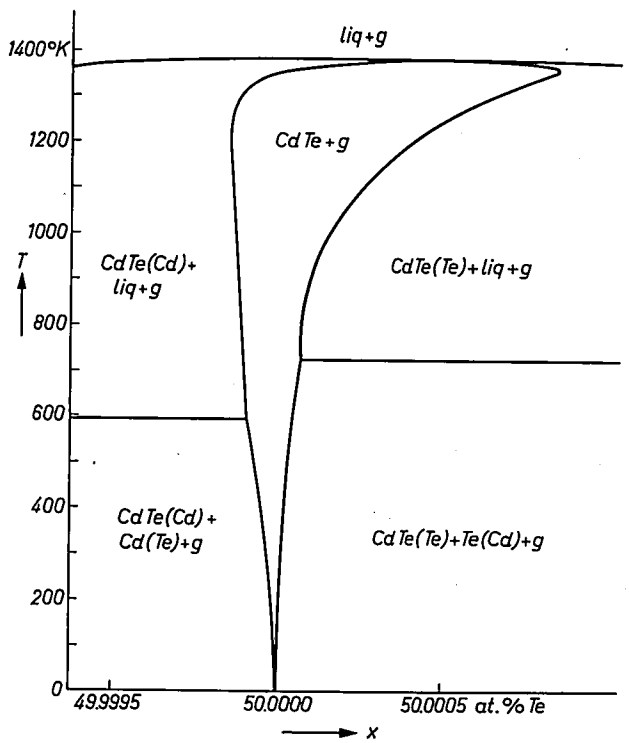


Fig. 19. The existence region (projection) of CdTe.

Fig. 20 shows the existence region of the semiconductor SnS [13], determined in a similar way, as an example of a compound for which the existence region lies entirely outside the stoichiometric composition at $T > 0$ °K.

It should be emphasized that the existence regions illustrated in figs. 19 and 20 are based on models indi-

[12] M. R. Lorenz, J. Phys. Chem. Solids 23, 939, 1962.
[13] H. Rau, J. Phys. Chem. Solids 27, 761, 1966.

cating the types of native point defect that occur in the crystal. Although the experimental data obtained so far are consistent with these models, the presence of other types of native point defect cannot be ruled out. There may, for example, be associates of neutral cadmium vacancies which are in equilibrium with the effectively charged cadmium vacancies V_{Cd}^{-} and V_{Cd}^{2-}, holes and electrons. The neutral double vacancy $(V_{Cd})_2$ could for

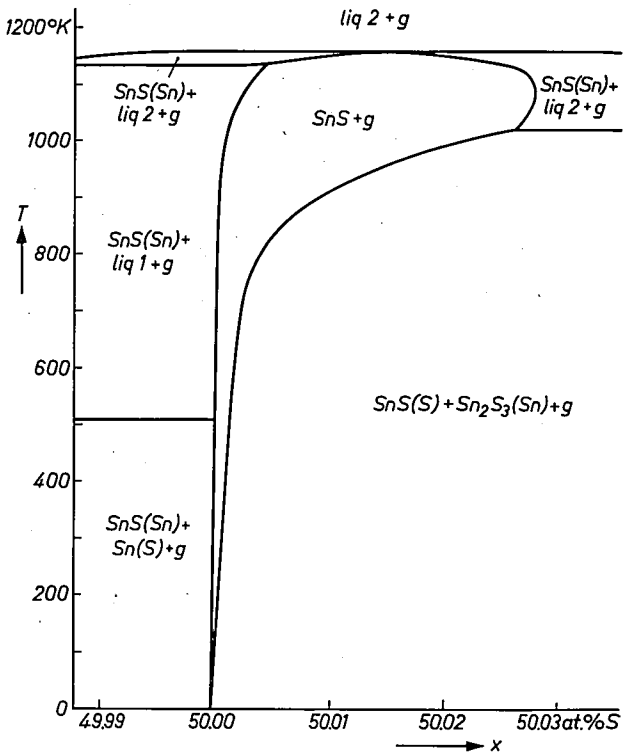


Fig. 20. The existence region (projection) of SnS. (After H. Rau [13].)

example be stabilized by covalent bonding between two neighbouring tellurium atoms (the general notation for these native point defects is $*Te_2$, see part II), corresponding to the fact that tellurium vapour contains Te_2 molecules as well as Te atoms. This is shown in schematic form in *fig. 21*. Multiple associates $(V_{Cd})_n$ might also occur. In this case the shape and position of the existence region would be different from that derived above. If, however, all the types of native point defects and the corresponding equilibrium constants are known, then the existence region can be accurately determined provided it is narrow.

Broad existence regions

The method discussed above is not suitable for determining broad existence regions because equations (31)-(33), (48), (49) etc. (from II) are not valid for large concentrations. However, the analysis of the coexisting

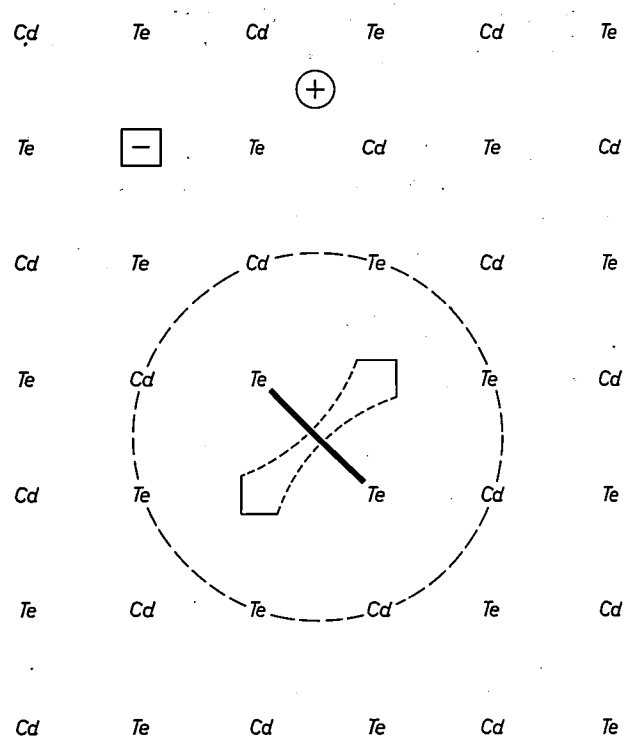


Fig. 21. Schematic diagram showing a CdTe crystal with atomic disorder. ⊕ is a free hole, ⊟ is a localized, effectively negatively charged cadmium vacancy V_{Cd}^{-}, and the ringed native point defect is an effectively neutral cadmium double vacancy $(V_{Cd})_2$. The thick oblique stroke symbolizes the covalent bonding between the two neighbouring tellurium atoms.

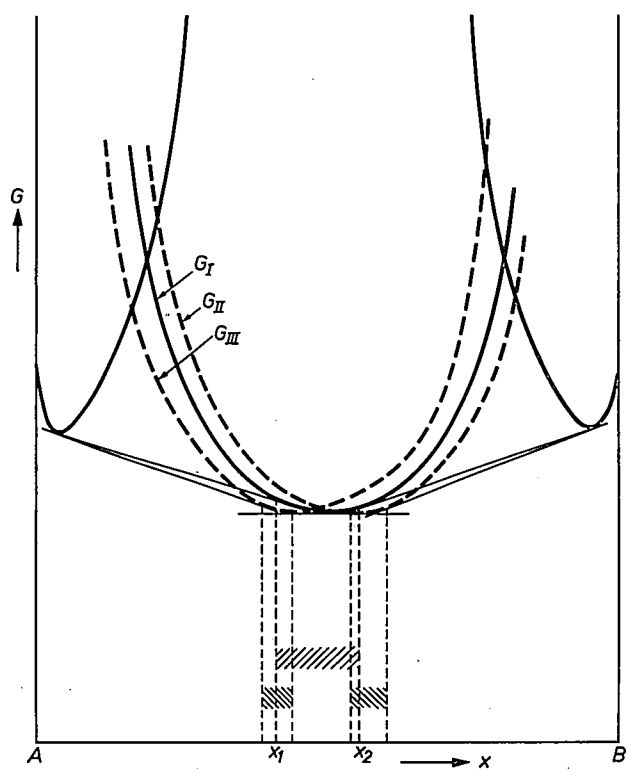


Fig. 22. Splitting up of a broad existence region into different existence regions due to the interaction of native point defects. G_I corresponds to the disordered structure, G_{II} and G_{III} to different types of ordering of the point defects.

gas phase is still of interest since the validity of equilibrium relations such as equations (60) and (61) in II s not affected.

i Information about the width of the existence region and the type of the dominant native point defects for compounds with broad existence regions can be obtained from chemical analysis, metal microscopy, and by determining the lattice parameters and the density as a function of temperature.

We shall conclude this article with a few remarks on the splitting up of a broad existence region into a number of narrower ones, which may occur at low temperatures. If AB contains a large number of dissolved native atoms, significant mutual interactions occur which, at certain concentrations (e.g. x_1 and x_2, see *fig. 22*), may lead to specific types of ordering (*II* and *III*) between the point defects. The occurrence of such superstructures is accompanied by a lowering of the Gibbs free energy ΔG_{super}, which means that below a certain temperature $T_0 = \Delta G_{\text{super}}/k$ the homogeneous region AB with randomly distributed dissolved native atoms (*I* in fig. 22) may be split up into a number of existence regions. This results in characteristic *T-x* diagrams of the type shown schematically in *fig. 23*. One system in which this takes place is the Cr-S system (see fig. 13), in which a series of stable phases of this kind can be indicated between the compositions CrS and Cr_2S_3 [14].

[14] F. Jellinek, Acta cryst. **10**, 620, 1957.

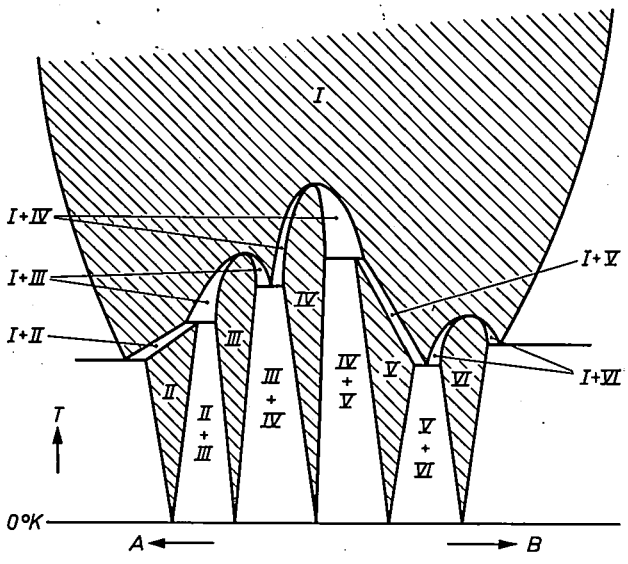


Fig. 23. Part of a *T-x* diagram showing the splitting up of a broad existence region into various smaller ones, which may occur at lower temperatures.

Summary. In the third and last part of this article on stoichiometry, methods of determining existence regions are discussed. It is explained why semiconductors and insulators usually have narrow existence regions while metallic compounds have broad ones. A calculation is given of the extremely narrow existence region of the semiconductor CdTe. Finally, the article briefly discusses the splitting up of a broad existence region into a number of narrower ones, which may occur at low temperatures.

Addendum to the article by A. L. Luiten, Superconducting magnets, Philips tech. Rev. **29**, *309-322, 1968 (No. 10).*

In this article, dealing with the technology of solenoids made from superconducting wire, a short historical introduction is given in which it is said that the first superconducting magnet producing a substantial field was made by Yntema in 1955. Professor E. Justi of Brunswick Technical University has now drawn our attention to a description of a superconducting magnet published by him as early as 1942. The magnet consisted of a ring of NbN, and from the load it could support, and other data it could be shown that the flux density was more than 15 000 gauss. The reader may care to refer to Proffessor Justi's article: E. Justi, Ein Dauerstrom-Elektromagnet, Elektrotechn. Z. **63**, 577-580, 1942.

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

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.

H. J. van den Berg: Interaction of bimolecular phospholipid leaflets with polyelectrolytes.
Adv. in Chem. Series No. 84, 99-103, 1968. *E*

H. Bosma, Th. Holtwijk & A. R. Miedema (Natuurkundig Laboratorium der Universiteit van Amsterdam): Ontwikkelingen en problemen in het toegepaste magnetisme.
Ned. T. Natuurk. **35**, 52-66, 1969 (No. 2). *E*

H. B. G. Casimir: In memoriam Prof. Dr. G. Holst.
Ned. T. Natuurk. **35**, 42-51, 1969 (No. 2). *E*

G. Diemer & J. H. Emck (Technical University, Eindhoven): Source Book of papers and statements prepared for the 1968 Eindhoven Seminar "The education of physicists for work in industry".
Publ. Centrex, Eindhoven 1969; 266 pp. *E*

J. A. W. van der Does de Bye: Quantum efficiency measurements using a 2π-geometry elliptic mirror.
Rev. sci. Instr. **40**, 320-321, 1969 (No. 2). *E*

B. B. van Iperen, W. Kuypers & H. Tjassens: A klystron frequency multiplier for submillimetre waves, generating its own driving power.
Nachrichtentechn. Fachber. **35**, 165-171, 1968. *E*

J. Jackson: Transient heat transfer and stability of superconducting composites.
Cryogenics **9**, 103-105, 1969 (No. 2). *M*

Y. Kamp: An equivalence between circuits with nonlinear inductances and nonlinear capacitances.
Archiv elektr. Übertr. **22**, 519-524, 1968 (No. 11). *B*

P. N. Kuin: The determination of nitrogen and boron in silicon carbide proton activation analysis.
Proc. 2nd Conf. on practical aspects of activation analysis with charged particles, Liège 1967, p. 31-42. *E*

H. O. Lincke (Nervenklinik der Universität München) **& W. J. Oosterkamp:** Angiographie und elektronische Farbgebung des Röntgenbildes.
Angiographie und ihre Leistungen, Tagung Berlin 1968, p. 76-78. *E*

F. K. Lotgering, R. P. van Stapele, G. H. A. M. van der Steen & J. S. van Wieringen: Magnetic properties, conductivity and ionic ordering in $Fe_{1-x}Cu_xCr_2S_4$.
J. Phys. Chem. Solids **30**, 799-804, 1969 (No. 4). *E*

J. Monin (Conservatoire National des Arts et Métiers, Paris), **J. Houdard & G.-A. Boutry:** Un nouvel ellipsomètre photoélectrique.
C.R. Acad. Sci. Paris **267B**, 1078-1080, 1968 (No. 20). *L*

A. Rabenau, H. Rau & G. Rosenstein: Chalkogenidhalogenide des Kupfers.
Naturwiss. **56**, 137-138, 1969 (No. 3). *A*

U. Schmidt & W. Thust: A 10-stage digital light beam deflector.
Opto-electronics **1**, 21-23, 1969 (No. 1). *H*

C. G. H. Scholten (Philips International Institute, Eindhoven): Elements for a new departure in air traffic control.
Thesis, Eindhoven 1969.

E. J. W. Verwey: Intermolecular forces in biochemistry.
Study Week on Molecular Forces, 1966 (Pontificiae Academiae Scientiarum Scripta Varia Vol. 31), p. 615-635; 1967. *E*

F. F. Westendorp & K. H. J. Buschow: Permanent magnets with energy products of 20 million gauss oersteds.
Solid State Comm. **7**, 639-640, 1969 (No. 8). *E*

E. E. Windsor: Construction and performance of practical field emitters from lanthanum hexaboride.
Proc. IEE **116**, 348-350, 1969 (No. 3). *M*

PHILIPS TECHNICAL REVIEW

VOLUME 30, 1969, No. 6/7

Accurate digital measurement of displacements by optical means

I. Displacement measurement with a reflection phase grating
II. Displacement measurement with a laser interferometer

In many developments taking place in the Philips Research Laboratories in Eindhoven and elsewhere in the world, the methods and devices employed for measuring displacements extremely accurately are of vital importance. For some years now these Laboratories have been making such measurements with an instrument comprising an optical grating (described in part I by De Lang, Ferguson and Schoenaker) which readily gives an accuracy of fractions of a micron and can even detect displacements of fractions of a nanometre. This kind of accuracy can also be achieved for displacements as large as several metres, by using interferometer methods with a laser as the light source. This was demonstrated with an equipment shown some months ago at the Physics Exhibition in London, and which is described in part II by De Lang and Bouwhuis. In both cases diffraction or polarization effects are used to subdivide the intrinsic digital unit of the measurement (given by the grating period and the wavelength respectively), while electronic methods are applied for supplementary linear interpolation.

I. Displacement measurement with a reflection phase grating

H. de Lang, E. T. Ferguson and G. C. M. Schoenaker

The problem of measuring linear displacements is encountered in engineering in a wide variety of situations, and the requirements for the accuracy of such measurements are becoming more exacting as technology advances. Dimensional tolerances of the order of a few microns are no longer unusual in workpieces with dimensions in decimetres. The precision lathe described in the previous number of this journal [1] and the recently described step-and-repeat camera for making photomasks [2] both demand very much smaller tolerances, amounting to a few tenths of a micron. Modern engineering also has its requirements for the *form* in which the measurement data are made available. In the production processes of today visual observation is out

Dr. H. de Lang, Dr. Ir. E. T. Ferguson and Ir. G. C. M. Schoenaker are with Philips Research Laboratories, Eindhoven.

[1] H. J. J. Kraakman and J. G. C. de Gast, A precision lathe with hydrostatic bearings and drive, Philips tech. Rev. **30**, 117-133, 1969 (No. 5).

[2] F. T. Klostermann, A step-and-repeat camera for making photomasks for integrated circuits, Philips tech. Rev. **30**, 57-70, 1969 (No. 3).

of place; what is now needed is a rapid, objective recording of measurement data, preferably as electrical signals of digital form to facilitate further processing. Digital output is of particular importance for numerically controlled machine tools [3].

These requirements can be met if the element (a carriage, say) whose displacement is to be measured is provided with some type of periodic structure, such as a grating, whose displacement generates an electrical signal of corresponding periodicity in a fixed scanning head. A measure of the displacement is obtained by electronically counting the number of periods of this signal that have elapsed. The measure is thus expressed in digital units which have the magnitude of the grating period.

Various kinds of grating can be employed. Systems using photoelectric scanning of a periodic *optical* structure, an optical grating, as proposed years ago [4], are particularly attractive. Such systems have been produced in the meantime in a variety of forms.

Let us first consider the simplest case, the scanning of an *absorption grating*, consisting of alternate opaque and transparent lines. If the grating period is large compared with the wavelength of the light, e.g. 0.1 mm, the effect of the grating can be completely described in terms of geometrical optics. *Fig. 1* shows an arrangement for measuring displacement by means of two such fairly coarse gratings. When one of the gratings moves

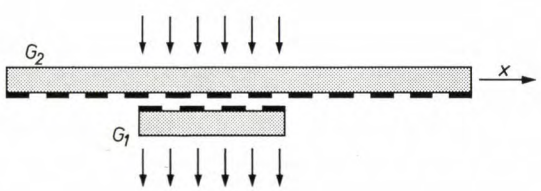


Fig. 1. Displacement measurement with two (not very fine) absorption gratings. The openings in the fixed reference grating G_1 are alternately covered and uncovered by the movement of grating G_2 in the x-direction.

in the x-direction, the openings in the stationary grating are alternately covered and uncovered so that the intensity of the transmitted light varies periodically. The period of the photoelectric signal obtained in this way corresponds to a displacement of one grating period, in this case 100 μm, and in order to measure accurately to within 0.5 μm interpolation down to 1/200 of the signal period is necessary. Interpolation of this order is possible (provided the grating, the guides etc. are sufficiently accurate), but it calls for fairly complicated electronic signal processing. It is therefore a more attractive proposition to use a *very fine* grating. For spectroscopic work diffraction gratings have long been made with a period of 1 μm or even less, and there are therefore no particular problems in making gratings with a period that could typically have a value of 8 μm like the ones used in the method of measurement we shall shortly describe. The operation of such fine gratings can no longer be described in terms of geometrical optics since it is now essential to take the diffraction effects into account. However, this is not a complication, since in fact in the method that we shall describe the diffraction is turned to good use. This method has other interesting features: the step of the digital measurement is 1/16th of the grating period, and the photoelectric information, relating both to the magnitude of the displacement and to its direction (forward or backward), is available in the form of an a.c. signal that can readily be processed and does not disappear when the movable grating comes to rest. This measuring system is used with success in the precision lathe [1] and the step-and-repeat camera [2] mentioned earlier.

Before describing the operation of our scanning system and the other equipment, we shall take a closer look at the special characteristics of optical gratings.

Scanning of an optical grating

When a light wave passes through a medium its amplitude and its phase are affected (by absorption and the difference in velocity of propagation respectively). The transmission coefficient of the medium, which accounts for both these effects, is written as $D = m \exp(-jn)$, and the action of a grating like the one in fig. 1 is thus completely described by a transmission function $D(x) = m(x) \exp\{-jn(x)\}$, where $m(x)$ and $n(x)$ are real periodic functions. In the case of the simple absorption grating mentioned above only m varies with x; the opposite case where only n varies with x (the phase grating) has also long been used in optics, and a grating of this kind is used in our system. The phase shifts which the light undergoes in passing through two phase gratings moving with respect to one another can be detected and put to use in the measurement by means of interference.

If the phase grating has a *short* period, the effect of *diffraction* becomes of primary importance, just as in an absorption grating. It will be useful first to give a brief quantitative description of this effect (the description is approximate but quite adequate for our purposes). Let us consider as a simple case a grating with a thickness that varies sinusoidally with the coordinate x with a period p; see *fig. 2*. A plane monochromatic light wave of wavelength λ and angular frequency ω is incident on the grating. The field strength E of the

wave in a plane V parallel to the plane of the grating has a phase which is independent of x:

$$E_V = A \sin \omega t.$$

After passing through the grating the wave front is no longer flat; the field strength in the plane W is:

$$E_W = A \sin (\omega t + b \sin 2\pi x/p). \quad . \ . \ . \quad (1)$$

Here $2b$ is the phase difference between light that has passed through a "trough" and light that has passed

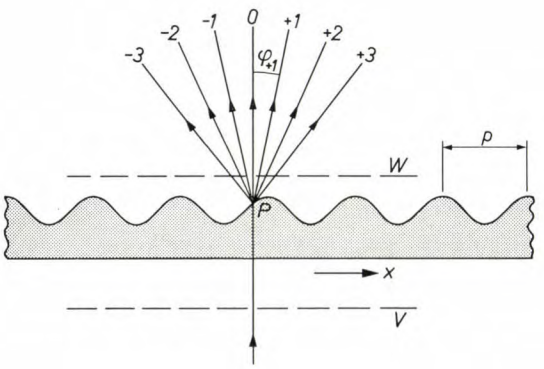


Fig. 2. Diffraction of the light (incident plane wave) in a phase grating. Here the phase grating is assumed to be a plate whose thickness varies sinusoidally; grating period p. The transmitted light has intensity maxima at certain diffraction angles φ_k.

through a "peak" of the grating surface; b is the parameter that characterizes the (sinusoidal) grating.

We can express E_W as a series:

$$\begin{aligned} E_W(x;t) &= Af_0 \sin \omega t + \\ &+ Af_{-1} \sin (\omega t + 2\pi x/p) + Af_{+1} \sin (\omega t - 2\pi x/p) + \\ &+ Af_{-2} \sin (\omega t + 2 \times 2\pi x/p) + Af_{+2} \sin (\omega t - 2 \times 2\pi x/p) + \\ &+ \ldots \\ &+ Af_{-k} \sin (\omega t + k \times 2\pi x/p) + Af_{+k} \sin (\omega t - k \times 2\pi x/p) + \\ &+ \ldots \ . \end{aligned} \quad . \ . \ . \quad (2)$$

The terms of this series can be interpreted as so many different plane waves, since at the plane W the phase of each term varies linearly with x. The direction of propagation of each wave can be derived from this variation; only the first term has the original direction of propagation, and represents the linearly transmitted wave. The whole series in fact represents the familiar effect of successive "diffraction maxima", and it can be shown from equation (2) that the diffraction angle φ_k for the wave (beam) of the order k is given by the condition:

$$\sin \varphi_k = k\lambda/p, \ \text{where} \ |k| < p/\lambda.$$

The relative intensity of the diffracted beams (with respect to A^2, the intensity of the incident beam) is f_k^2; in our case of a sinusoidal grating the coefficients f_k are related to the parameter b of the grating by the Bessel functions:

$$f_k = f_{-k} = \mathrm{J}_k(b). \quad . \ . \ . \ . \ . \ . \quad (3)$$

After this brief description of the diffraction effect, we must now ask what happens to the phase of the diffracted beams when the grating is displaced through a distance ξ in the x-direction. The reference point P (on the axis of the incident light beam) is kept stationary. In eq. (2) we must now substitute $x - \xi$ for x. This means that the phase of the beam of order (say) -1 has decreased by $2\pi\xi/p$, while the phase of the beam of order $+1$ has increased by the same amount:

$$\left.\begin{aligned} E_{-1} &= Af_{-1} \sin\left(\omega t + \frac{2\pi x}{p} - \frac{2\pi\xi}{p}\right), \\ E_{+1} &= Af_{+1} \sin\left(\omega t - \frac{2\pi x}{p} + \frac{2\pi\xi}{p}\right). \end{aligned}\right\} \quad . \ . \quad (4)$$

The resultant phase difference $4\pi\xi/p$ between the two beams, which can be detected if the beams are caused to interfere with each other, is a measure of the displacement. For the following it is also important to note that the phase difference between the two beams (-1) and $(+1)$ varies by 4π, i.e. by *two* full periods, if the grating is displaced through a distance equal to one period p (with respect to the fixed reference point P). Another essential feature is that the phase difference is independent of the wavelength of the light used.

To round off our introductory remarks we ought to mention that eq. (3) shows that the beams $(+1)$ and (-1) have the intensity $\mathrm{J}_1^2(b)$. This has a maximum when the grating parameter $b \approx 1.8$, the value of the maximum being 0.34; the zero-order beam then has an intensity of only 0.11. All these numerical values apply for a grating whose thickness varies sinusoidally. Eq. (2) is valid for any given grating structure of the same period p, but the coefficients $f_{\pm k}$ are different. In practice we use phase gratings with a "square-wave" variation in thickness. These are easier to make and also allow a higher maximum intensity to be obtained for the two first-order beams, while the zero-order beam has exactly zero intensity.

Besides the phase gratings for transmitted light considered above, there are also *reflection* phase gratings, with completely analogous characteristics. This is the type used in the arrangement for our measuring system,

[3] See for example T. J. Viersma, Some considerations on the numerical control of machine tools, Philips tech. Rev. **24**, 171-179, 1962/63.
R. Ch. van Ommering and G. C. M. Schoenaker, The COBRA, a small digital computer for numerical control of machine tools, Philips tech. Rev. **27**, 285-297, 1966.

[4] J. Guild, The interference systems of crossed diffraction gratings: theory of Moiré fringes, Clarendon Press, Oxford 1956.
J. Guild, Diffraction gratings as measuring scales, Oxford University Press 1960.

but for the purposes of the next section we shall continue to consider the transmission grating.

Scanning with an optical imaging system

We have seen that when the phase grating is displaced through one grating period, a certain phase difference (in our case 4π) is caused between two emergent diffracted beams of light. We have also mentioned that the phase difference, and hence the displacement, can be determined by means of interference between the beams. How is this interference brought about?

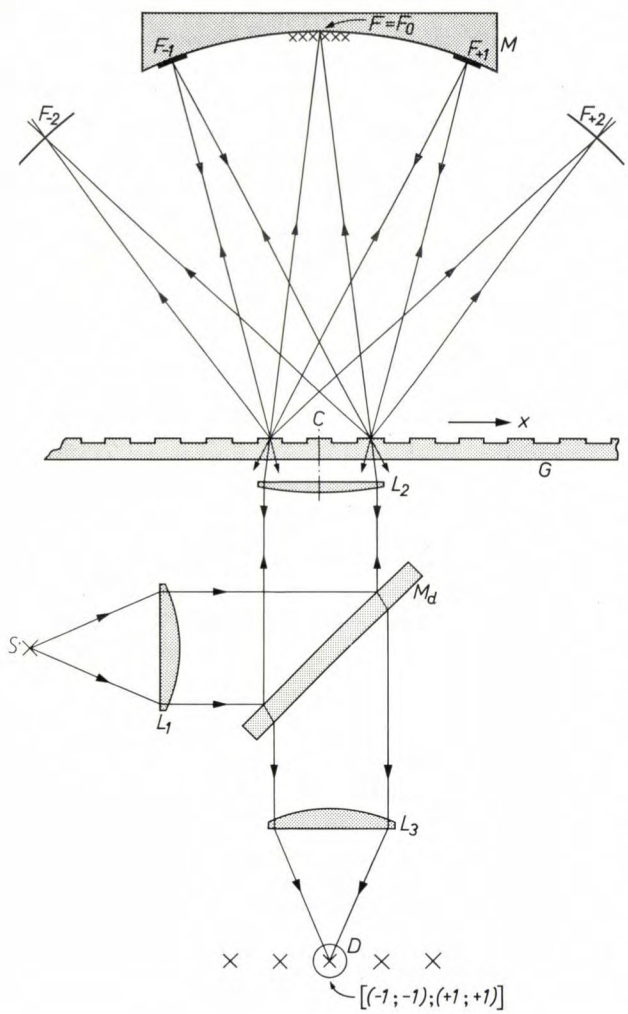


Fig. 3. Principle of the scanning of a phase grating G by means of a concave mirror M which forms an image of the grating on the grating itself. S light source. M_d semi-transparent mirror. D photodetector.

Unlike the first method, in which the displacement was detected with the aid of a second, stationary grating, the method now to be described requires only one grating, since *an image of the grating is formed on the grating itself* by means of an optical system. We shall illustrate this with reference to the diagram in *fig. 3*. For simplicity we take the image-forming system here to consist of a spherical concave mirror M, whose centre of curvature C lies in the plane of the grating G. The light from an incandescent lamp S is collimated by a lens and projected on to the grating via a semi-transparent mirror M_d. Placed near the grating is a lens L_2 whose focal point F lies on the concave mirror. Owing to diffraction from the grating, however, the light does not only strike the mirror at F but also at a number of other points: we can thus speak of the focal points F_0, F_{-1}, F_{+1}, etc., corresponding to the diffraction maxima of orders 0, —1, +1, etc. The mirror is made small enough for all the higher orders to fall outside it, and it is blackened at the position F_0. Thus, only the light of orders —1 and +1 is reflected back to the grating. Here again, diffraction takes place in beams of the different orders. The lens L_2 collimates these beams and projects them on to the semi-transparent mirror M_d. The light which this transmits is projected by a lens L_3 on to a photodetector D, at which an image of the incandescent lamp S is thus produced in a series of orders. The detector picks up only the central image, formed by contributions from the beam of order —1 delivered by the light reflected at F_{-1} and the beam of order +1 originating from the light of F_{+1}. We thus have interference between the beams (—1;—1) and (+1;+1). If A is again the amplitude of the incident light, we can calculate the intensity I of the resultant emergent beam as follows.

Let us first consider the diffraction of order —1 and assume that the grating has undergone a displacement ξ. The emergent diffracted beam is given by (see eq. 4):

$$Af_{-1}\sin\left(\omega t+\frac{2\pi x}{p}-\frac{2\pi\xi}{p}\right)=Af_{-1}\sin\left[\omega t+\frac{2\pi(x-\xi)}{p}\right].$$

This strikes the mirror, and the beam reflected back to the grating is given by:

$$Af_{-1}\sin\left(\omega t-\frac{2\pi x}{p}-\frac{2\pi\xi}{p}\right)=Af_{-1}\sin\left[\omega t+\frac{2\pi(-x-\xi)}{p}\right].$$

(Because of the action of lens L_2 the phase also contains a quadratic term in x, but this can be neglected here.) This beam is again diffracted by the grating, whence:

$$E_{(-1;-1)}=Af_{-1}^2\sin\left[\omega t+\frac{2\pi}{p}(-x-\xi)+\frac{2\pi}{p}(x-\xi)\right]$$

$$=Af_{-1}^2\sin\left[\omega t-2\frac{2\pi}{p}\xi\right]. \qquad \ldots\ (5a)$$

Similarly we find for the other beam reaching the exit aperture:

$$E_{(+1;+1)} = Af_{+1}^2 \sin\left[\omega t + 2\frac{2\pi}{p}\xi\right]. \quad \ldots \quad (5b)$$

If the grating profile is symmetrical, we have $f_{-1} = f_{+1} = f_1$, so that from eqs. (5a) and (5b):

$$E_{(-1;-1)} + E_{(+1;+1)} = 2Af_1^2 \sin\omega t \cos\left(2\frac{2\pi}{p}\xi\right), \quad (6)$$

and therefore the intensity is:

$$I(\xi) \propto A^2 f_1^4 \left[1 + \cos\left(4\frac{2\pi}{p}\xi\right)\right]. \quad \ldots \quad (7)$$

Thus, when the grating is displaced through a distance ξ equal to one period p, the optical signal goes through *four* full periods. This factor of four reducing the basic unit of measurement is made up of two factors of two, one due to the superposition of the diffraction maxima of order +1 and —1, as we have already seen in the previous section, and the other arising because the light is diffracted by the grating *twice*, once on the way there and once on the way back.

For comparison it is useful to note that earlier measuring systems with phase gratings made use of the interference of the two beams of order (0;0) and (1;1)[5]. The reduction in the basic unit, which allows the same accuracy to be obtained with a coarser grating, amounts to a factor of only 2 in this case. Apart from this our system has a number of other advantages, most of which stem from the use of the interference of *symmetrically diffracted beams*. With a symmetrical grating profile the beams (+1;+1) and (—1;—1) always have the same amplitude, irrespective of the wavelength of the light, and therefore the photosignal obtained is fully modulated. Next, the symmetry rules out systematic errors of measurement that might arise from slight changes in the position of the grating in a direction perpendicular to its surface. The phase difference of the interfering beams, which varies with the displacement of the grating in the x-direction, is also independent of wavelength, as was stated in the preceding section; i.e. the system works as an *achromatic interferometer*. Finally, the direction of the emergent beams is independent of wavelength, because the dispersions of the diffractions on the outward and return paths exactly compensate one another.

There is in fact some dispersion in the position of the focal points of the concave mirror. This means that when light with a wide spread of wavelengths is used the focal points are extended into spectra. This would appear to have no significance since the dispersion is compensated when the beams emerge. There is however the residual effect that the spread of the foci usually determines the tolerance for displacements of the grating in a direction perpendicular to its surface (the "depth of focus"). In fact, the modulation depth of the signal is found to decrease if the grating surface does not pass exactly through the centre of curvature C. Analysis shows that this loss of modulation depth increases not only with the angular aperture of the incident beam (determined by the dimensions of the light source S with respect to the focal length of the lens L_1), but also with the dispersion in the diffraction angle. In our particular case of symmetrically diffracted beams, however, even this second effect does not occur.

In concluding this paragraph we should stress two features of our scanning system. The first is that — as is generally the case with scanning using an image-forming system — the action of the mirror can be regarded as the formation of a *reversed image* of the grating surface on that same surface (or, to be more exact, of the second harmonic of the grating structure, since only the diffraction maxima of order +1 and —1 are used). Since the image is reversed, rotation of the grating in its own plane about the point C will not have any effect on the photosignal. This is of great importance, since the guide in which the grating runs is never absolutely perfect and the grooves of the grating are in practice never exactly parallel with one another over the full length of the grating ("fan error"). These imperfections do no harm because of the insensitivity of the reversed-image scanning system to rotation of the grating in its own plane. Incidentally, it is a general feature of the grating methods usually used that slight *damage* to the grating (perhaps with the complete loss of a few lines) does not result in complete loss of the signal.

The second special feature of our scanning method is that it is very easy to use a *reflection grating* instead of a transmission grating. The complete optical system can then be kept on one side of the grating, which offers considerable advantages. This feature is used in the practical design, which we shall now describe.

Practical design of a scanning head

Fig. 4 shows the final version of a scanning head for our measuring system; a diagram of the head is shown in *fig. 5*. Fixed to the carriage whose displacement is to be measured is a reflection grating with a grating period of 8 μm. The grating consists of a glass substrate on which an aluminium film has been deposited by vacuum evaporation; the grating lines, which have a square-wave profile, are photo-etched into the film. The etch depth is 0.37 μm. To prevent damage to the aluminium film a glass cover plate (about 1 mm thick) is cemented to it.

[5] H. de Lang, Dutch Patent No. 112407, applied for 21st January 1958.
D. A. Palmer, J. sci. Instr. **37**, 261, 1960.

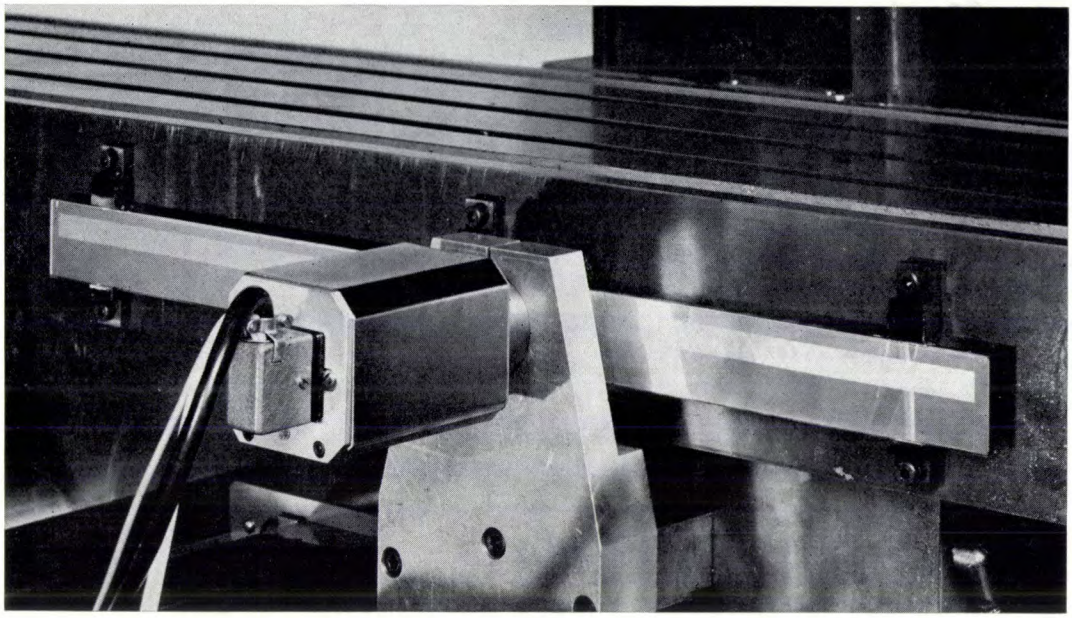
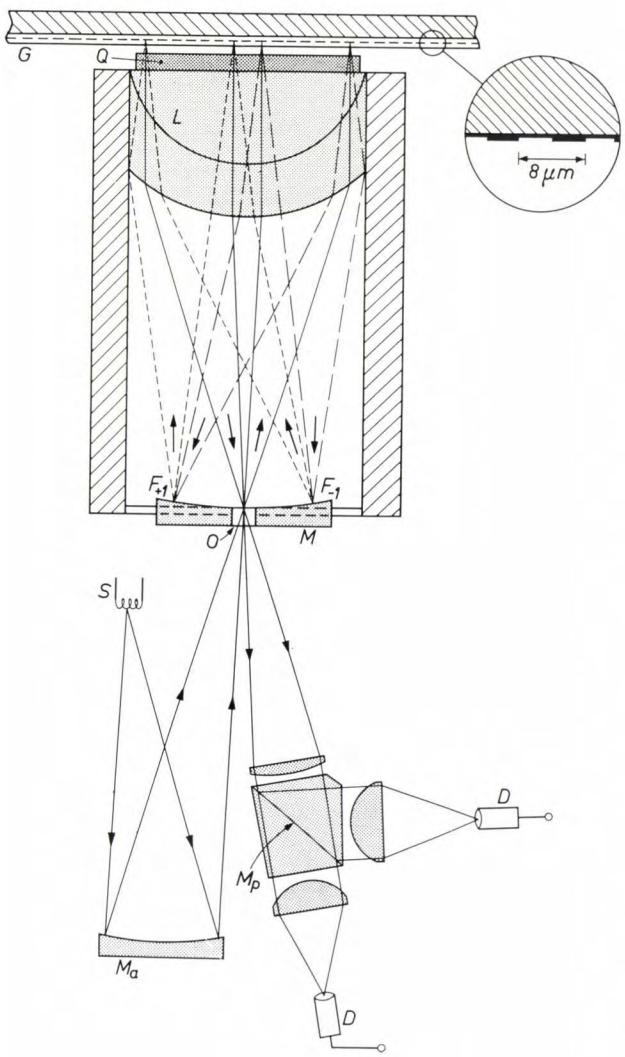


Fig. 4. Scanning head and grating of the measuring system in its final form. The system is shown here mounted in a numerically controlled milling machine in one of the Philips works.



The image-forming system consists of a concave mirror, which has a small hole in the middle of it, combined with a plano-convex correction element consisting of two component lenses, whose spherical surfaces are concentric with the mirror. A small mirror condenser focuses the light from an incandescent lamp on to the hole in the concave mirror, and this light passes through the correction element and strikes the part of the grating in the right-hand half of the visual field of the mirror, as shown in fig. 5. The light diffracted and reflected by the grating arrives at the mirror at the focal points F_{-1} and F_{+1}. The mirror reflects this light back to the grating, but now to the left-hand half of its visual field (the image of the right-hand half of

Fig. 5. Schematic diagram of the scanning system in its final form. The grating G, fixed to the moving object, is a reflection grating, which means that all the optical elements can be situated on one side of it. S light source. M_a condenser mirror replacing the lens L_1 in fig. 3. The light enters the actual scanning system through an aperture O in the main mirror M. This mirror forms an image of a part of the grating on another part via a two-component correction lens L, and the doubly diffracted light again emerges through the aperture O. Owing to the difference between the directions in which the light enters and leaves the system there is now no need for the semi-transparent mirror in fig. 3. However, the light is split by the quartz plate Q into two components polarized at right angles to one another, and these two components are now separated by a polarizing beam splitter M_p and separately detected by the two photodetectors D. This enables a discrimination to be made between forward and reverse movements of the grating. By making the mirror M oscillate about an axis parallel to the grating lines at a fairly high frequency the electronic processing of the photosignals is considerably simplified.

the grating thus appears on the left-hand half). Here the light is for the second time diffracted and reflected by the grating, so that the beams (—1;—1) and (+1;+1) pass through the hole in the mirror to the photodetector system, which observes the interference.

This scanning head has essentially the same characteristics as the optical system in fig. 3. It contains some extra elements, however, by means of which a) forward and reverse displacements of the grating can be distinguished, b) a signal is obtained even when the grating is stationary, and c) there is digital interpolation down to a basic unit of 1/16th of a grating period, i.e. down to 0.5 μm. We shall now briefly discuss how this is achieved.

Between the scanning head and the grating there is a plate of crystalline quartz Q. The optical axis of the quartz crystal is tilted through 45° with respect to the surface of the plate, which has the result that the image formed by the optical image-forming system of the one part of the grating on the other part is separated into two components with polarizations at right angles to one another. Because of the bi-refringence of the quartz these two components are displaced slightly with respect to one another, and by giving the quartz plate (through which the light travels twice) the appropriate thickness and orientation it can be arranged that the displacement in the direction of measurement is 1/8th of the grating period p. This means that the phase difference between the two emergent interfering diffraction beams (—1;—1) and (+1;+1) differs for the one direction of polarization by an amount $\pi/2$ from the phase difference for the other direction of polarization. (Since a relative displacement of the images by the period of the second harmonic of the grating structure, i.e. $\frac{1}{2}p$, would correspond to a phase difference of exactly 2π.) The two components with these directions of polarization are then separated by a polarizing beam splitter M_p [6] and the two separated beams are applied to two separate photodetectors, giving *two* photosignals s_1 and s_2, with a phase difference of $\pi/2$:

$$s_1 = K_1 + L_1 \sin (8\pi\xi/p), \quad . \ . \ . \quad (8a)$$

$$s_2 = K_2 + L_2 \cos (8\pi\xi/p), \quad . \ . \ . \quad (8b)$$

where K_1, K_2, L_1 and L_2 are constants.

Once these two signals have been obtained it is in principle possible to measure the displacement of the grating in digital steps of $p/16$, and also, as we shall see, to distinguish between forward and reverse displacements. However, we make use of a little trick that greatly simplifies the actual measurements: the concave mirror of the image-forming system is made to oscillate about an axis parallel to the grooves of the grating, so that the image of the grating carries out a simple harmonic oscillation in the direction of displacement. Without the oscillation the two photosignals would be at a steady value when the grating is stationary, and could therefore only be processed by the cumbersome and in our case barely adequate methods of d.c. amplification. The oscillatory movement modulates the phase of the signals s_1 and s_2 sinusoidally, and if Ω is the angular velocity and mp the amplitude of the image oscillation we obtain the following expressions:

$$S_1 = K_1 + L_1 \sin \left[8\pi\left(\frac{\xi}{p} + m \sin \Omega t\right)\right], \quad . \quad (9a)$$

$$S_2 = K_2 + L_2 \cos \left[8\pi\left(\frac{\xi}{p} + m \sin \Omega t\right)\right]. \quad . \quad (9b)$$

Thus, even when ξ is constant, we still have a.c. signals that can be amplified in the normal way. After amplification of the signal the phase can be measured digitally in units of $\pi/2$ by counting the number of times the signal passes through zero (this is dealt with in the next section).

We should say something more about the optics of the measuring head, and in particular about the correction element used in it (L in fig. 5). This has various functions. In the first place it makes the ray path at the location of the grating telecentric, i.e. the external principal point of the system goes to infinity. This makes the scanning system insensitive to any tilting of the grating about an axis in the plane of the grating. Secondly — and this function is even more important — the correction element removes the aberrations over a wide field of the concave mirror. It can easily be shown that in forming an image of the grating surface upon itself the aberrations of the mirror give rise to errors in the phase relation between the interfering beams, and that they therefore reduce the signal modulation due to the displacement of the grating. The aberrations, expressed in optical path lengths, should therefore not be more than a fraction of the wavelength of the light used. A very good image can be obtained over a wide field of view with the aid of a single correction lens alone [7]. However, to meet the very strict requirements imposed here, it was useful to make the correcting element from two lenses of different types of glass cemented together, as shown in fig. 5. This construction also has the great advantage that the correction of the image of the light source formed on the photodetector is obtained in a very easy way.

Before going into the digital processing of the signals S_1 and S_2 (eq. 9) it is perhaps interesting to mention the exceptionally sensitive detection of displacements that can be obtained with these signals. If ξ in eq. (9) is a whole multiple of $\frac{1}{8}p$, then the oscillogram of the phase-modulated signal S_2, for example, is a periodic curve consisting only of even harmonics. With a minute displacement of the grating odd harmonics appear in the frequency spectrum of S_2 (changing in sign according to whether the displacement is to the right or to the left), and this gives rise to an alternating peak height.

[6] See note [6] of the article II by H. de Lang and G. Bouwhuis, this issue, p. 165.

[7] J. Dyson, J. Opt. Soc. Amer. 49, 713, 1959.

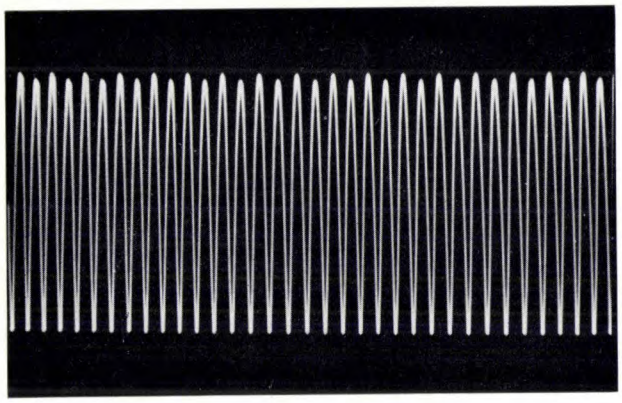


Fig. 6. A record of the phase-modulated signal S_2 (eq. 9) illustrating the possibility of detecting displacements of fractions of a nanometre. In the case shown here the grating displacement deviated by 5 nanometres from an integral multiple of $\frac{1}{8}p$.

Fig. 6 shows the result of a displacement through 5 nanometres. With appropriate electronic aids, it would be possible in this way to detect displacements as minute as hundredths of a nanometre (fractions of an ångstrom) and thus to maintain a particular average position with extraordinarily high accuracy.

Electronic signal processing

The way in which our measuring system derives the required information from the photoelectric signals will now be briefly considered, without going into details about the logic circuits, the design of the amplifiers, etc. The processing can be divided into two steps: digital indication and, where it is required, analogue interpolation.

Digital measurement

We shall first describe the more or less conventional method by which the digital information on the displacement and the information on the displacement direction can be obtained from the signals s_1 and s_2 of eq. (8), originating from the grating interferometer system when the mirror does *not* oscillate.

First of all, the constants K_1 and K_2 have to be eliminated from the signals by means of a compensation method using two extra photoelectric cells and two difference amplifiers. After amplification and peak-clipping, which eliminates the effect of the constants L_1 and L_2 by making the difference between them imperceptible, two signals s_{1c} and s_{2c} are obtained which vary with time in the way shown in *fig. 7a* when the grating is displaced uniformly. These signals are differentiated in a separate circuit, giving two signals s_{1d} and s_{2d} consisting of alternate positive and negative pulses; see fig. 7*b*. If the waveforms in figs. 7*a* and *b* refer to a displacement to the right, then the signal waveforms shown in *fig. 8a* and *b* will be obtained for a displacement to the left: for the signals s_{1c} and s_{2c} only the time axis has been reversed, but for the differentiated signals the ordinate has been reversed as well.

The pulses s_{1d} and s_{2d}, which appear every time the signals s_{1c} and s_{2c} go through zero, are now passed to a digital counter which will count a step forwards or backwards at every pulse. If we make the direction of counting depend in the appropriate way on the polarity of the signal s_{1c} or s_{2c} present at the instant when the counting pulse appears, we can arrange that the counter

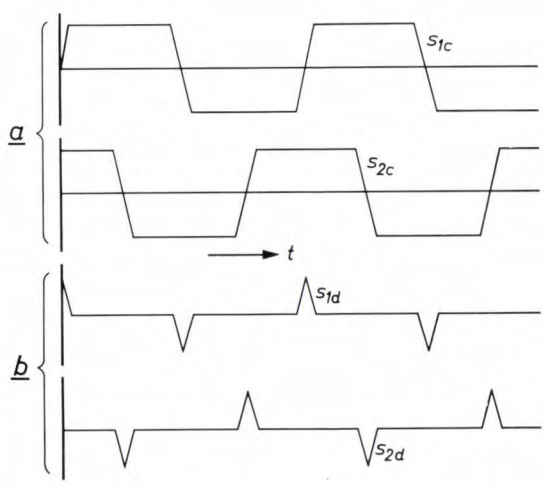


Fig. 7

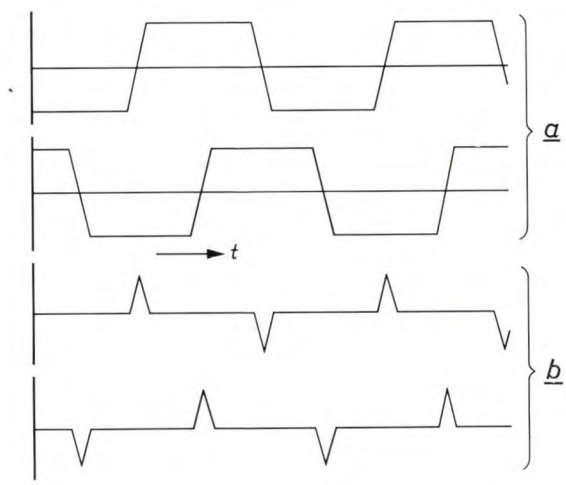


Fig. 8

Fig. 7. The signals obtained from the two photocells during a displacement, shown as a function of time. *a*) Signals after amplification and peak-clipping, *b*) after subsequent differentiation. A complete period of the signals corresponds to a displacement of $\frac{1}{4}p$.

Fig. 8. As fig. 7, for a displacement in the opposite direction.

always counts forwards in fig. 7 and always backwards in fig. 8. It can easily be verified that the following "truth table" is applicable:

	forwards	backwards
s_{1d} positive	s_{2c} positive	s_{2c} negative
s_{1d} negative	s_{2c} negative	s_{2c} positive
s_{2d} positive	s_{1c} negative	s_{1c} positive
s_{2d} negative	s_{1c} positive	s_{1c} negative

The decoding of the signals in accordance with this table is a simple matter using logic circuits. During one signal period, i.e. — as we have seen from the derivation of eq. (8) — during a displacement through one quarter of the grating period p, we obtain four counting pulses, as can be seen in fig. 7 (or 8). The counter thus indicates the displacement digitally in units of $p/16$, i.e. in units of 0.5 μm with the grating period that we have used.

This method of extracting the required information from the signals s_1 and s_2 would suffice if it were in fact possible to eliminate the constants K_1 and K_2 in the way described. To do this one would have to have identical photoelectric cells and d.c. amplifiers with absolutely no drift. In practice this is not feasible, problems arising in particular from the differences in the temperature dependence of the photoelectric cells. Because of these difficulties it was decided to adopt the system mentioned earlier in which the concave mirror of the scanning system is made to oscillate. This converts the signals into a.c. signals S_1 and S_2, as given by eq. (9), and the d.c. components K_1 and K_2 are automatically eliminated in the normal a.c. couplings in the circuits.

Signal processing in this case is somewhat different. In the first place, similar signals S_{1c}, S_{2c}, S_{1d}, S_{2d} are again derived by amplification, clipping and differentiation, and these again yield the required counting pulses upon decoding with the same truth table. What is new, however, is that the counter still receives pulses even when the grating is stationary. This is due to the oscillation of the mirror, a number of pulses for the forward direction being received when the mirror swings to one side, and an equal number of pulses for the reverse direction when it swings back again. (This is easily understood if one considers that a displacement of the grating and a rotation of the mirror are optically equivalent, neglecting second-order effects.) The amplitude of the oscillation of the mirror is now given a value such that when the grating is stationary, the train of measuring pulses has alternately seven forward pulses and seven reverse pulses; the repetition rate of these pulses is of course equal to $f = \Omega/2\pi$, the frequency of oscillation of the mirror (see eq. 9). In the block diagram in *fig. 9* we see that this train of pulses, processed in the manner described and decoded for direction, is not applied directly to the counter C but to an intermediate counter H, which has eight positions: —3, —2, —1, 0, 1, 2, 3, 4. The intermediate counter keeps the gates G_a and G_b to the main counter closed, except in positions 4 and —3: in the position 4 it opens G_a and in position —3 it opens G_b.

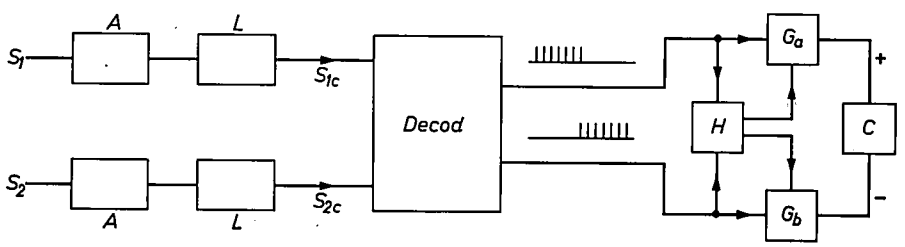


Fig. 9. Block diagram of the circuit for the digital processing of photodetector signals S_1 and S_2. *A* amplifiers. *L* limiters. In the circuit *Decod* the signals are differentiated, and they are decoded in accordance with the truth table given in the text; this means that a forward movement over a unit distance of $p/16$ produces a positive pulse, while such a movement in the reverse direction produces a negative pulse. Superimposed on this, as a result of the oscillation of the mirror, there continuously appear "forward" pulses and "reverse" pulses. These are counted in the intermediate counter H, which eventually opens the gate G_a or the gate G_b to let through the overflow pulses. Consequently, only the overflow pulses are counted in the main counter C and thus supply a direct digital indication of the displacement in units of $p/16$, with the proper direction.

When the grating is stationary, the intermediate counter continuously counts backwards and forwards between positions +4 and —3. The periodic opening of G_a has no effect (nor has the opening of G_b), because after position 4 has been reached the intermediate counter does not receive any positive pulse that could reach the main counter through G_a. The main counter is thus not actuated and it retains the latest information relating to the position of the grating.

If the grating now makes a movement "forwards", with a speed of say exactly $p/16$ per oscillation period $(2\pi/\Omega)$ of the mirror, then the intermediate counter does not receive 14 pulses in the time $2\pi/\Omega$, but 15, alternately 8 forwards and 7 backwards. Every time position 4 is reached there is thus one overflow pulse which is transferred through the opened gate G_a to the main counter — the intermediate counter H remains in position 4 until a reverse pulse arrives. In the case in question the main counter thus makes one step (equal to 1/16th of a grating period) in each interval $2\pi/\Omega$.

If the movement of the grating is slower, the intermediate counter H will make several complete runs backwards and forwards before an overflow pulse is transferred; if it is much faster, several overflow pulses are transferred for one complete run forwards and backwards of the intermediate counter. The maximum speed at which the main counter gives a reliable indication is *not* limited by the oscillation frequency of the mirror, as might perhaps be thought. If the movement of the grating is faster than the speed that would correspond to the apparent movement to and fro which is equivalent to the mirror oscillation, the intermediate counter continuously receives pulses in one direction, only more and fewer in alternate half-periods of the mirror, and *all* the pulses go to the main counter. The limit is therefore set only by the frequency characteristics of the photoelectric cells and circuits. At a grating speed of 50 cm/s the counting frequency required is 1 MHz, a value which is not difficult to obtain. To avoid disturbance by unwanted mechanical vibrations the mirror frequency is made fairly high, e.g. $\Omega = 10^4$ rad/s, i.e. 1500 oscillations per second.

To explain our choice of mirror amplitude, resulting in the alternate counting of seven steps forwards and backwards, we must take a closer look at eq. (9). Let us consider the term

$$\cos\,[8\pi(\xi/p + m \sin \Omega t)]$$

in eq. (9b). This has the same form as (1) and can be expanded in the following series (cf. eq. 2):

$$\left[J_0(8\pi m) + 2J_2(8\pi m)\cos 2\Omega t + \ldots\right]\cos\frac{8\pi\xi}{p} -$$

$$-\left[2J_1(8\pi m)\sin\Omega t + 2J_3(8\pi m)\sin 3\Omega t + \ldots\right]\sin\frac{8\pi\xi}{p}.$$

Here J_0, J_1, ... are Bessel functions. It can be seen from the series that when the grating is stationary (or moving slowly) the signal contains in addition to K_1 a d.c. component (or quasi-d.c. component) $J_0(8\pi m)\cos(8\pi\xi/p)$, which in principle is eliminated in the a.c. amplification. If there is an abrupt change of speed, however, this d.c. component may then lead to erroneous information processing as a result of switching transients in the a.c. amplifier. We avoid this by giving the mirror amplitude m a value for which $J_0(8\pi m)$ is zero. An infinite number of values of m satisfy this condition; the smallest is $m_1 = 2.43/8\pi$ rad $\approx$ 140°, and the next is $m_2 = 5.53/8\pi$ rad $\approx$ 317°. We have chosen $m = m_2$, at which about 7 counts are made forwards and 7 backwards per mirror period (there are of course not exactly 7, but this does not affect the function of the intermediate counter; this is easily understood by considering the case of a rapid displacement where the forward and backward counting is noticed only as a periodic speeding up and slowing down of the stream of overflow pulses to the main counter). The choice $m = m_1$ could also have been made, but it would have been less suitable for several reasons and particularly in view of the analogue interpolation (see below).

Apart from the term with Bessel function $J_0(8\pi m)$ the term $2\sin(8\pi\xi/p)\,J_1(8\pi m)\sin\Omega t$ can also theoretically cause unwanted effects. We may write this term as:

$$J_1(8\pi m)\left[\cos\left(\Omega t - \frac{8\pi\xi}{p}\right) - \cos\left(\Omega t + \frac{8\pi\xi}{p}\right)\right].$$

If a grating speed $\xi/t = p\Omega/8\pi$ should happen to occur, the first half of this expression takes the constant value $J_1(8\pi m)$, so that we again have a d.c. component. For $m = m_1$ this is about 1½ times larger than for $m = m_2$ (the J_1 values being 0.51 and −0.34 respectively). We were in fact able to demonstrate that irregularities due to this d.c. component can occur: in our case, with $p = 8$ μm and $\Omega = 10^4$ rad/s, this could happen at a speed of about 3 mm/s, but in practice it does not prove troublesome.

Analogue interpolation

In many cases the digital indication of a position in units of 0.5 μm will be sufficiently accurate. By interpolation of the digital basic units a further refinement is possible, and with the present-day quality of gratings a refinement by a factor of perhaps 10 is certainly meaningful. A method for this interpolation has been devised, which we shall describe with reference to *fig. 10*.

This figure shows how the phase $\varphi/8\pi$ of the signals S_1, S_2 varies as a function of time, due to the oscillation of the mirror, the grating remaining stationary. Owing to this phase variation the intermediate counter

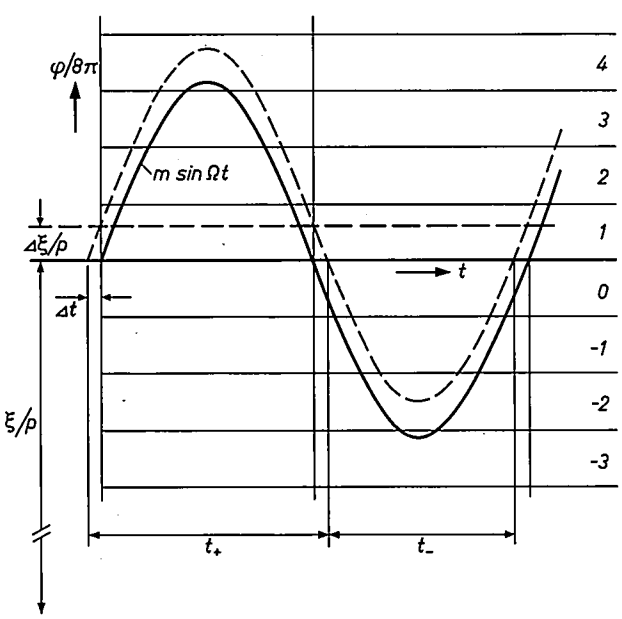


Fig. 10. Illustrating the method of interpolation.

goes through all its possible positions from —3 to +4; this is indicated in the figure by the division of the ordinate by the horizontal bands.

The solid curve relates to a "digital" grating position, i.e. a position that corresponds exactly to the indication of the main counter ($\xi = np/16$, n being a whole number, interpolated value zero). This curve is symmetrical with respect to the axis $\varphi/8\pi = \xi/p$, and the intermediate counter remains for the same length of time in the positions +1, +2, +3, +4 as it does in 0, —1, —2, —3. If the grating has been displaced by a distance $\Delta\xi$, then the periodic phase variation will assume a form like that of the dashed curve; the intermediate counter remains in the positions +1 ... +4 for a longer time than it does in the others. The intermediate counter is now given an individual output for each of the positions +1 ... +4 and all these outputs are connected to an OR circuit (*fig. 11*). The output

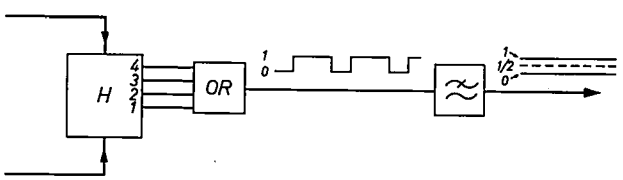


Fig. 11. Circuit used for the interpolation, i.e. for determining $\pm\Delta\xi$ situated between 0 and $\pm p/32$. The positions +1 ... +4 of the intermediate counter H in fig. 9 each have their own output; these outputs are connected to an OR circuit. The average value of the output voltage from this circuit is a measure of $\Delta\xi$.

signal of this circuit has a constant value of 1 when a signal is applied at one (or more) of the inputs, otherwise the output is zero. We thus obtain a rectangular waveform, which has the value 1 for the same length of time that it has the value 0 only when $\Delta\xi = 0$, and whose average value is a simple measure of the unknown $\Delta\xi$. This value can be determined by means of a low-pass filter; an ordinary moving-coil meter, which in itself acts as a filter, can be used to provide an indication, its scale being calibrated in such a way that the interpolated value can be read off directly.

The interpolation is not absolutely exact, but the error is extremely small. It may be calculated as follows.

The average value of the rectangular-wave voltage, which is determined by means of the circuit shown in fig. 11, is in fact a measure of the difference between the time t_+, during which the phase $\varphi/8\pi$ of the signal is greater than ξ/p, and the time t_-, during which it is smaller than ξ/p (see fig. 10). This amounts to measuring the small time interval Δt. In fig. 10 it can be seen that the required displacement $\Delta\xi$ is related to Δt by:

$$\frac{\Delta\xi}{p} = m \sin(\Omega\Delta t),$$

whereas in the interpolation $\Delta\xi$ and Δt are assumed to be linearly related (with a slope equal to the initial slope of the sine curve):

$$\frac{\Delta\xi}{p} = m\Omega\Delta t.$$

Since the interpolation need go no farther than $\Delta\xi = p/32$ (midway between two "digital" grating positions), the maximum interpolation error found is the relative difference between $1/(32m)$ and arc sin $[1/(32m)]$. At the chosen mirror amplitude of $m = m_2 = 5.53/8\pi$ this is an error of 0.35% (at $m = m_1 = 2.43/8\pi$ the error would be 1.9%). Theoretically — if there were no grating defects, etc. — it would be possible in this way to take the interpolation down to a factor of 300; for an interpolation factor of 10 the method employed is therefore amply sufficient.

An important point is that our method of interpolation only uses the *phase angle* of the signals S_1 and S_2, and not their amplitude. This means that, just as in the digital measurement described earlier, the result is completely independent of variations in photocell characteristics etc. due to temperature fluctuations and other influences.

The final indication of the grating position is made up from the digital indication in units of $p/16$, in our case 0.5 μm, and the indication, for example from a moving-coil meter, of a value continuously varying between —0.25 and +0.25 μm, which has to be added to the digital indication. This interpolated part can also be fed to an analogue-digital converter if desired, and if an interpolation factor of 10 is used the position can be given in a purely digital indication in units of 0.05 μm.

Another use for the interpolated measuring signal, which may often be of great practical interest, is as the control signal in a closed control loop. In fact, the interpolation provides an analogue signal, and if the grating is fixed to the carriage of a machine tool, for example, a control loop using this signal will be able to hold the carriage very accurately to a programmed position. In a closed control loop using only digital measuring information it has been found in practice — and also from theoretical considerations — that the carriage will continually oscillate back and forth through a distance that may be as large as one measuring unit. In a high-precision lathe, for example, this is not acceptable, even when the measuring unit has been made as small as 0.5 μm.

Applications and further developments

The application of the measuring system described here in a precision lathe and in a step-and-repeat camera used for making photomasks for integrated circuits has already been discussed in this journal in the recent articles noted earlier [1] [2]. The gratings referred to in these articles have a period of 8 μm and total lengths of 20 cm and 6 cm respectively. In the step-and-repeat camera a measurement with interpolation

down to 0.05 μm is used for the (intermittent) displacement in one coordinate direction. For the (continuous) displacement in the coordinate direction perpendicular to this, the appropriate measuring signal is used dynamically for firing the flash lamps at the proper places; the accuracy achieved, which depends on the oscillation frequency of the mirror and on the speed of the carriage, is 0.12 μm [2].

A measuring system operating on the principles described here, for use with machine tools, will shortly be marketed by Philips.

In a further development of the system described in this article it will perhaps be desirable to use an electro-optical effect instead of the mechanical oscillation of the mirror, which has its disadvantages (the equilibrium position can drift and the oscillation frequency is limited). In such a system an alternating electric field would be used to apply a sine-wave modulation to the phase difference between two components of a light beam with polarizations at right angles to one another, although with existing electro-optical materials this would require fairly high voltages.

Another future possibility is that of combining the scanning system of the grating with a control system enabling the grooves of the grating to be used as *guides*: this is perhaps an easier solution than using mechanical slideways, and where extreme accuracy is required a better one. One might also contemplate a combination of two such systems using gratings with sets of grooves at right angles (or even a combination of these two gratings in a chequer-board pattern) for accurate two-dimensional location, which is required in the step-and-repeat camera referred to above, for example, and in various other areas of technology.

Summary. A method has been developed for measuring displacements up to a few decimetres with very high accuracy, using a single reflecting phase grating with a typical grating period of 8 μm. This grating, fixed to the moving object, is scanned by a fixed scanning head incorporating an optical system with a concave mirror, by means of which an image of the grating is formed upon the grating itself. A beam of light is therefore diffracted *twice* by the grating, and the beam diffracted twice to the order +1 is allowed to interfere with the beam diffracted twice to the order −1. This reduces the basic unit of the measurement by a factor of 4. The process described is performed with two incident beams polarized at right angles and the signals are displaced with respect to one another by 1/16th of a grating period; this results in a further reduction of the basic unit by a factor of 2. Electronic counting of the number of times the two signals go through zero gives a digital indication of the displacement in units of 1/16th of a grating period, i.e. 0.5 μm, and forward and reverse movements can also be distinguished. The concave mirror is made to oscillate. This has the advantage that the electronic circuits for the digital indication and interpolation only have to process a.c. signals and the indication does not vanish when the grating is stationary. The article concludes with a description of a simple way of interpolating by a factor of 10 or higher.

II. Displacement measurement with a laser interferometer

H. de Lang and G. Bouwhuis

The measurement of displacement by means of an interferometer is the most accurate method of determining a length. A good piece of evidence for this is the fact that interferometry is the basis for the definition of the international standard unit of length: 1 metre = 1 650 763.73 wavelengths *in vacuo* of a specific spectral line of the krypton isotope ^{86}Kr.

In the classical arrangement for measuring lengths with an interferometer, two beams of light from the same source (and thus to a certain extent coherent) are allowed to interfere after a system of mirrors has made them travel along different paths. The resultant intensity depends on the relative phase of the waves and will thus vary periodically if the optical path length of one of the two beams is made to vary in magnitude uniformly, e.g. by displacing one of the mirrors. A period here corresponds to a change in path length of the magnitude of one wavelength of the light.

For the extremely accurate digital measurements of displacements required in engineering, as discussed in the preceding article (I), the use of a direct interferometric method of this type — without a grating and thus involving no diffraction effects as in I — is very attractive. In this case the wavelength serves as the natural digital unit of measurement. The advent of the laser has now made this application a practical proposition. With the quasi-monochromatic light sources used earlier the requirements imposed on the coherence length (small optical bandwidth) and the local coher-

Dr. H. de Lang and G. Bouwhuis are with Philips Research Laboratories, Eindhoven.

ence (sharp beaming) of the light in order to measure path differences of several decimetres could only be met at the expense of intensity, giving intensities that were much too low for measurements outside the laboratory, even when photoelectric detectors were used. With a gas laser, however, it is readily possible to reach intensities which are 10^6 times higher even with a modest total power, since the radiation is automatically sharply beamed. Since, moreover, the coherence length of the laser light is practically unlimited, it is possible to measure path differences of many metres.

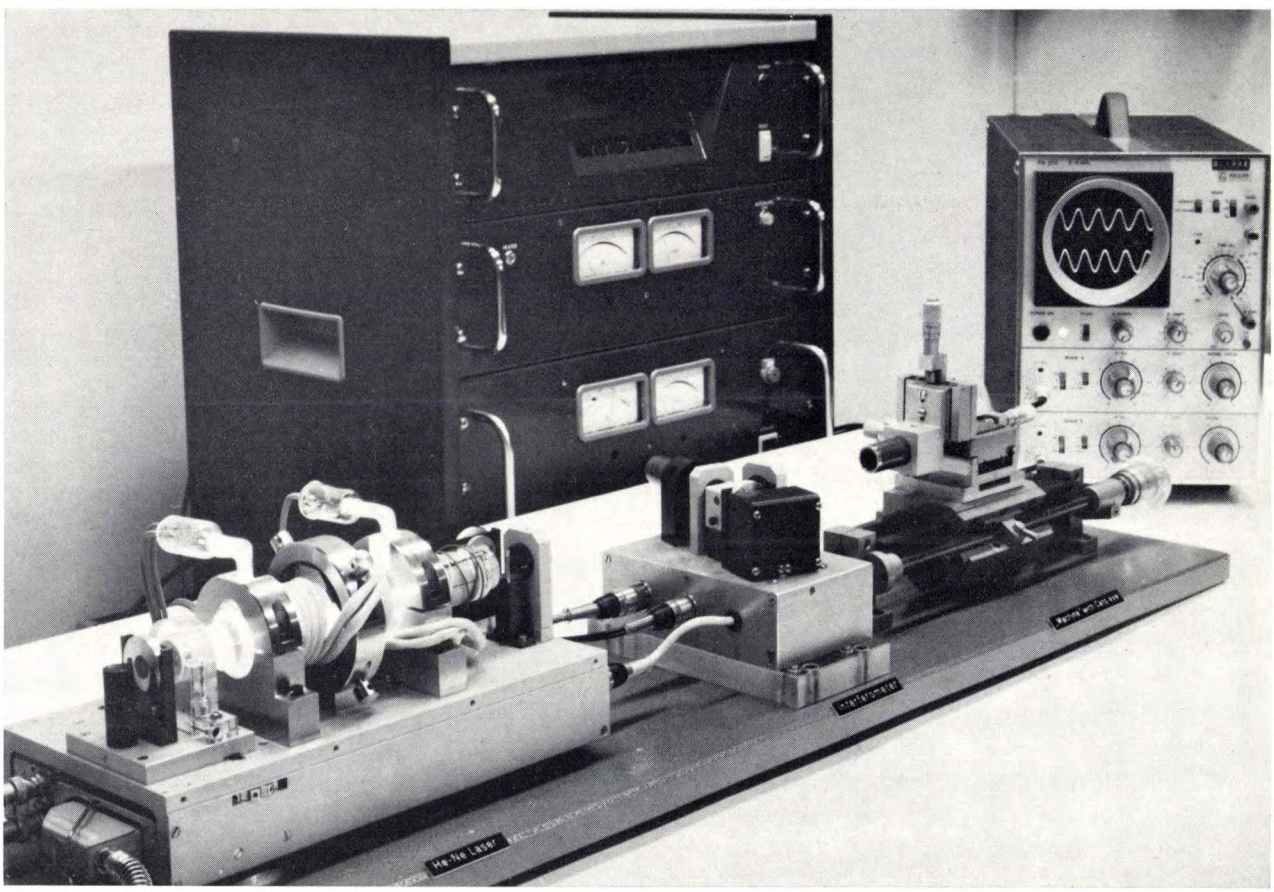


Fig. 1. Arrangement for the highly accurate measurement of displacements using a stabilized helium-neon laser. (As demonstrated at the Physics Exhibition 1969 in London.)

In the following we shall describe an interferometer, made at these Laboratories, for measuring displacements with a helium-neon laser (see *fig. 1*). The outstanding features of the interferometer are entirely comparable with those of the grating interferometer described in I, such as: accuracy to within fractions of a micron, simple digital indication and discrimination between forward and backward displacements. It can therefore be used for similar purposes, although the ranges of displacement for which the two devices can best be used are rather different. Compared with the grating interferometer, the laser interferometer is at present still fairly expensive. For small displacements, say up to 25 cm, the grating interferometer is therefore the best choice; for lengths greater than about 50 cm, on the other hand, it becomes very difficult if not impossible to make a suitable grating, and here therefore the laser interferometer comes into its own. Incidental features may decide the choice of where the fields of application of the two methods overlap. In some applications, for instance, the thermal expansion of the grating may be a disadvantage (although in the step-and-repeat camera recently described in this journal [1] it was seen to be an asset). In other applications difficulties may arise because the laser interferometer also "measures" the changes in the refractive index of the air along the path of the light. The laser interferometer also has a rather awkward *non-metric* basic digital unit

[1] F. T. Klostermann, A step-and-repeat camera for making photomasks for integrated circuits, Philips tech. Rev. **30**, 57-70, 1969 (No. 3).

of measurement (the wavelength of the light or a fraction of it); set against this is the fact that the basic unit can easily be 5 times smaller than in the grating method, so that interpolation may not be necessary. Another aspect — which is purely mechanical but may sometimes be decisive — is that a grating has to be fixed rigidly to the object whose displacement is to be measured. This is not always possible, whereas it is not usually very difficult to form a flat reflecting surface on the object, e.g. by evaporating a metal film on to it.

In describing our laser interferometer we have to deal with two quite separate parts of the equipment: the helium-neon laser used as the light source, and the actual interferometer system.

Stabilization of the laser frequency

The mirror system of our gas laser [2] is adjusted in such a way that only one laser mode is obtained; as is well known, the radiation of such a mode is limited to an extremely narrow frequency band, and it is therefore usual to speak of *the* laser frequency. This frequency is determined by the distance between the mirrors and would remain highly constant for constant temperature and other conditions; its actual value, however, cannot be predicted with sufficient accuracy because of the Doppler effect of the gas molecules. As a result of this effect, the spectral line of the gas at which the laser action occurs, e.g. the neon line of wavelength 1.15 μm, has a width of about 4×10^{-6} μm, and the much narrower laser mode may take up any arbitrary position within this band. Its frequency can therefore vary by more than 1 in 10^6. For our displacement measurements, where it is the wavelength that determines the base unit, it is necessary to fix the laser frequency at a specific location in the Doppler profile with an accuracy of say 1 in 10^7: this will enable us to measure displacements of 1 metre with an accuracy of better than 0.1 μm.

To fix the laser frequency to this degree of accuracy we make use of a stabilizing method developed at these Laboratories as long ago as 1964 [3], which makes use of the Zeeman effect. The He-Ne laser is placed in an axial magnetic field. The effect of this is to split each spectral line, including the neon line used in the laser, into two components circularly polarized in opposite senses; each of these has its own Doppler profile; see *fig. 2*. The laser mode also splits into two components circularly polarized in opposite senses. The two components also differ in frequency by a small amount [4], of the order of 1 MHz; this does not matter for our stabilization method, but we shall come back to this point in the next section. Because of the relative shift of the two Doppler profiles the two laser mode components will now have different intensities. In the situation sketched in fig. 2, where the (average) mode frequency f_m is *lower* than the undisturbed line frequency f_0, the left-hand circular component will be stronger than the right-hand one. Now the laser can be tuned to a certain extent by means of electrostrictive and thermal changes in length. If when the laser is tuned the mode frequency f_m becomes *larger* than f_0, the difference in intensity of the two mode components changes sign, as can be seen in fig. 2. This provides a simple criterion for adjustment, and the frequency f_m can be stabilized to the value f_0 by measuring the intensity difference and using this measurement to correct the tuning until the difference is zero. The arrangement used for this is shown in *fig. 3*. The laser beam passes through an electro-optical crystal of potassium dihydrophosphate (KDP), whose bi-refringence is modulated with an a.c. voltage. A photoelectric cell in the transmitted beam, placed behind an analyser oriented at 45°, receives alternate contributions from the right-hand and the left-hand circular mode components. The magnitude and phase of the resultant alternating signal thus depends upon the difference in the intensity of the two components, and by comparing this signal with the alternating voltage

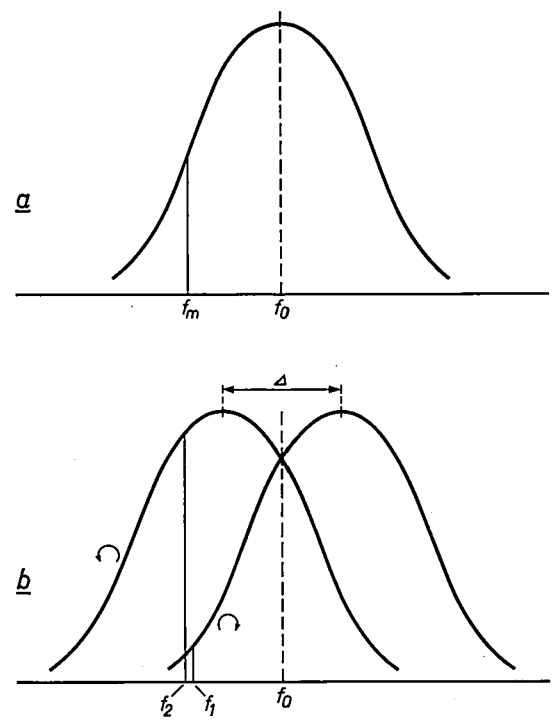


Fig. 2. *a*) Doppler profile of the spectral line 1.15 μm of neon (unperturbed frequency f_0). The mirror configuration of the laser is such as to give a (monochromatic) laser mode somewhere within this profile (frequency f_m).
b) By means of a magnetic field the spectral line is split into two components (Zeeman splitting Δ), each with its own Doppler profile, and the laser mode separates into a right-hand and a left-hand circularly polarized component, differing slightly in frequency. The difference in intensity of the two components depends on the difference between the (mean) mode frequency f_m and the line frequency f_0.

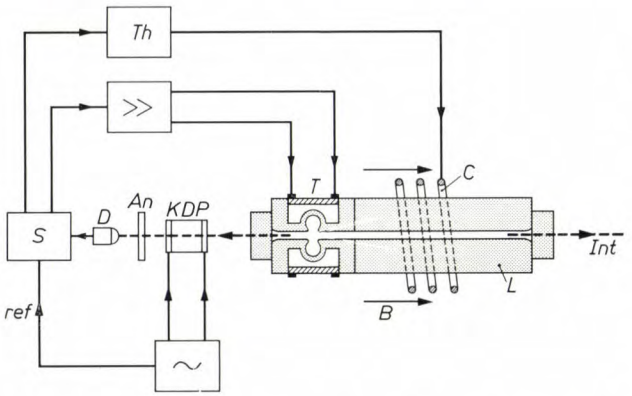


Fig. 3. Arrangement with laser L, where the laser frequency (f_m in fig. 2) is fixed at the unperturbed neon-line frequency f_0. The magnetic field B causes the Zeeman splitting. The intensity difference between the two circularly polarized mode components (f_1 and f_2 in fig. 2) is detected by means of an electro-optical crystal (KDP) whose bi-refringence is modulated, an analyser An and a photodetector D. By comparing (in S) the detector signal with the a.c. voltage on the crystal, which serves as reference signal, a control signal is obtained which controls the voltage on an electrostrictive tuning element T in the laser. The signal also controls a thermal tuning of the laser, via Th and a heating element C, in order to keep within the operating range of the electrostrictive element. The laser beam is directed on to the interferometer (at Int).

applied to the KDP crystal a control voltage is obtained which is applied after amplification to an electrostrictive element in the laser body [5] for the required length correction. A second, much slower, control operated by the temperature of the laser as a whole ensures that we keep within the operating range of the electrostrictive element.

In this way it is not difficult to keep the laser frequency stable for many hours with a variation of less than 1 in 10^9. This is a hundred times better than the requirement we have quoted, and in fact this amount of stabilization is more than can really be of use here because of the existence of certain other effects: variations in wavelength due to variation of the refractive index in air, and variations of the neon-line frequency f_0 due to the effect of the gas pressure and other parameters, which can differ considerably from one laser to another. These various effects can be expected to give variations of about 1 in 10^7.

Arrangement of the interferometer

As a result of the method of stabilization, the laser supplies a beam of light consisting of two components of equal intensity that are circularly polarized in opposite senses and have slightly different frequencies. These various circumstances are all turned to useful effect in the interferometer described here.

The arrangement is shown schematically in *fig. 4*. The laser beam first strikes a semi-transparent mirror. The *reflected* radiation is detected by a photoelectric cell behind an analyser, which thus passes only light polarized linearly in a particular direction. The superposition of the two circularly polarized components with slightly different frequencies f_1 and f_2 corresponds to a linearly polarized wave whose direction of polarization rotates "slowly", at half the difference frequency (i.e. about 1 MHz). The detected signal thus has the difference frequency $f_1 - f_2$; it is used as a reference signal.

The part of the radiation which has passed through the semi-transparent mirror passes through a quarter-wave plate, which converts each of the two circularly polarized components of opposite sense into a linearly polarized wave; the two waves have directions of polarization at right angles to each other, as indicated in fig. 4. For convenience we shall refer here to the "perpendicular" component (frequency f_1) and the "parallel" component (frequency f_2).

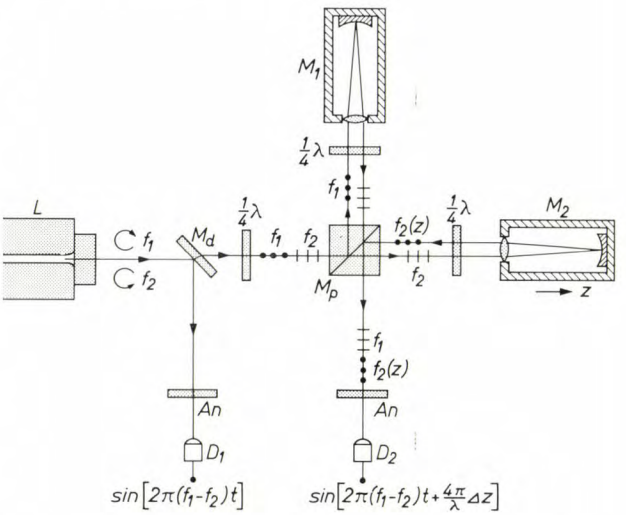


Fig. 4. Diagram of the laser interferometer. The radiation from the laser L, in two components polarized circularly and in opposite senses and with slightly different frequencies f_1 and f_2, passes through a semi-transparent mirror M_d, which separates a fraction that goes to an analyser An and the photodetector D_1 for a reference signal. The two components are then converted into two *linearly* polarized beams by a quarter-wave plate (the "perpendicular" and the "parallel" beams denoted in the figure by dots and dashes). A polarizing beam splitter M_p directs these beams to the fixed and the variable interferometer arms respectively, each with a "cat's-eye", M_1 and M_2. The displacement of M_2 in the direction of the beam (z-direction) is now determined by the phase difference shown by the beat signal, generated in the photodetector D_2 by the two reflected components, with respect to the reference signal at D_1.

[2] The laser is basically the same as the one previously described in this journal: J. Haisma, S. J. van Hoppe, H. de Lang and J. van der Wal, Philips tech. Rev. **24**, 95, 1962/63.

[3] An almost identical method has been subsequently described by I. Tobias, M. L. Skolnick, R. A. Wallace and T. G. Polanyi, Appl. Phys. Letters **6**, 198, 1965.

[4] H. Statz, R. Paananen and G. F. Koster, J. appl. Phys. **33**, 2319, 1962.

[5] An element of this type has been described by J. Haisma, Construction and properties of short stable gas lasers, Thesis, Utrecht 1966; Philips Res. Repts. Suppl. 1967, No. 1.

These two components now enter the interferometer proper. Here they strike a polarizing beam-splitter consisting of a number of vacuum-evaporated layers on which the beam is incident at the Brewster angle. The parallel component here is almost entirely transmitted and the perpendicular component almost completely reflected [6]. As a result, the component with the frequency f_1 goes through the fixed arm (shown pointing upwards in fig. 4) and the one with the frequency f_2 goes through the variable arm of the interferometer (shown pointing to the right).

We shall first follow the "perpendicular" component (f_1). This passes through a quarter-wave plate oriented at 45°, to a fixed mirror system M_1, consisting of a lens with a mirror located in its focal plane (a "cat's-eye"). With a system of this kind the beam is reflected in a direction exactly opposite to the direction of incidence, irrespective of the angular position of the system; this means that the angular position does not have to be adjusted [7]. The reflected beam passes through the quarter-wave plate again and back to the polarizing beam splitter. Since passing twice through the quarter-wave plate rotates the plane of polarization through 90°, the beam is now transmitted. The radiation then passes through an analyser, oriented at 45°, to a photodetector.

The "parallel" component (f_2) of the original beam traverses the variable arm of the interferometer in a similar way. This time the reflection is from the cat's-eye M_2, which is fixed to the object whose displacement Δz is to be measured. Here again the plane of polarization is rotated through 90° after the beam has passed through the quarter-wave plate twice, so that the returning beam in this case is mainly reflected by the polarizing beam splitter and reaches the photodetector via the analyser together with the other returning beam (f_1).

The analyser, oriented at an angle of 45°, transmits the same fraction of both beams, and since the beams have different frequencies the result is a wave with a beat frequency $f_1 - f_2$ whose phase depends on the path difference between the components and hence on the coordinate z of the moving cat's-eye M_2. This results in a photosignal whose a.c. component has the frequency $f_1 - f_2$ and a phase which is a linear function of z: when M_2 is displaced over a distance Δz equal to half the wavelength, a phase shift in the signal of $\Delta\varphi$, equal to 2π, is produced, so that in general:

$$\Delta\varphi = \frac{4\pi}{\lambda}\Delta z. \qquad \ldots\ldots \quad (1)$$

This phase shift can be determined in magnitude and direction (forward or reverse displacement) by comparing the phase of the resultant photosignal with that of the reference signal mentioned earlier. In the electronic signal processing, which takes place in very much the same way as described in I for the measuring system using a grating, one positive or negative pulse is obtained every time the phase of the signal increases or decreases by an amount π. The displacement is thus measured by counting unit steps of the magnitude of a quarter of a wavelength, i.e. about 0.29 μm (see equation 1).

Instead of this digital measurement an "analogue" measurement can also be carried out and changes in optical path length very much smaller than the basic unit step can then be determined. With simple means an accuracy of a thousandth of a wavelength is obtained, i.e. 1 nanometre, which is well below the limit that might possibly be considered for displacement measurements in engineering.

An important practical point is that the interferometer described delivers an a.c. signal even when both cat's-eyes are stationary. Just as in I, this means that d.c. amplifiers are not required. Moreover the measurement is not affected by variations in the laser intensity or the magnetic field. We have already mentioned that very large displacements can be measured with this interferometer. The displacement can also take place at *high speed*: if for example the speed is 0.5 metre per second the photodetector must be capable of following a frequency of 2 MHz, which is certainly not exceptionally high. (A higher magnetic field is required in this case to give the frequency difference $f_1 - f_2$ of the laser mode components a value high enough to prevent the frequency of one of the photodetector signals from assuming the value zero — see fig. 4.)

By means of a modification shown in *fig. 5* [8] the base unit can be reduced by a factor of 2. Instead of immediately separating the "perpendicular" and the "parallel" radiation components by a polarizing prism, both components (f_1 and f_2) are allowed to pass through each of the two interferometer arms before the separation is effected. A quarter-wave plate in one of the two arms rotates the direction of polarization of both components by 90°, and it can be seen from fig. 5 that this will have the result that a phase increase in one detected beat signal, e.g. the one with perpendicular polarization, due to a displacement Δz, will be associated with an equally large phase decrease in the other beat signal, with parallel polarization. The relative phase of the two signals therefore varies twice as quickly as in the first arrangement.

Concluding remarks

The merits of the grating interferometer and of the laser interferometer have already been compared in the foregoing. In conclusion we should like to emphasize

that there are other potential applications of the laser in metrology — some of them still well beyond the needs of present-day engineering. Some examples have been described elsewhere [9] [10]. Only one of these should be mentioned here. In the preceding article (I, page 155) we referred in passing to the exceptionally sensitive detection of displacement which is possible with the signals of the grating interferometer. As might be expected, our laser interferometer is equally capable in this respect: what is really astonishing, however, is the ease with which a few components can be rigged together to give extremely high sensitivities once the laser beam is available. Thus with a simple arrangement including what is known as a confocal etalon, we were easily able to detect an oscillation with an amplitude of 0.06 nanometres [10].

It seems fair to say that the present progress of technology towards the vanishingly small — witness the integrated circuit — will not at any rate be hampered by any shortcomings in metrology.

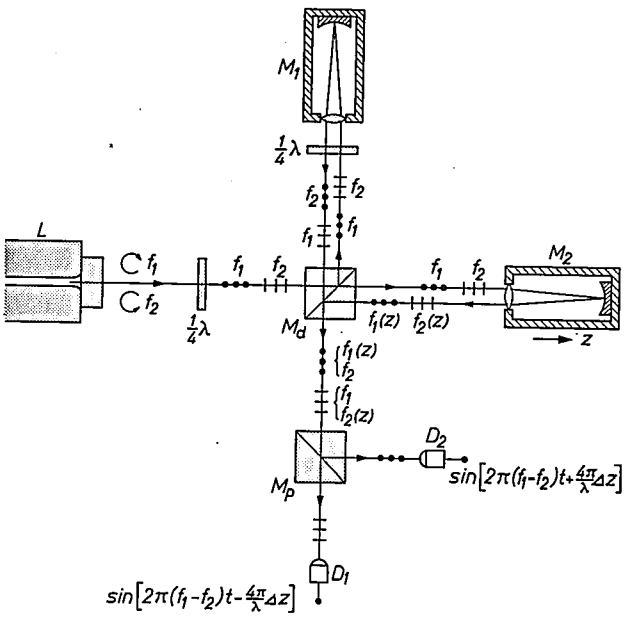


Fig. 5. Modification of the laser interferometer, which further subdivides the basic digital unit of measurement by a factor of 2. The letters have the same significance as in fig. 4.

[6] The possibility of using a multilayer mirror for almost completely separating light components polarized at right angles to each other has been pointed out in: H. de Lang and G. Bouwhuis, Colour separation in colour-television cameras, Philips tech. Rev. **24**, 263-271, 1962/63, footnote [4] on p. 267.

[7] Another system possessing this feature has three flat mirrors at right angles to one another (a "triple mirror"). This is not suitable in our case because of its anisotropy. The cat's-eye is also simpler and cheaper.

[8] This modification was proposed by E. T. Ferguson.

[9] H. de Lang, G. Bouwhuis and E. T. Ferguson, Physics Letters **19**, 482, 1965. See also H. de Lang, Polarization properties of optical resonators passive and active, Thesis, Utrecht 1966; Philips Res. Repts. Suppl. 1967, No. 8.

[10] G. Bouwhuis, Interferometrie met gaslasers (in Dutch), Ned. T. Natuurk. **34**, 225-232, 1968 (No. 8).

Summary. For measuring displacements of the order of 1 metre or greater it becomes difficult to make and use gratings that are long enough. In such cases the classical method of measuring lengths with an interferometer is indicated. If a laser is used as the light source, the exceptionally great coherence length of the laser radiation enables a distance of many metres to be covered, while a digital measurement using the wavelength as the basic unit (e.g. 1.15 μm with a helium-neon laser) can readily be carried out. In the arrangement described in this article the laser frequency is accurately fixed by making use of a Zeeman splitting of the laser mode and a control circuit with electrostrictive tuning of the laser. In the actual interferometer the Zeeman splitting gives two a.c. signals whose phase difference depends linearly on the displacement. A phase shift of π corresponds to a displacement of a quarter wavelength (i.e. about 0.29 μm). This digital unit can be halved by making a simple modification.

A Pirani gauge for pressures up to 1000 torr and higher

L. Heijne and A. T. Vink

In vacuum technique there is a need for a pressure gauge covering a range of 7 to 8 decades, with the upper limit in the neighbourhood of atmospheric pressure. The authors of this article have designed a new and very simple instrument, based on the Pirani gauge, that fulfils these requirements.

The Pirani gauge makes use of the pressure-dependence of the thermal conductivity of a gas. The gas whose pressure is to be measured is admitted to an enclosure containing an electrically heated wire which forms one arm of a Wheatstone bridge. The cooling of the wire due to the thermal conduction of the gas changes the electrical resistance of the wire, and the change is a measure of the gas pressure. This method of measurement, described by Pirani in 1906 [1], can be used at pressures from about 10^{-4} torr (about 10^{-6} torr when special measures are taken) up to 100 torr at the most; at higher values the thermal conductivity of gases is virtually independent of the pressure. The Pirani gauge has the advantages that the measurement is electrical, no gas is removed from or supplied to the system in which the pressure is being measured, and the instrument can be used for practically all kinds of gas. Moreover, the lower limit of the measuring range conveniently overlaps the upper limit of the widely used ionization gauges [2]. A feature which the Pirani and ionization gauges have in common is that the calibration curve depends on the kind of gas.

Unfortunately the upper limit of the measuring range is still well below atmospheric pressure, and this means that a third type of instrument, such as a mercury gauge, has to be used when pressures above 100 torr in vacuum equipment have to be measured. This would not be necessary if the range of the Pirani gauge could be extended up to about 1000 torr. Various investigators have tried to do this by making use of the pressure dependence of additional cooling due to *convection* of the gas near the hot wire, this convection being initiated by a second heating element, by using a very hot wire, or by a blower [3]. We have succeeded in obtaining a reproducible pressure-dependent additional cooling of the wire due to convection *without introducing extra components* and with the wire no hotter than about 300 °C: this has been achieved through a careful choice of the temperature of the wire and the dimensions of the envelope and the wire. By combining this manometer tube with a special control circuit we have produced a very simple pressure gauge that will operate in a range of about 0.0001-1000 torr.

Fig. 1 shows the complete pressure gauge set up for measurement, and *fig. 2* shows the design of the manometer tube. The wire is coiled and mounted horizontally some way away from the convex upper wall of the envelope. Mounting the wire horizontally usually gives the strongest natural convection [4] [5];

Fig. 1. The new Pirani gauge set up for measurement. The control unit with the meter can be seen on the left.

Dr. L. Heijne and Ir. A. T. Vink are with Philips Research Laboratories, Eindhoven.

coiling the wire makes tension springs unnecessary and also assists the convection. The wire material is an alloy of platinum with 10% rhodium; this material was chosen because a) it does not react with reactive gases such as oxygen (whereas tungsten, which is often used, does), b) it has a low coefficient of emission for heat radiation (which limits the power taken by the wire *in vacuo*), and c) it possesses sufficient mechanical stability at the working temperature to be used in the form of a horizontally tensioned coil (pure platinum is too soft for this).

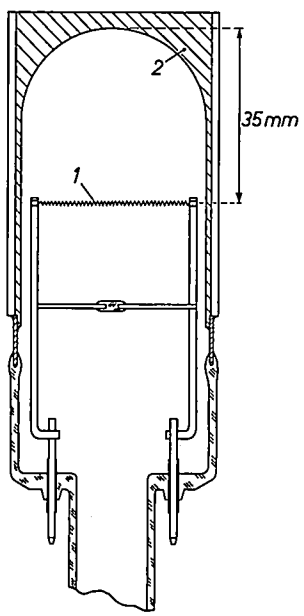


Fig. 2. Design of the new manometer tube, with convex upper wall, for pressures of about 0.0001-1000 torr. The operation below about 10 torr depends on the thermal conduction of the gas, as in the original Pirani gauges; above 10 torr it depends on convection. *1* coiled wire of platinum with 10% rhodium, diameter 25 μm, wound on a 1 mm mandrel. Part *2* of the tube is at room temperature; it is made of copper, has a black matt surface and is fitted with cooling fins. The tube must be used in the position shown, with wire *1* horizontal and wall *2* at the top.

Sensitivity, stability of the convection pattern, and measuring range all imposed different requirements on the design of the new manometer tube, and a compromise had to be found between them. In general the *sensitivity* increases with the distance between the wire and the upper wall and with the temperature of the wire; this temperature therefore sets a lower limit to this distance. On the other hand the *stability* of the convection sets an upper limit to this distance, particularly for heavy gases (e.g. CO_2) and high pressures. We have found experimentally that an adequate sensitivity — in other words, a measurable heat loss due to convection — is obtained if the distance between the wire and the upper wall is at least 15 mm at pressures of up to 200 torr, and the wire temperature is higher than about 200 °C. Above 200 torr a distance of about 10 mm would be sufficient, and in that case a lower wire temperature would also be permissible. We also found that adequate stability is assured if the following conditions are satisfied:

1) The distance between wire and upper wall should be 40 mm at the most for pressures up to about 1 bar (760 torr).
2) The inner surface of the envelope above the wire should be smoothly contoured, with no sharp corners, edges or protrusions.

The dimensions of envelope and wire used in our manometer tube ensure the necessary stability and also a smooth transition between the two mechanisms of heat transport, so that a wide range of pressures can now be measured with a single gauge.

The convection principle can also be applied for pressures above 1 bar, but to ensure stable convection at the greater density of the gas the distance between the wire and the upper wall must be smaller, which raises the lower limit of the measuring range. For the range from about 1 to about 11 bars we have made an experimental tube with a cylindrical envelope and a wall distance of 7 mm; some information about this tube is given in the caption to *fig. 3*.

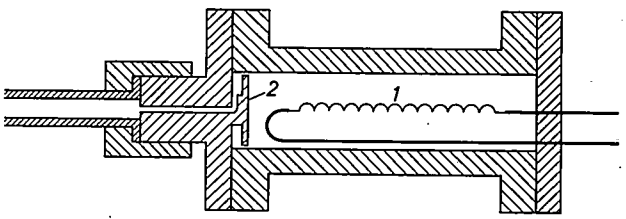


Fig. 3. Schematic cross-section of a manometer tube with cylindrical envelope, for measuring pressures from 1 to 11 bars. The wire *1* is roughly parallel to the axis of the cylinder, but need not coincide with it. Distance between wire and upper wall is about 7 mm. Wire supports and contact pins are of gas-tight material. Draught effects due to pressure changes, which are troublesome at these high pressures, are largely eliminated by a screen *2* in front of the gas inlet. The stability is very good up to pressures of 11 bars; the maximum measurable pressure is probably much higher.

[1] M. v. Pirani, Verhandl. Dtsch. Phys. Ges. **8**, 686, 1906.
[2] See for example S. Dushman and J. M. Lafferty, Scientific foundations of vacuum technique, Wiley, New York 1962.
[3] J. B. Johnson, Rev. sci. Instr. **27**, 303, 1956.
P. E. Seiden, Rev. sci. Instr. **28**, 657, 1957.
J. A. McMillan and T. Buch, Rev. sci. Instr. **28**, 881, 1957.
A. W. Smith, Rev. sci. Instr. **30**, 485, 1959.
K. Diels and R. Jaeckel, Leybold Vakuum-Taschenbuch, Springer, Berlin 1962, p. 105.
[4] M. Jakob, Heat transfer, Wiley, New York 1949, part I, chapters 3 and 22-26.
[5] H. von Ubisch, Appl. sci. Res. A **2**, 364-430, 1949/51.

Fig. 4 shows a manometer tube in a bridge circuit; M is a high-resistance voltmeter, R_p is the resistance of the hot wire which changes when the gas is admitted. The following well-known and essentially different methods can be used for the measurement.

1) At a constant bridge voltage V_b the meter reading can be taken as a measure of the pressure [6].
2) The temperature of the wire, and hence its resistance R_p, can be adjusted by varying the supply voltage V_b until the voltmeter reads zero, the voltage required being a measure of the pressure.

In method (1) the temperature of the wire decreases with increasing gas pressure; if it were to approach the temperature of the envelope, the calibration curve would become too flat, because of the insufficient temperature difference between wire and envelope (*not* because the pressure dependence of the thermal conductivity is too small). Method (2) does not show this type of saturation [5]; this is therefore the obvious method for our pressure gauge, which works with a fairly low wire temperature and is intended to measure both high and low pressures.

It is perhaps useful to look for a moment at method (1), which is widely used with the classical Pirani gauge [7]. We shall consider three limiting cases, in which we choose:

1a) $R_0 \ll R_p$ (and hence $R_1 \ll R_2$); at constant bridge voltage V_b the *voltage* across the wire then remains virtually constant.

1b) $R_0 \gg R_p$ (and hence also $R_1 \gg R_2$); at constant V_b the *current* through the wire now remains virtually constant.

1c) $R_0 \approx R_p$; at constant V_b the *power* dissipated in the wire remains virtually constant (decreasing R_p by 20% reduces the power by 1.2% and a 40% reduction in R_p decreases the power by 6.25%).

If a material with a positive temperature coefficient (e.g. a metal) is used for R_p, the saturation in the calibration curve caused by an insufficient temperature difference between wire and envelope occurs soonest with method (1b) (I_p constant) since the power dissipated in the wire ($I_p^2 R_p$) then decreases with increasing gas pressure. In methods (1a) (V_p constant) and (1c) (power constant) the saturation in this case is slower.

With the method we have chosen (2) it is desirable to keep the wire temperature constant *automatically* [5]. Not only does this provide the advantage of a direct-reading instrument, but it also avoids the risk of the wire burning out if the operator reduces the voltage V_p too late when the pressure drops suddenly to a low value. For this purpose we use an electronic control unit. The circuit diagram is shown in *fig. 5*. The bridge difference voltage ΔV is amplified by a differential amplifier and then used as the supply voltage for the bridge. If the gas pressure increases, the wire temperature falls slightly, causing an increase in ΔV and hence in V_b that almost completely offsets the cooling. In addition to the resistors R_0, R_1 and R_2 of the bridge

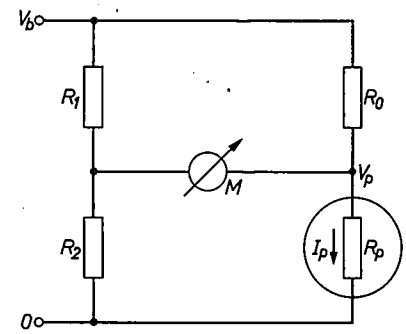


Fig. 4. Circuit of the tube in a Wheatstone bridge. R_p resistance of the hot wire. R_0, R_1 and R_2 resistors. V_b bridge voltage. V_p voltage across the hot wire. M high-resistance voltmeter.

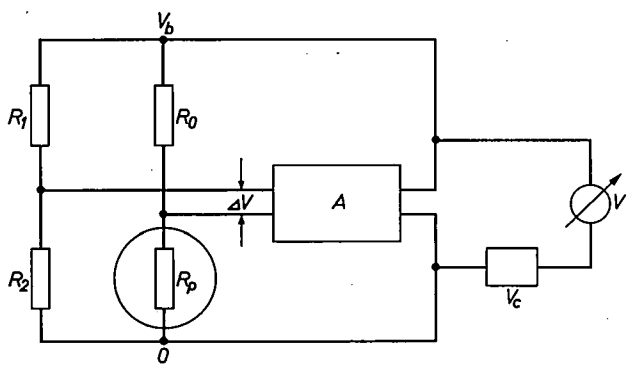


Fig. 5. Measuring circuit with automatic control of the wire temperature. A is a differential amplifier for d.c. voltage; amplification about 500×. V_c compensation for supply voltage in vacuum. The voltmeter V indicates $V_b - V_c$, the increase in the bridge voltage after the admission of gas; it is calibrated separately for each type of gas. Other symbols are as in fig. 4.

circuit and the differential amplifier, the control unit (on the left in fig. 1) contains a voltmeter for various voltage ranges, which is connected to the voltage V_b. The zero reading of the meter is made to correspond to vacuum by subtraction of a constant, but adjustable, voltage V_c: in a vacuum the wire still loses heat by radiation and by conduction via the mountings.

The calibration curve of the combined control unit and manometer tube of fig. 2 is given in *fig. 6* for nitrogen. As already noted, the calibration curve depends on the kind of gas being measured; light gases like helium and hydrogen have a higher thermal conductivity. The calibration curve shown for nitrogen is also valid for air and oxygen to a very good approximation. The dashed curve presents for comparison the calibration curve of a "classical" Pirani gauge, with identical wire, in the region where it differs from that of the new type.

[6] Instead of reading the voltmeter, one can also adjust the resistance R_1 or R_2 (or both of them) until the voltage across M is zero again, and use this resistance value as a measure of the pressure.

[7] See for example M. Pirani and J. Yarwood, Principles of vacuum engineering, Chapman & Hall, London 1961, p. 102. Following Pirani, only methods (1a) and (1b) are generally used besides method (2). Method (1c) is not regarded as important in the literature.

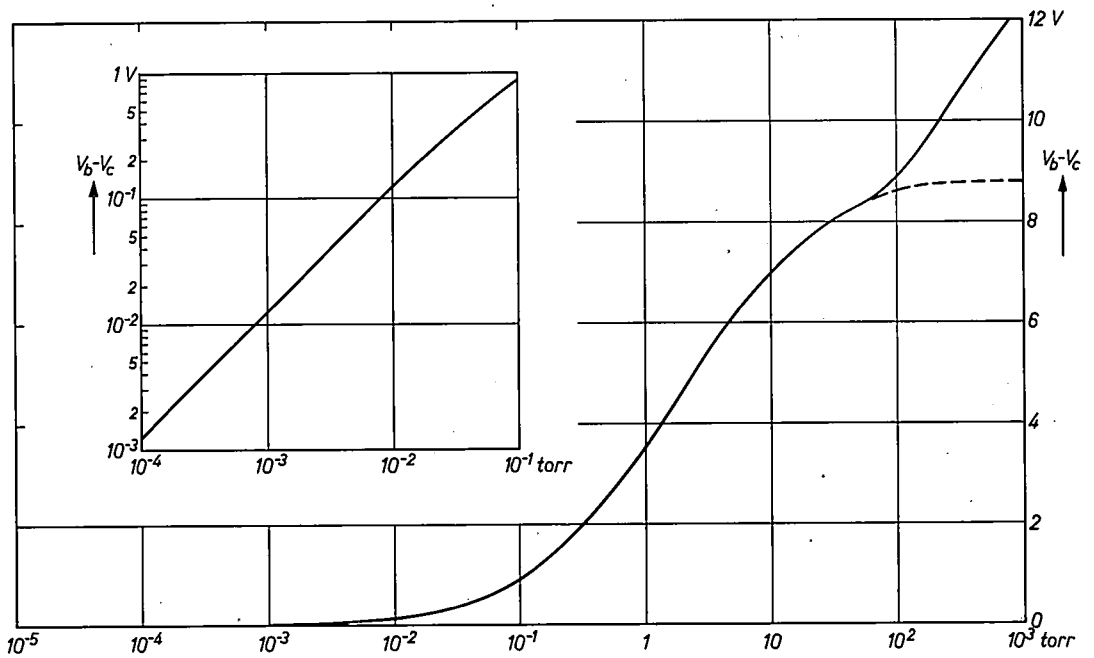


Fig. 6. Nitrogen calibration curve for the gauge unit using the manometer tube of fig. 2. The range below 10^{-1} torr is given with a logarithmic ordinate in the inset. The gas pressure is plotted horizontally, the bridge voltage less the value for a vacuum is plotted vertically ($V_b - V_c$, see fig. 5; $V_c \approx 1$ volt). The dashed curve relates to a corresponding "classical" Pirani gauge in the range where it differs from the curve for the new type.

Finally *fig. 7* shows the calibration curve of the tube in fig. 3 for pressures above 1 bar.

In designing the combined manometer tube and control unit our first concern was to make the equipment simple, reliable and easy to use over the whole pressure range. No attempt was made to achieve the highest possible accuracy and sensitivity in specific pressure ranges, and therefore no attempt was made to stabilize the wall temperature. Nevertheless, the accuracy obtained is sufficient for most applications. With careful use the voltage established at a certain pressure is found to be reproducible to within about 1% of the full scale deflection of the measuring range in use; the corresponding accuracy of the pressure measured depends on the pressure range and can be determined with the aid of the calibration curve.

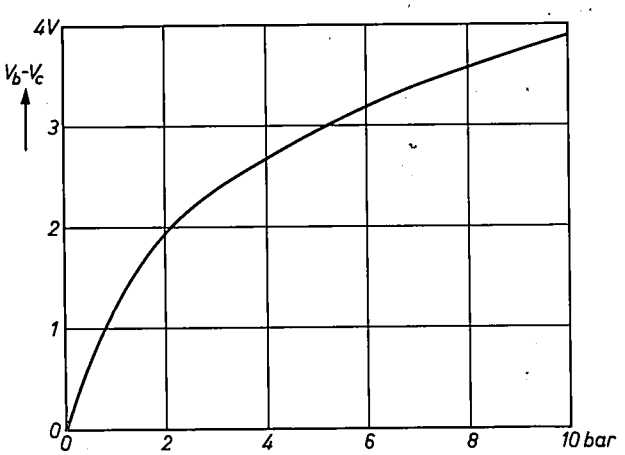


Fig. 7. Nitrogen calibration curve for the gauge unit using the tube of fig. 3; the horizontal scale gives the *excess pressure* in bars. The voltage V_c corresponds here to the bridge voltage V_b at 1 bar (about 12 volts).

Summary. The operation of the Pirani gauge relies on the pressure dependence of the thermal conductivity of a gas. When a gas is admitted into the gauge near a hot wire the change in the electrical resistance of the wire, due to the cooling caused by the thermal conduction of the gas, is used as a measure of the gas pressure. The measuring range (previously from about 0.0001 to 100 torr at the most) can be extended upwards by making use of the *convection* of the gas around the wire. The article describes a Pirani gauge for a pressure range from about 0.0001 to about 1000 torr in which this principle is applied. The design is extremely simple because of the appropriate choice of wire temperature and of the dimensions of tube and wire. The temperature and hence the electrical resistance of the wire are kept automatically constant in a Wheatstone bridge by means of an electronic control unit; the voltage required is a measure of the gas pressure. The design parameters are briefly discussed in connection with stability and sensitivity. A variant of the tube is suitable for measuring pressures from 1 to 11 bars and perhaps higher.

PTC thermistors as self-regulating heating elements

E. Andrich

The resistance of PTC thermistors varies with temperature in such a way that they can be used as heating elements with particularly interesting characteristics. When a sufficiently large voltage is used, they reach a temperature near their ferroelectric Curie point with a high initial current in a few seconds, assume a high value of resistance at this temperature and act from then on like thermostats. These heating elements thus automatically regulate their power consumption to suit the heat requirements. Investigations with laboratory models of soldering irons and domestic hotplates have given impressive demonstrations of the special advantages of these elements.

Introduction

PTC thermistors are resistors made of semiconducting materials with large positive temperature coefficients. Ceramic semiconductors like barium titanate or a mixture of barium titanate with strontium and lead titanate are usually used. These materials have a transition temperature (Curie temperature) at which the crystallites change from the tetragonal to the cubic phase. This transition is accompanied by a marked change in electrical properties. In particular, the resistance increases by several powers of ten when the temperature is raised to the vicinity of the Curie point. The abrupt increase in resistance is steepest in semiconducting barium titanate, whose Curie point lies at 120 °C. The temperature coefficient in this case reaches a maximum of 60% per degree (*fig. 1*). In mixed titanates, whose PTC characteristics lie in other temperature ranges, depending upon the composition of the material, the maximum change of resistance per degree is generally between 10 and 20% (*fig. 2*).

An obvious application of these substances with this sharp increase in resistance is in temperature-dependent switching processes: if the temperature of the thermistor increases to the neighbourhood of the Curie temperature, perhaps because of an increase in the ambient temperature or an increase in the electrical power converted in the device, it then restricts the current flowing through it. This effect can be used, for example, to reduce the power consumption of a load connected in series with it, or to switch over a relay.

A special case is encountered when the thermistor itself is the load and the power supplied to it is converted into a useful heat output. In this case it combines various functions and operates as a self-regulating

Dipl.-Phys. E. Andrich is with the Aachen laboratory of Philips Zentrallaboratorium GmbH.

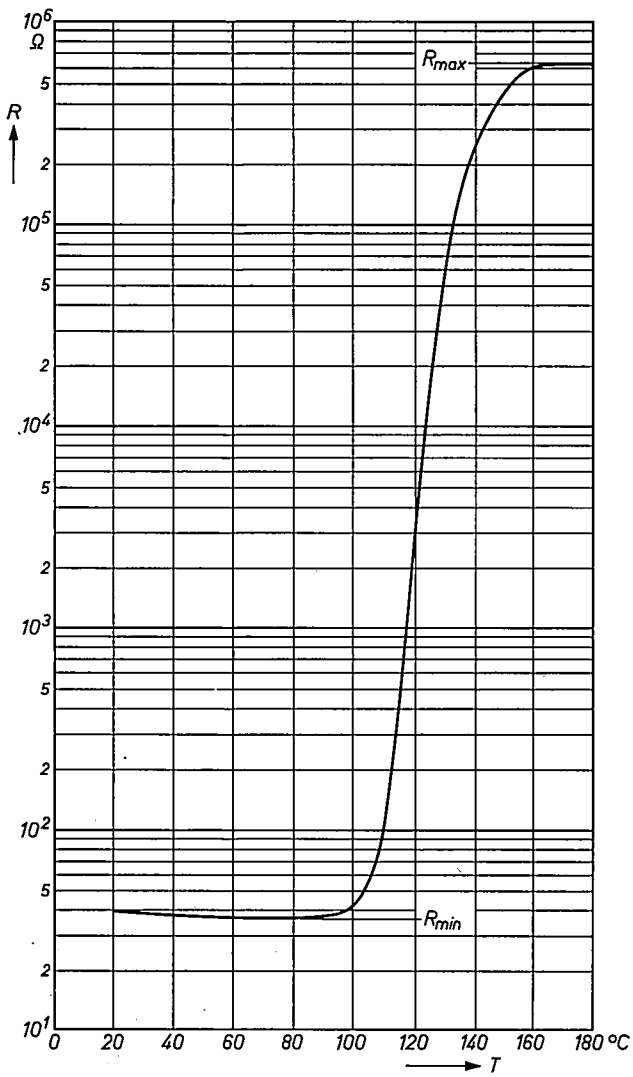


Fig. 1. Resistance-temperature characteristic of a PTC thermistor of semiconductor $BaTiO_3$. Its Curie temperature is 120 °C. The temperature coefficient also has its maximum here, with about 60% change in resistance per degree.

heat source, in other words as a thermostat. At a given voltage the PTC thermistor initially draws a high current, since at first it is still cold and its resistance is low. Its temperature then rises because of Joule heating and it would become white hot in a few seconds if the sharp increase in resistance at the Curie temperature did not immediately restrict the power absorbed. A state of equilibrium then arises in which the power absorbed adjusts itself to be equal to the heat dissipated; the thermistor tries to keep its temperature in the vicinity of the Curie point. This can often be a useful feature of a heating element that is subjected to a varying heat drain. For example, a soldering iron fitted with such an element adjusts its power intake to suit the conditions of the soldering operations. It heats up quickly but does not become overheated when not in use. Because of this there is less corrosion of the bit and also, where there is a large loss of heat, as when soldering a chassis, the heat loss is quickly compensated [1].

Experimental soldering irons of this type have been made with some success at the Philips laboratories in Aachen (there is no production of these irons as yet). Other applications now being studied include a fast-heating hotplate for domestic use, and a simple thermal conductivity meter incorporating a PTC thermistor as a heat source of virtually constant temperature for laboratory use. This instrument can be so designed that the power or current consumption are exactly proportional to the thermal conductivity of the test object.

The use of PTC thermistors as heating elements presents a number of technical problems. The elements should not crack when heated up quickly, they should be capable of withstanding high electrical fields and should have an adequate life at operating temperatures between 300 and 400 °C. These requirements are partly met by using special methods for fabricating these ceramic semiconductors and applying the contacts. The work at the Aachen laboratories has eliminated the problems encountered in the reproducible manufacture of these soldering irons with their heating thermistors.

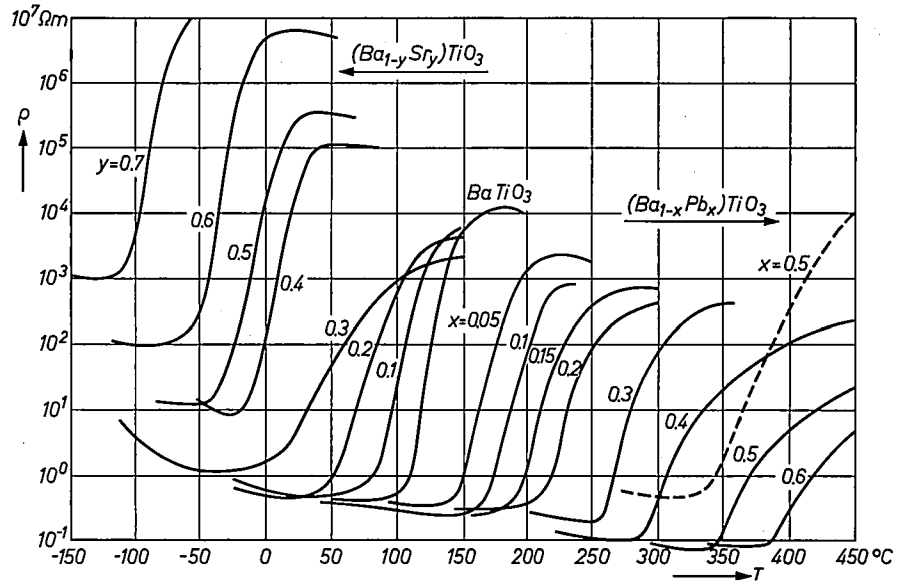


Fig. 2. Resistance-temperature characteristics of PTC thermistors sintered in the laboratory. The distribution of the curves depends on the composition of the material, indicated by the value of the subscripts x and y. The two curves for $x = 0.5$ differ only because different sintering methods were used for the two thermistors.

In this article we shall show how many of the characteristics of PTC thermistors make them particularly suitable for application as heating elements. We shall also indicate that there are certain limitations in such applications because of particular properties of the materials.

Properties of semiconducting titanate

By mixing $BaTiO_3$ with $SrTiO_3$ or with $PbTiO_3$ in the manufacture of the ceramic elements, materials are obtained with a wide range of Curie temperatures and thus a corresponding range of different PTC characteristics. Depending on the proportion of $SrTiO_3$ or $PbTiO_3$ in these mixed titanates the Curie point is shifted either downwards or upwards (fig. 2). The inclusion of 0.3 mol% of $LaTiO_3$ makes these substances

[1] E. Andrich, Properties and applications of PTC thermistors, Electronic Appl. 26, 123-144, 1965/66.

semiconducting. The sharp increase of resistance at the Curie temperature does not take place uniformly throughout the whole ceramic body but only at the grain boundaries between the densely sintered crystallites [2]. Barrier layers are formed at these boundaries, whose resistance depends not only on the temperature but also on the applied voltage. The resistance increase with positive temperature coefficient extends over a range of about 100 degrees. Outside this range the temperature coefficient is negative (*fig. 3*).

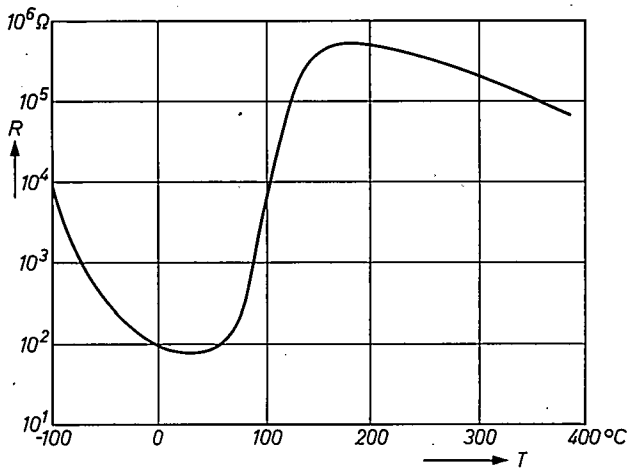


Fig. 3. Variation of the resistance R of a PTC thermistor in the temperature range from —100 to +390 °C. Outside the PTC range (50 °C to 180 °C) the temperature coefficient is negative.

It should be noted that the curves in figs. 1, 2 and 3 were obtained with a test voltage that was smaller than 2 V. At the higher voltages that are applied when the device is used as a heating element there is a significant decrease in the resistance with increasing voltage, as in the well-known voltage-dependent resistor (VDR). The behaviour of the PTC thermistor at higher voltages can be readily understood if it is assumed to have a VDR shunted across it. The resistance ratio R_{max}/R_{min} (fig. 1) of the thermistor when it is being used as a heating element may therefore be considerably smaller than that observed at the test voltage; nevertheless it can still be greater than 100:1, and this is still perfectly adequate for temperature-dependent control of the power in the heating element.

The current-voltage characteristic of the PTC thermistor, which shows the behaviour of the device when it is carrying a current, is therefore of importance; this is the non-linear curve in *fig. 4*. The dashed curve in fig. 4 indicates the dependence of this characteristic on ambient temperature or heat loss. The PTC thermistor can thus be used as a switching element that severely limits the current through the resistive load R in series with it as the ambient temperature rises from T_1 to T_2. This situation corresponds to the operating point P_2' in fig. 4. It should be borne in mind that this is usually a single switching process; when the ambient temperature goes back to T_1 the operating point does not go back beyond P_2. Generally speaking, it is only possible to switch the thermistor back to the cold resistance with the operating point P_1 if the supply of current is temporarily disconnected or the ambient temperature falls far below freezing point, so that the complete current-voltage characteristic of the thermistor rises above the load-line of the load. This means that there is little point in using a series-connected thermistor to give temperature-dependent control of the power in a resistive load. In such a circuit the thermistor works merely as a delayed-action fuse [1].

This brings us to another important characteristic of this device, the current-time characteristic. In *fig. 5* it can be seen that the relatively high initial current drops steeply after a certain time to a low residual value; this happens when the Curie temperature is reached, and the higher the initial current the earlier it occurs. Measurements of this time show that it may be as short as a fraction of a second [1]. If such short self-arresting heating-up processes are possible, and PTC thermistors with Curie points higher than 400 °C can be made (see fig. 2), then it should be possible to build heating units which reach their operating temperature immediately after being switched on and then hold this temperature constant. We shall see presently (page 174) that in fact it is not possible to heat up

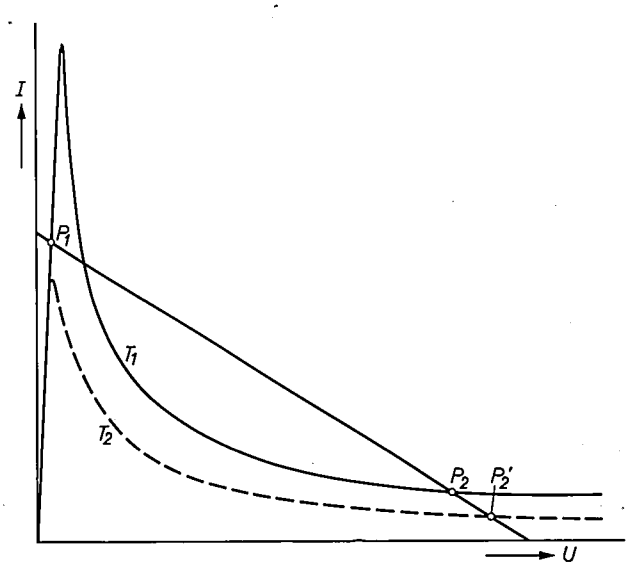


Fig. 4. Current-voltage characteristic of a PTC thermistor at different ambient temperatures T_1 and T_2. At low ambient temperature T_1 the characteristic has two stable points of intersection P_1 and P_2 with the load-line of a load R connected in series with the thermistor. At the operating point P_1 a relatively high current flows through the circuit. At higher ambient temperature T_2 there is only one point of intersection P_2' at a low current, indicating that the thermistor has severely restricted the current through the load.

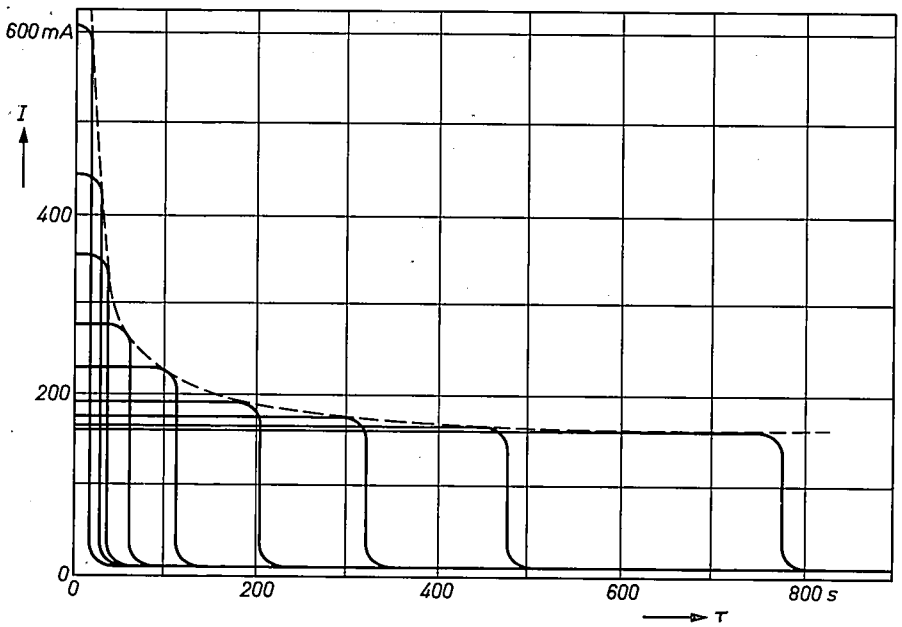


Fig. 5. Current-time characteristics of a PTC thermistor. The higher the initial current the sooner the thermistor reaches its Curie temperature, at which the current drops to a small residual value. The dashed curve represents the switching time τ as a function of the initial current I.

the entire system within fractions of a second, for we have not yet taken into account the thermal resistance of the thermistor and the material that only plays a passive part in the heating-up process, or their heat capacities. Moreover, the initial electrical power available is usually limited.

The technology of (Ba,Pb)TiO_3

The (Ba,Pb) titanates have high Curie temperatures and are therefore particularly interesting for use as heating elements. The method of producing these mixed titanates will now be considered in some detail. For technical reasons the two components $BaTiO_3$ and $PbTiO_3$ are prepared separately. The $BaTiO_3$ is pre-fired at about 1100 °C to give the reaction

$$BaCO_3 + TiO_2 \rightarrow BaTiO_3 + CO_2,$$

and the $PbTiO_3$ is pre-fired at about 950 °C to give the reaction

$$PbO + TiO_2 \rightarrow PbTiO_3$$

(a diagram of the process is given in *fig. 6*). The pre-firing processes produce a powder consisting of fine grains of doped titanates; these are ground in a ball mill for about 16 hours and at the same time thoroughly mixed in the specified barium-lead ratio. The powder is then compacted by cold pressing into pellets for the heating elements. The pellets are finally sintered at a temperature from 1300 to 1400 °C to form a bluish-grey ceramic. The volume of the pellets shrinks by about 20% during this process.

Difficulties arise during sintering since PbO evaporates readily at these high temperatures. The PbO also reacts with the ceramic elements usually used for substrate, boat and furnace tube in the sintering process. To avoid such reactions it is necessary to use platinum substrates in the preparation of these titanates and to use closed platinum vessels similar to those used in the production of piezoelectric ceramics containing PbO.

The sintering conditions (i.e. the temperature and atmosphere) and the preliminary treatment of the powder all have a considerable influence on the end-product of the mixture (Ba,Pb)TiO_3 (more so than with pure $BaTiO_3$ semiconductors) and very many experimental runs were therefore needed before a reproducible method of preparation was found.

The PTC heating element

Accuracy of temperature control

The operating temperature of a PTC heating element is not very much affected by a variation in operating voltage. For example, a thermistor rated at 10 V can still be operated at 20 V with no significant increase in its temperature. An example considered in conjunction

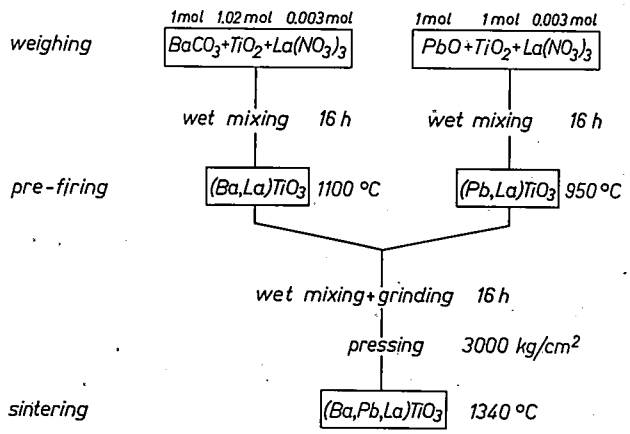


Fig. 6. Diagram of the method of preparing lanthanum-doped (Ba,Pb)TiO_3.

[2] E. Andrich and K. H. Härdtl, Investigations on $BaTiO_3$ semiconductors, Philips tech. Rev. **26**, 119-127, 1965.

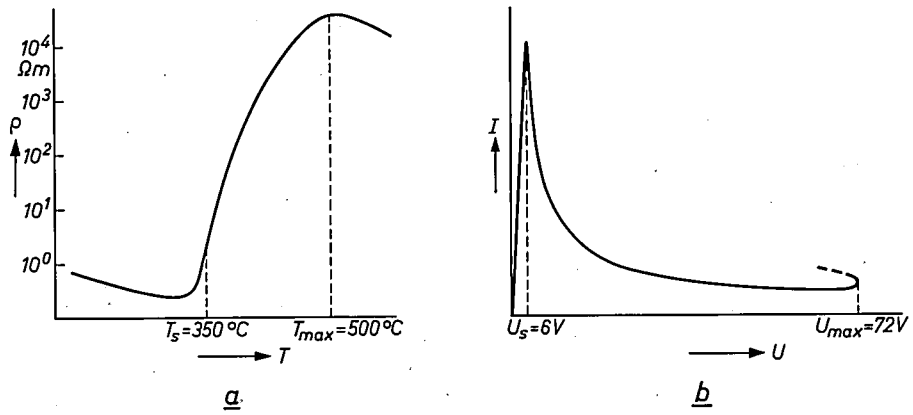


Fig. 7. *a*) Resistance-temperature characteristic and *b*) current-voltage characteristic of a $(Ba_{0.5}Pb_{0.5})TiO_3$ thermistor. The temperature interval T_{max}-T_s in (*a*) corresponds to the voltage interval V_{max}-V_s in (*b*).

with the two characteristics in *fig. 7* will make this clear. From the resistance-temperature characteristic (*R-T* characteristic) of a thermistor of $(Ba_{0.5}Pb_{0.5})TiO_3$, as shown in fig. 7*a*, we find that the temperature interval over which the temperature coefficient is positive is $\Delta T = 150$ °C, and from the current-voltage characteristic (*I-V* characteristic) shown in fig. 7*b*, we find that the corresponding voltage interval is $\Delta V = 66$ V. The ratio $\Delta T/\Delta V = 2.3$ °C/V then gives the average change in temperature per volt. If the maximum current is reached at 350 °C with 6 volts, then for 10 V we can calculate an operating temperature of about 359 °C in the thermistor and an operating temperature of 382 °C for 20 V.

On the other hand, it can be seen that a fine adjustment of the temperature in the thermistor can be attained by varying the applied voltage. Of course, a simplified calculation like this can only give an approximate picture of the behaviour of the thermistor. A complete picture would have to take into account the heat lost to the environment, the non-linearity of the characteristic and also the geometry of the thermistor. However, although the calculations can be made more complete in this way they still have to be based on certain simplifying assumptions, and this means that there is not much improvement in the agreement with the measured results.

The temperatures we are referring to are only to be found inside the thermistor, since the thermal conductivity of the thermistor material is low ($\lambda \approx 1$ W/m°C). The temperature stabilization at the surface is therefore not as perfect as one might expect from the electrical characteristics of the thermistor. Moreover, when the device is incorporated in an actual instrument, such as a soldering iron, it has to be surrounded with electrical insulation, which is usually also a poor conductor of heat. Finally, the heat-dissipating metal casing has a fairly high heat capacity and consequently a delayed response to the temperature stabilization of the thermistor. The measures taken to minimize the adverse effects of these factors on the temperature control include the use of a thin PTC element and the careful choice of insulator and metal jacket, but obviously some compromises have to be made.

The various layers — semiconductor, insulation and metal — should not be rigidly bonded, although this would give better heat transfer. However well the thermal coefficients of expansion are matched to one another, there would still be relative movement of the materials against one another because of the considerable temperature differences that can arise between them during heating-up or when heat drain is large. The heat-transferring surfaces between the different materials must therefore fit flush with each other, with no air gaps between them.

Even with an ideal design for the soldering iron, the two unavoidable, passive quantities — thermal resistance and heat capacity — are still present in the system. These quantities cause a time delay in the heat control, and a lower temperature at the bit; the temperature of the bit is also highly dependent upon the heat loss to the work. A controlled heating element does not therefore of itself give a stable temperature at the bit, independent of the heat loss to the work. This is a fact that is easily overlooked and which applies equally to soldering irons with other kinds of temperature control, such as a bimetallic strip.

Operating voltage and thermal strength

The limits of the operating voltage are not simply set by the maximum voltage that can be applied V_{max} (see fig. 7*b*). If the minimum resistance of a heating thermistor below its Curie point is 1 Ω, it will absorb 1 kW of power when the operating voltage is still only 32 V, which means that the element heats up in less than 1 second. Owing to the thermal contact with the heat-dissipating metal jacket, this gives rise to a high temperature gradient in the semiconductor which can cause the ceramic material to crack. To avoid this the initial power when the element is switched on must not be greater than 300 to 400 W per element; this value corresponds to a heating element with a surface area of 1 cm².

If, on the other hand, we apply a lower voltage and wait until the thermistor and the system to be heated have reached an almost steady temperature near the

Curie point, we can then raise the voltage nearly to the limit of the breakdown voltage, i.e. to about 70 volts in our case.

It has also been found possible to make $(Ba,Pb)TiO_3$ semiconductors that can be operated at the mains voltage. Laboratory models made from semiconducting $(Ba_{0.5}Pb_{0.5})TiO_3$ withstand fields up to 600 V/mm without breakdown. The *R-T* characteristic of these thermistors shows a resistance increase above the Curie temperature of greater than 10^5. Unfortunately certain other features of these thermistors have made them unsuitable so far for certain applications such as soldering irons. In particular, these thermistors usually have too low a resistance below their Curie temperature: 5 Ω at 1 mm electrode spacing and 1 cm^2 cross-section. At an operating voltage of 220 V this would require an initial power consumption of almost 10 kW, which is definitely too high for these semiconductors. On the other hand, if a high-resistance material with a ϱ of several tens of Ωm is used, the negative temperature coefficient will have a marked effect between room temperature and the high Curie temperature. For example, if a $(Ba,Pb)TiO_3$ thermistor has a resistance of 300 Ω just below a Curie temperature of 340 °C, its resistance at room temperature will be as much as ten times higher, i.e. about 3000 Ω [3]. This behaviour increases the heating-up time, so that this is no longer just a few seconds.

Another feature that causes difficulties, particularly in PTC thermistors supplied from the mains, is the non-uniform distribution of the voltage along the current path. Because of the low thermal conductivity of this ceramic material there is a fairly high temperature gradient inside the thermistor (*fig. 8*, lower curve), since the heat is only conducted away via the contact surfaces. As can be seen from fig. 1, the resistance in the PTC region increases more or less exponentially with the temperature. When it is drawn on a linear scale the resistance curve looks like the one shown in fig. 8. Consequently most of the voltage drop appears across a fraction of the semiconductor thickness inside the material. Most of the heat is generated here, and this has the result that the temperature gradient increases still further in the heating process, which in turn leads to a further narrowing of the zone where the heat is generated: the whole device is unstable until almost all of the voltage drop is confined to a thin layer inside the thermistor [4].

The same instability is found in the series arrangement of two similar PTC thermistors; when a voltage is applied, the one that was initially a little warmer or slightly better insulated heats up while the other remains cold. Bearing in mind that the abrupt increase of resistance takes place at the grain boundaries, the thermistor can be thought of as a series arrangement of very thin layers of PTC resistance. Since the greater part of the voltage concentrates at a few grain boundaries, very high fields will arise, particularly with mains operation: there will be local overheating and finally the ceramic will fracture because of mechanical stresses at the places where the temperature gradients are steepest. Even with a thickness of one millimetre these elements can still split into two layers. The investigations of these effects have not yet been completed.

Because of these effects the application of mains-operated PTC devices appears at present to be limited

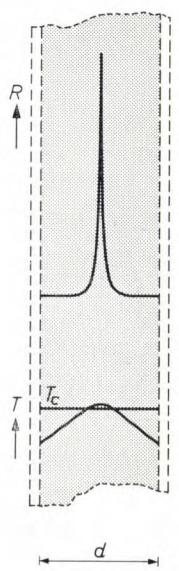


Fig. 8. Temperature curve (lower curve) and resistance curve (upper curve) in a PTC heating thermistor along the current path for the same heat loss on left-hand and right-hand sides. *d* thickness of the thermistor.

to cases in which the initial power is limited and the required heat dissipation per unit surface area is not too great. It will be shown below that such PTC devices are however suitable for use in domestic hotplates and that they are more suitable for this application than the conventional heating elements in spite of the various difficulties we have noted.

Sensitivity to ambient atmosphere; contacts

At operating temperatures above 250 °C the PTC thermistor is highly sensitive to the surrounding atmosphere presumably because at higher temperatures there is a significant diffusion of oxygen in the semiconductor. If it reaches such temperatures when it is carrying a current and it is kept in a reducing atmosphere or in a vacuum, it will soon lose its positive temperature

[3] The resistance of low-resistivity thermistors ($\varrho \approx 0.1$ Ωm) at room temperature, on the other hand, is only about 2.5 times higher than the resistance minimum below the Curie point.

[4] The NTC resistor behaves in the opposite way; when a voltage is applied, the *current* becomes locally concentrated inside the device and can even form a breakdown path. This is discussed in E. Andrich and P. L. Gillessen, Eigenschappen en toepassingsmogelijkheden van PTC-thermistors, Polytechn. T. **E 19**, 592-597, 613-619, 1964 (in Dutch).

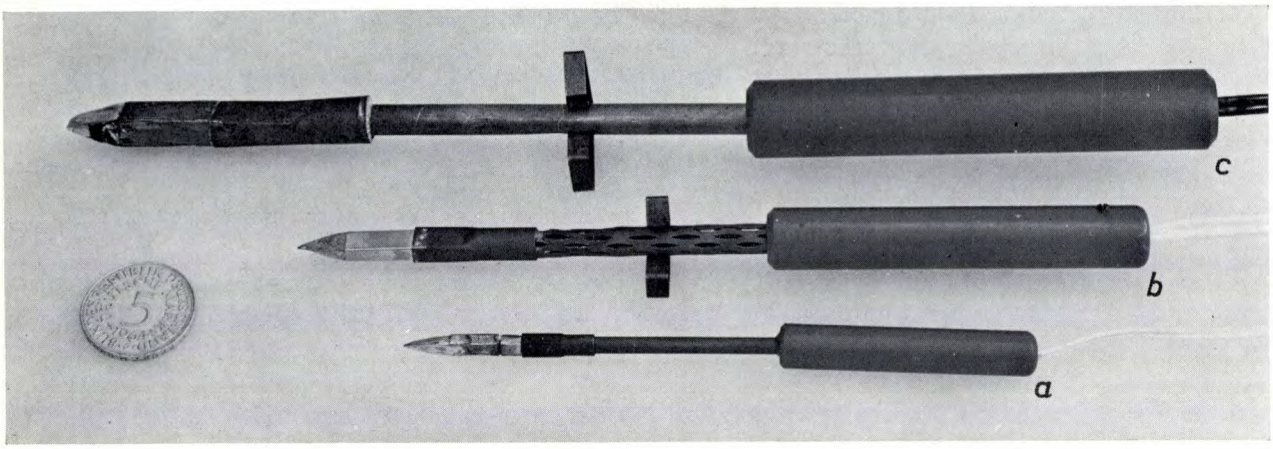


Fig. 9. Three laboratory models of soldering irons with thermistor heating for 5-20 V operating voltage. The maximum powers are *a*) 23 W, *b*) 50 W, and *c*) 120 W.

coefficient because of reduction, greatly exceed its Curie temperature and be destroyed by the avalanche-like increase in current. On the other hand, at the same current in an oxidizing atmosphere such as air the semiconductor does not lose its electrical characteristics. After slow cooling following sintering, or after tempering in air at about 550 °C, it reaches the peak of its PTC characteristic and then its characteristics remain stable, even after months of use, provided air is still used as the surrounding atmosphere. This has been verified in life tests of several months on semiconductors of the composition $(Ba_{1-x}Pb_x)TiO_3$, x having values up to 0.6. The only sign of ageing observed was at the metallic contacts, which oxidize slowly. Noble metals cannot be used for the contacts of PTC thermistors because barrier layers at the surface of the thermistor give rise to contact resistances which in some cases can be quite large. Experience has shown that the contact metal that ensures a good ohmic connection must have a certain affinity to oxygen. Although this implies a gradual oxidation of the contact at high operating temperatures and hence a deterioration of the connection, experiments in the laboratory have shown that a life of more than 3000 hours can nevertheless be achieved with vacuum-evaporated nickel-chrome contacts.

Soldering irons and hotplates using PTC heating elements

Soldering irons using PTC heating elements have been made in the laboratory in various versions for different power ratings; three of them can be seen in *fig. 9*. The design is the same in all cases; a rectangular $(Ba_{0.5}Pb_{0.5})TiO_3$ pellet 1 mm thick with vacuum-evaporated metal electrodes is placed between two metal strips which act as the current leads (see *fig. 10*). The electrical insulation is now provided by mica on both sides, and the whole assembly is placed in a copper tube. The end of the tube tapers to a point.

The soldering irons are rated for an operating voltage that may vary from 5 to 20 V; at a voltage of 10 V the medium sized soldering iron (*b* in fig. 9) requires a warming-up power of 100 W, which brings it to the required operating temperature within 10 seconds. The maximum power consumption, e.g. for soldering to a chassis, is 50 W, and when the iron is idle (i.e. not being used for soldering but switched on) it draws 10 W. The temperature at the bit is then about 300 °C. The other two versions shown, *a* and *c*, draw maximum powers of 23 W and 120 W respectively and the idling powers are 7 W and 20 W. A heating power of up to 30 W/cm² can be obtained from these thermistors with careful design. The life of the irons is determined by the gradual oxidation of the contacts mentioned above, but as we noted earlier is greater than 3000 hours.

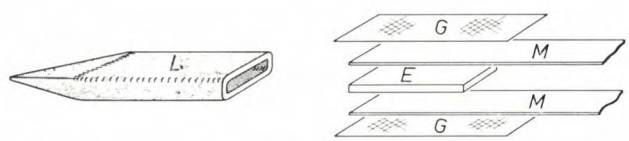


Fig. 10. Design of the soldering iron. *E* PTC heating element. *M* metal strips serving as current leads. *G* mica. *L* bit.

There are also temperature-controlled soldering irons of a different design, in which the control is effected by means of a bimetallic strip, a thermocouple or the ferromagnetic Curie point. These have the same good features as the soldering iron with thermistor heating: fast heating-up, automatic compensation of high heat losses and a low idling temperature. An advantage of the PTC element which these others do not possess is that all these useful functions are characteristic of the heating element itself; moreover they give continuous regulation, they have no switching contacts and thus do not cause interference, and only two leads from the current source are required.

A second application of PTC heating elements, now the subject of experiments in the laboratory, is a domestic hotplate fitted with mains-operated thermistors. These, however, are still in the laboratory stage. *Fig. 11* shows a laboratory model and the two elements that heat the plate. Since these heating elements automatically limit their own temperature, the hotplate can be mass-produced in an inexpensive thermoplastic housing whose resistance to deformation can be guaranteed only up to a specified temperature. If a resistive heating element without external temperature control is used, this is only possible if the element is given a sufficiently high resistance, but it then takes a relatively long time for the plate to reach the required temperature. *Fig. 12* shows that the use of a thermistor permits faster heating up. Thermistor heating is here compared with the maximum permissible ohmic heating of the same system, i.e. with a constant power, corresponding to the steady-state power of the thermistor. It can be seen that the temperature of the thermistor-heated plate reaches 90% of its final temperature of 120 °C in slightly more than 4 minutes, whereas the plate with ohmic heating takes 25 minutes to reach this temperature. Since the insulation has to be better on account of the higher voltage and the plate has a much greater heat capacity, it takes longer to heat up than the thermistor-heated soldering iron and its control action is not nearly as good; nevertheless the principle again has some definite advantages in this application.

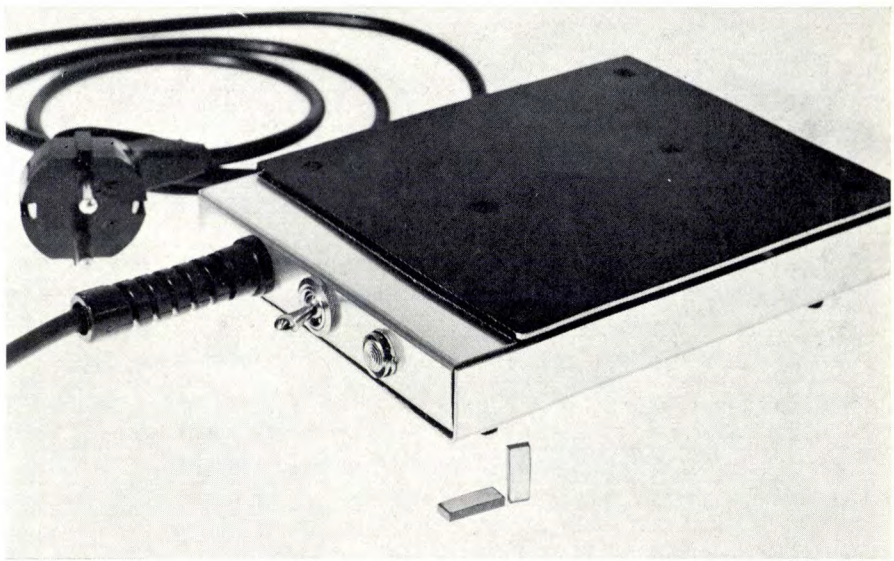


Fig. 11. Laboratory model of a domestic hotplate for mains-voltage operation. Continuous load 33 W, plate temperature 120 °C. The plate is heated by two PTC thermistors (shown in the foreground).

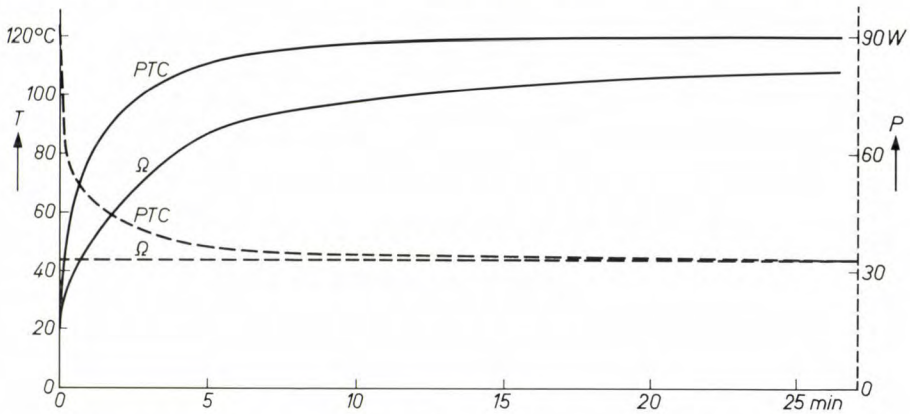


Fig. 12 Temperature variation (solid curves) and power consumption (dashed curves) of the hotplate after being switched on. *PTC*: Temperature variation and power consumption when heating with PTC thermistors. *Ω*: Temperature variation for constant power consumption, corresponding to ohmic heating.

In conclusion we can say that the experience gained in our laboratory indicates that the PTC thermistor is clearly a useful new type of self-regulating heating element. It will probably be most suitable where low heating powers, i.e. up to about 100 W, are required. Thermistors for low-voltage operation with working temperatures up to 400 °C can already be made with readily reproducible characteristics.

Summary. There are several interesting applications as self-regulating heating elements for PTC thermistors, which are ceramic semiconductors that give a sharp increase in resistance above their crystallographic transition temperature. When a voltage is applied they heat up quickly, drawing a high current which is then severely restricted when the Curie temperature is reached. The thermistor then automatically regulates its power consumption to match the heat loss. The device gives continuous control, and acts as heating element, temperature sensor and power controller at the same time. It can be operated at twice the rated working voltage with no significant change in its temperature. These devices are particularly useful and economical for low-voltage operation (between 6V and 40 V) and for powers lower than 100 W. Impressive results have been obtained with thermistor-heated soldering irons. They are ready for use within 10 seconds, and if required can increase their power consumption by five times to compensate for the heat drain through the soldering iron to the work. Other possible applications are under investigation, including a domestic hotplate for mains operation and an instrument for measuring thermal conductivity.

Electrostatic motors

B. Bollée

The magnetic forces between moving electric charges, which underlie the operation of all present-day electric motors, are much smaller than the electrostatic forces between the same charges. As electrostatic motors are based on the second effect it appears surprising, at first sight, that they are not in common use. This article explains why, under ordinary conditions, the electrostatic motor is at a disadvantage compared with the electromagnetic motor. The most likely area of application appears to lie in the construction of very small motors. It is shown that, by using the precision techniques available in a modern laboratory, it is possible to make various types of electrostatic motors with torques of practical interest.

Introduction

One aspect of the present trend towards miniaturization is the search for methods of making smaller electric motors which are still able to deliver a useful mechanical power output. If the motor is used in an apparatus with a built-in energy source (dry cells, storage battery, energy paper [1], solar cell), then the efficiency of the motor is of great importance as this will determine the size and useful life of the source. In assessing the performance of miniaturized motors the power per unit volume and the efficiency are the two most important criteria. In the case of the traditional electromagnetic motor with all its modern variants, both the power per unit volume and the efficiency are reduced when the same design is produced in a scaled-down version. This comes about because the energy dissipation in the magnet coils relative to the electric power consumed becomes greater as the motor becomes smaller. The operation of the *electrostatic* motor is based on the forces which electric fields exert on electric charges. It contains no coils, and the dissipation is very low, being that due to charge transport in an arrangement of short electrodes connected in parallel, and to dielectric losses. It is therefore to be expected that the ratio of dissipation to power consumption in a very small electrostatic motor will compare favourably with that in an electromagnetic motor of the same size. With this idea in mind we instituted a fresh investigation into the possibilities of making electrostatic motors.

Like electromagnetic motors, electrostatic motors can be synchronous and asynchronous and we have studied both types. A brief outline of the principles of operation is given below, followed by a more detailed discussion of particular aspects.

Principles of the synchronous and asynchronous electrostatic motor

The synchronous motor we considered is shown in its most elementary form in *fig. 1*. It is a variable (rotary) capacitor with a square-wave voltage applied across the plates. When the motor is running at the correct speed (i.e. synchronously), the rotor turns half

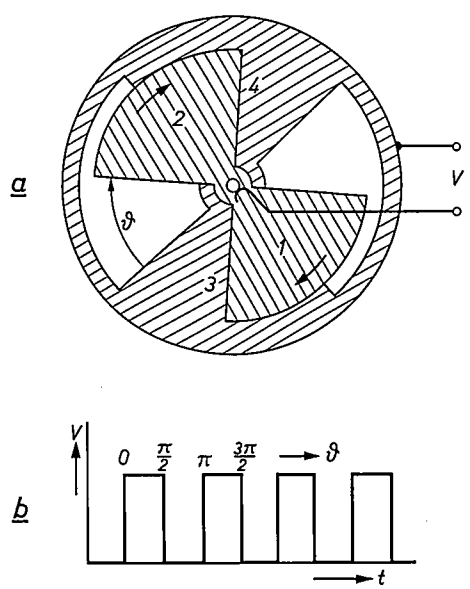


Fig. 1. The principle of the synchronous electrostatic motor is illustrated by a variable (rotary) capacitor connected to a voltage of rectangular waveform. Between $\vartheta = 0$ and $\vartheta = \pi/2$ there is a voltage between the rotor and the stator. The rotor blade *1* is then attracted by the stator blade *3*, and *2* by *4*. In the next quarter-revolution the voltage is zero and the rotor continues turning under its own inertia.

Ir. B. Bollée, formerly with Philips Research Laboratories, Eindhoven, is now in charge of the University Workshops in Utrecht.

[1] P. A. Boter and M. D. Wijnen, Energy paper, Philips tech. Rev. **28**, 298-299, 1967.

a revolution in one cycle of the voltage. In the quarter-revolution when the rotor and stator blades are approaching each other there is a voltage between them and they attract each other. During the next quarter-revolution the voltage is zero, and the rotor continues to rotate by its own inertia. Although the torque in one period shows a marked fluctuation, this motor is synchronous in the sense that it rotates in step with the supply voltage.

The choice of electrode configuration and number of electrodes can be varied considerably. *Fig. 2* shows a small motor we have made which has sixty axially oriented electrodes.

The operation of the *asynchronous* electrostatic motor is based on the fact that a cylinder of a material which is slightly electrically conducting (a "poor insulator") is subjected to a torque when placed in a rotating electric field. This is illustrated schematically in *fig. 3*. When a voltage is applied to the capacitor and the capacitor is rotated about the axis of the cylinder *A*, a torque is exerted on the cylinder. The explanation is that the field gives rise to induced charges on the cylinder. Owing to the resistance these lag behind the rotating field. Consequently the field exerts a torque on the charges and hence on the cylinder. A cylinder made of a material which is not electrically conductive but which shows dielectric hysteresis also experiences a torque in a rotating field.

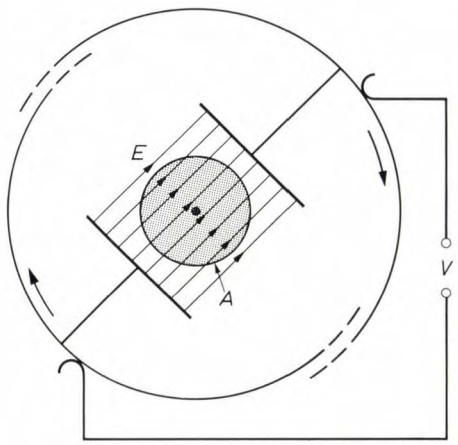


Fig. 3. The operation of the asynchronous electrostatic motor depends on the fact that a cylinder *A* which is slightly conductive (or shows dielectric hysteresis) experiences a torque when placed in a rotating electric field *E*.

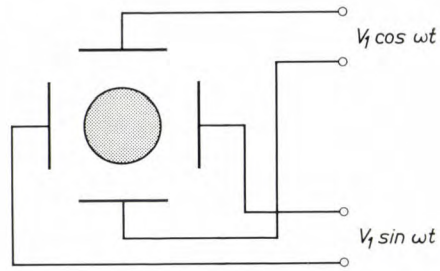


Fig. 4. An arrangement of stationary electrodes and two alternating voltages differing 90° in phase, used for generating a rotating field.

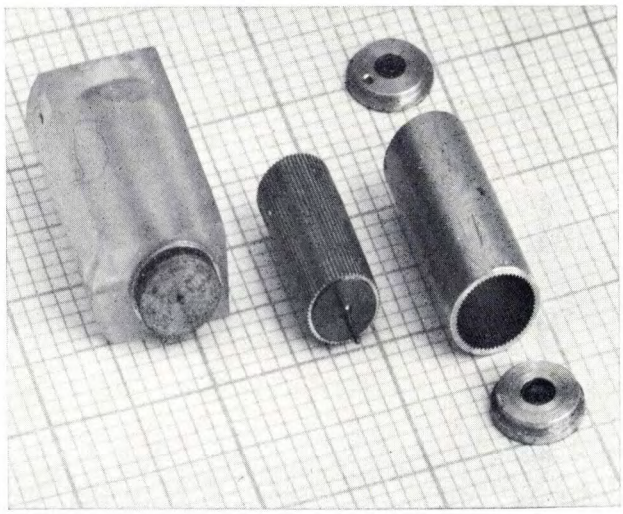
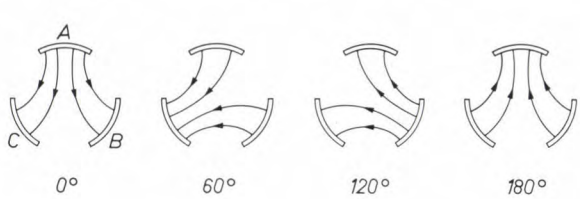


Fig. 2. An electrostatic synchronous motor with 60 electrodes. Length of the rotor 10 mm, diameter 4.5 mm, gap between stator and rotor electrodes 10 μm. The rotor and stator are hollow aluminium cylinders in which rectangular grooves have been cut. The shaft runs in sapphire bearings. The maximum power is about 100 μW at 220 V and 50 Hz.

In order to use this effect to make an electric motor, the rotating field must of course be obtained by electrical rather than mechanical means. This can be done by using phase-shifted voltages on a number of stationary electrodes. The simplest arrangement is that shown in *fig. 4*, where two pairs of electrodes are connected to voltages which differ in phase by 90°. An arrangement which we have actually constructed is shown in *fig. 5*. It consists of three electrodes connected to three different voltages with a mutual phase difference of 120°.

The motors shown in figs. 4 and 5 are called bipolar;



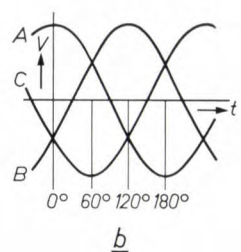


Fig. 5. A configuration of three electrodes for connection to a three-phase voltage.
a) Successive configurations of the electric field corresponding to phase increments of 60°.
b) The voltages on the three electrodes as a function of time.

the fields generated by the electrodes are such that the rotor will have one pair of poles. If a motor is built with a ring of N electrodes, with the successive electrodes energized by voltages of successive phases, then the number of pole-pairs is $p = N/m$, where m is the number of phases.

The torque on the cylinder in fig. 3 is present because the rotating field has an angular velocity with respect to the cylinder. When the motor is operating, the angular velocity of the rotor must be smaller than that of the rotating field (i.e. the operation is "asynchronous"), and the torque and efficiency depend on the ratio between these angular velocities.

The performance that can be obtained from an electrostatic motor is limited by the *breakdown* that occurs if too high a voltage is applied to a given motor (or, for a given voltage, if the distance between the plates is made too small). A "normal" *breakdown field strength* (in air of atmospheric pressure with a gap of 1 mm) is about 3×10^6 V/m, so that for a synchronous motor with a plate spacing of 100 μm the voltage must remain below 300 V. One way of raising this limiting value is to operate the motor in a vacuum, in which field strengths of 3×10^8 V/m are possible. The breakdown field strength also depends on the distance between the electrodes. As the electrode spacing decreases, the breakdown field strength in air at atmospheric pressure begins to increase sharply at about 7.5 μm, and at 2 μm it reaches a value of 1.7×10^8 V/m, which is close to the value in a vacuum [2].

Earlier investigations of electrostatic motors

Although the principles underlying the electrostatic motor, both synchronous and asynchronous, are simple and have long been known, remarkably little attention has been paid in the past to the possibility of making motors of this type. The earliest publication known to us that deals with the synchronous electrostatic motor is a thesis by Petersen, which appeared in 1907 [3]. The next study devoted to it, by Trump [4], did not appear until 1933. He was considering machines of high power and high voltage *in vacuo*, and built a 55-watt motor for experimental purposes. He did not investigate small motors. Thereafter the synchronous electrostatic motor was forgotten.

The asynchronous motor was investigated as long ago as 1893 by Arno [5]. The phenomenon of dielectric hysteresis in insulating material (glass, ebonite, etc.) was already known and he used it to build a motor operating at 3800 V, 40 Hz, which reached a speed of 250 rev/min. After the turn of the century there followed a period in which the principle was widely used for studying the dielectric hysteresis of materials [6], but not for making motors. It was not until 1931 that the asynchronous motor again came up for discussion, in a thesis by Strobl [7]. He checked his theoretical calculation on a 2000 V motor built for that purpose, but an operational motor was never constructed. After this the asynchronous motor also fell into oblivion.

As previously mentioned, there is now a need for miniature motors in a wide variety of devices ranging from toys to satellites. The electrostatic motor has features that make it of interest for such purposes, especially as many new materials and technologies have become available since the thirties which could be turned to good use in the manufacture of such small motors. It will be apparent from the work discussed below that it is in fact possible to produce electrostatic motors with a volume of no more than a few cubic centimetres which have torques of practical interest (of the order of 5 μNm, i.e. about 50 mg-force × cm).

The synchronous electrostatic motor

Our investigations into the synchronous electrostatic motor relate mainly to obtaining an optimum design. The essential features to be considered are the capacitance variation (and hence the form and arrangement of the electrodes) and the waveform of the alternating voltage (sinusoidal, square-wave, etc.). First the torque and power are calculated as a function of the capacitance variation and the voltage. Initially the losses are left out of account, and are dealt with in a separate section. Finally we consider the question of starting, which is always a problem with synchronous motors.

Calculation of torque and power

In a synchronous electrostatic motor the capacitance C of the capacitor formed by stator and rotor is a function of the angle of rotation ϑ of the rotor. Be-

---

[2] J. Biermanns, Hochspannung und Hochleistung, Hanser, Munich 1949, p. 74; W. O. Schumann, Elektrische Durchbruchfeldstärke von Gasen, Springer, Berlin 1923, p. 25.

[3] W. Petersen, Elektrostatische Maschinen, Thesis, Technical University of Darmstadt, 1907.

[4] J. G. Trump, Vacuum electrostatic engineering, Thesis, Massachusetts Institute of Technology, 1933.

[5] R. Arno, Über ein rotirendes elektrisches Feld und durch elektrostatische Hysteresis bewirkte Rotationen, Elektrotechn. Z. **14**, 17-18, 1893.

[6] V. von Lang, Versuche im elektrostatischen Drehfelde, Wiener Berichte **115**, IIa, 211-222, 1906.
A. Lampa, Über Rotationen im elektrostatischen Drehfelde. Ein Beitrag zur Frage der dielektrischen Hysteresis, Wiener Berichte **115**, IIa, 1659-1690, 1906.
E. von Schweidler, Die Anomalien der dielektrischen Erscheinungen, Handbuch der Elektrizität und des Magnetismus, Barth, Leipzig 1918, part I, pp. 232-261 (with an extensive list of references up to the end of 1912).
J. Schaffit, Über ein neues elektrostatisches Drehfeld-Voltmeter von Prof. Peukert, Thesis, Brunswick 1913.
P. Lertes, Untersuchungen über Rotationen von dielektrischen Flüssigkeiten im elektrostatischen Drehfeld, Z. Physik **4**, 315-336, 1921.

[7] K. Strobl, Der elektrostatische Asynchronmotor, Thesis, Technical University of Vienna, 1931.

tween the stator and rotor a voltage V is applied that varies periodically with time. When the motor is turning, ϑ is a function of time and hence C is also. Electrical power is then converted into mechanical power. This power is calculated as follows.

If I is the current supplied through the terminals, then the electrical energy supplied in a time interval dt is given by:

$$P_e dt = IV dt.$$

The (varying) charge on the electrodes is CV, and the change in this is:

$$I dt = d(CV) = C dV + V dC,$$

so that

$$P_e dt = CV dV + V^2 dC.$$

Part of this energy is stored in the electric field. The instantaneous field energy W_e is $\frac{1}{2} CV^2$, and this increases in the time dt by an amount

$$dW_e = CV dV + \tfrac{1}{2} V^2 dC.$$

The rest of the energy supplied,

$$P_e dt - dW_e = \tfrac{1}{2} V^2 dC,$$

is delivered to the rotor as mechanical energy $P_m dt$. The mechanical power is thus:

$$P_m = \tfrac{1}{2} V^2 dC/dt. \quad \ldots\ldots\ldots \quad (1)$$

The torque T acting on the rotor is:

$$T = P_m dt/d\vartheta = \tfrac{1}{2} V^2 dC/d\vartheta. \quad \ldots \quad (2)$$

Between T and P_m there exists the relation:

$$P_m = \omega_m T, \quad \ldots\ldots\ldots\ldots \quad (3)$$

where $\omega_m = d\vartheta/dt$ is the angular velocity of the rotor.

Let the number of electrodes be N. The angle between two electrodes, the "electrode pitch", is then $2\pi/N$. In the theory of electric motors it is the usual practice to denote the position of the rotor by the variable $\Theta = N\vartheta$; this quantity Θ varies by 2π when the rotor moves on by one electrode. The capacitance C is a periodic function of Θ with a period 2π.

Suppose now that the motor rotates uniformly (e.g. due to the flywheel effect of the rotor) and that the rotor moves on by one electrode in one period of the supply voltage. In that case by definition we call the speed "synchronous" (even in those cases in which the motor is unable to supply power at this speed); other speeds will be discussed later. The synchronous angular velocty ω_m equals ω_s/N, where $\omega_s = 2\pi f_s$ is the angular frequency of the supply voltage. C is then a periodic function of time with the same period as the supply voltage. From (1) and (2) it can be seen directly when a positive power and torque can be obtained. This occurs,

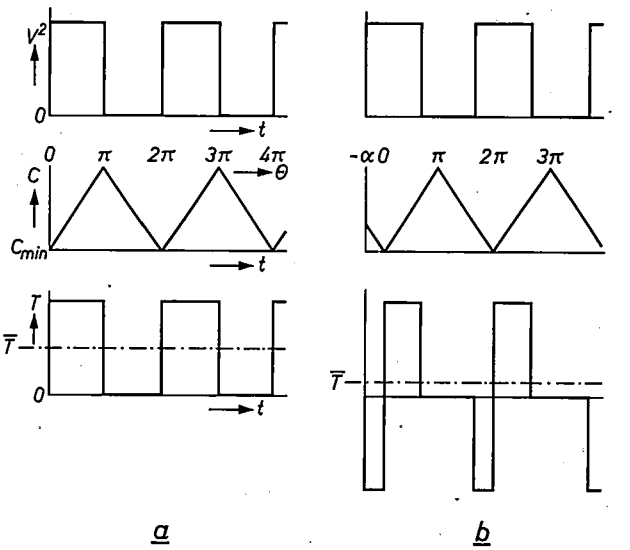


Fig. 6. A rectangular voltage $V(t)$ and a triangular variation of capacitance $C(\Theta)$ are shown. The corresponding torque T, which acts on the rotor when it is turning at the synchronous speed, is derived from equation (2). The angular position of the rotor is given by $\Theta = \omega_s t - \alpha$. In (*a*) $\alpha = 0$, in (*b*) $\alpha \neq 0$. In case (*a*) the average torque T is maximum.

for instance, if we have a rectangular voltage with $V \neq 0$ in that part of the period in which dC/dt is positive, and $V = 0$ where dC/dt is negative. This is illustrated in *fig. 6a* for a triangular variation of capacitance.

In general the rotor will lag in phase behind the voltage. The angle will be given by:

$$\Theta = \omega_s t - \alpha$$

(see fig. 6*b*). For part of the cycle the torque and power are then negative, and the average torque and power are smaller. The angle α assumes a value such that the average torque is equal to the torque required by the load (*fig. 7*). If the required torque is greater than the maximum torque of the motor, the motor falls out of synchronism and stops.

In the case of a square-wave voltage, if the parts of the cycle for which $V^2 \neq 0$ and $dC/dt > 0$ are of equal length, then the average power is a maximum

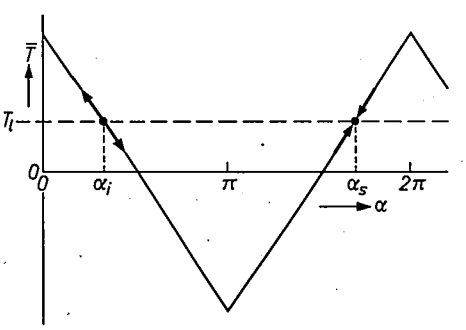


Fig. 7. The average torque $\overline{T}$ as a function of the angle α by which the rotor lags behind the voltage. For a load with a torque T_l, the value α_l is unstable while α_s is stable.

when these parts coincide (C_{min} coincides with the beginning of the square wave, C_{max} with the end, as in fig. 6*a*). The maximum average power in that case is:

$$\overline{P}_{max} = \overline{\tfrac{1}{2} V^2 \, dC/dt} = \tfrac{1}{2} V_0^2 (\omega_s/2\pi)(C_{max} - C_{min}), \quad (4)$$

where V_0 is the peak square-wave voltage. The maximum average torque is:

$$\overline{T}_{max} = (N/4\pi) V_0^2 (C_{max} - C_{min}). \quad . \quad . \quad . \quad (5)$$

The following points may be noted in connection with the equations:

1) P and T are proportional to V^2. This implies that the sign of the voltage is irrelevant. To the motor the voltage shown in *fig. 8b* is equivalent to the square-wave voltage of fig. 8*a*. The motor cannot operate with a voltage as shown in fig. 8*c*. With the voltage of fig. 8*a* the motor runs in both directions equally well.

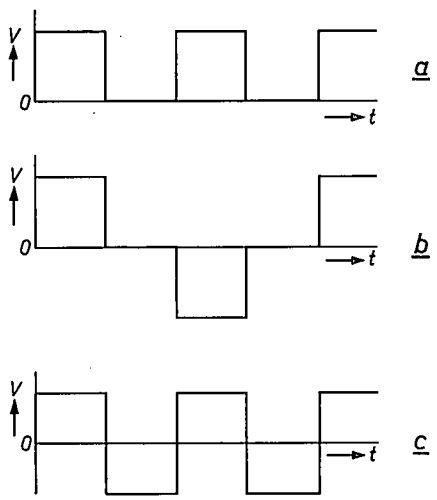


Fig. 8. As the torque and power of the synchronous electrostatic motor are proportional to the square of the voltage, the voltage forms (*a*) and (*b*) are equivalent, whereas the motor cannot run with the voltage waveform (*c*).

2) Given the magnitude V_0 and frequency f_s of the square-wave voltage, and the capacitance difference $C_{max} - C_{min}$, the maximum average power $\overline{P}_{max}$ is independent of the number of electrodes N. The torque is proportional to N, and the angular velocity is inversely proportional to N. A motor with many electrodes is thus slow but powerful. In practice, there is an upper limit to N. If we start distributing more and more electrodes over a given rotor surface area, $C_{max} - C_{min}$ will gradually decrease (see *fig. 9*), and the expected increase of torque will not be attained. We shall return to the relation between $C_{max} - C_{min}$ and N later.

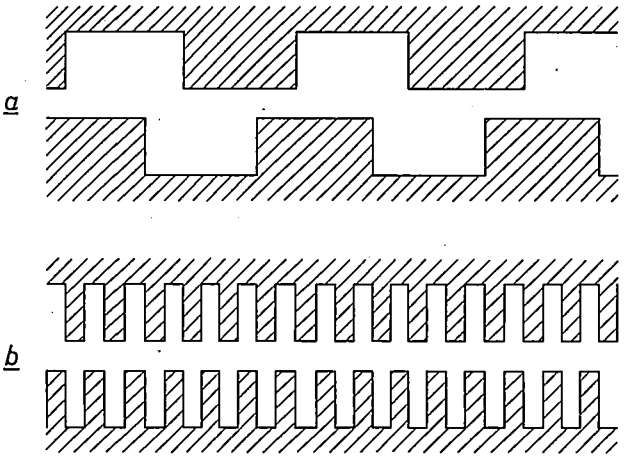


Fig. 9. *a*) As long as the gap width is much smaller than the electrode pitch, then, for a given rotor surface area, the capacitance variation is independent of the number of electrodes (N) into which this area is divided. *b*) If N is so large that the pitch becomes as small as the gap width, hardly any capacitance variation can be expected.

3) Design details of the motor, such as gap width, electrode height, etc., are not explicitly contained in the equations. They are accounted for in a single factor $C_{max} - C_{min}$. To determine the maximum torque in an actual case it is not necessary to know internal structural details, but it is sufficient to measure the capacitance as a function of the position of the rotor. This can be done by mounting the stator on a graduated turntable and fixing the rotor in position with respect to the frame of the turntable. Such a measurement of $C(\Theta)$ is also a check on the precision of manufacture of a motor.

The synchronous electrostatic motor compared with an analogous electromagnetic motor

Equation (4) provides a basis for comparing the electrostatic synchronous motor with variable capacitance, as discussed above, with the magnetic synchronous motor possessing variable reluctance, which is its analogue. Both are presented schematically in *fig. 10*. This comparison explains why the development of the

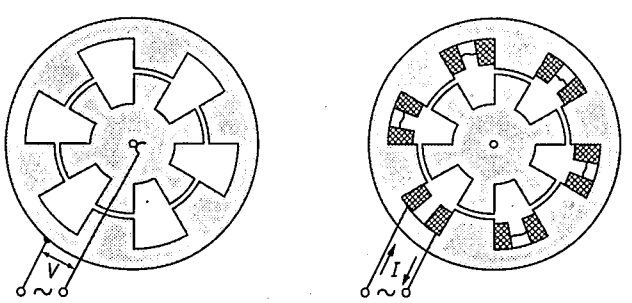


Fig. 10. An electrostatic synchronous motor with variable capacitance and a magnetic synchronous motor with variable reluctance, shown schematically to illustrate the analogy between them.

electromagnetic motor has made such great strides whereas the electrostatic motor has not yet found practical application. It also indicates the circumstances in which the electrostatic motor might well compete with the electromagnetic type.

Let us assume that C_{min} in eq. (4) can be neglected with respect to C_{max}. If the rotor turns at a constant voltage V_0, the field energy in the motor varies between zero and a maximum value $W = \frac{1}{2} C_{max} V_0^2$. Eq. (4) takes the simple form:

$$\overline{P}_{max} = f_s W. \quad \ldots\ldots\ldots \quad (6)$$

Since an equation of this form also applies to the electromagnetic motor in fig. 10, the power of the two types of motor can be compared by considering the field energy in the two cases. This is

$$W = As(\tfrac{1}{2} DE) = \tfrac{1}{2} As\varepsilon_0 E^2 \text{ (electrostatic)}, \quad \ldots \quad (7a)$$

$$W = As(\tfrac{1}{2} BH) = \tfrac{1}{2} AsB^2/\mu_0 \text{ (magnetic)}, \quad \ldots \quad (7b)$$

where A is the total area of the poles or electrodes and s the gap width.

The upper limit of the power that can be produced with a motor of a given size is determined by the practical limit of E or B provided the motor is not too small. In the magnetic case the upper limit for B, as determined by the *saturation of the magnetic material*, is roughly 1 Wb/m^2.

In the electrostatic case the power is limited by electrical breakdown. If the electrostatic motor is to rival the magnetic motor in field energy per unit volume of field space, the required field strength is $E = B/\sqrt{\varepsilon_0\mu_0}$ (cf. eq. 7a and 7b). For $B = 1$ Wb/m^2, we obtain $E = 3 \times 10^8$ V/m. At a "normal" breakdown strength of 3×10^6 V/m (see page 180) the permissible field energy (for the same A and s) is thus 10^4 times smaller than in the magnetic case. By using thin plates more field space can be accommodated in an electrostatic motor than in a magnetic-reluctance motor of the same size (see, for example, fig. 12). In spite of this advantage which may be 10 to 100 times, the maximum power per unit volume of the whole motor is still a factor of 100 to 1000 below that of the magnetic motor. Therefore the electrostatic motor will not generally be a competitor of the magnetic type.

As previously mentioned, the operation of the electrostatic motor in a vacuum gives it a much higher breakdown field strength (3×10^8 V/m) and would enable it to compete with the magnetic motor. A proposal along these lines was made by Trump in 1933 [4], but practical versions of vacuum motors have never been made.

We have previously referred to the heat generated in the coils. In *small* magnetic motors this may become comparable to the input power. The efficiency steadily diminishes with decreasing size and the power limit is no longer determined by the magnetic induction but by the necessity of dissipating the heat generated in the coils. Thus the magnetic motor steadily loses its advantage as it gets smaller.

In quantitative terms this can be seen as follows. Consider a magnetic motor of the type in fig. 10, with the linear dimension l and in which the current density is J. Let us consider the effect on the efficiency and on the maximum power output of changing all the linear dimensions in the same proportion. Assume that the magnetic circuit outside the gap is completely short-circuited; if the gap width is s, the field strength in the gap is then $(1/s) \times$ the number of ampere turns, and is thus proportional to $(1/l) \times Jl^2 = Jl$. From (6) and (7) the mechanical power delivered, P_m, is proportional to W, and hence to $As(Jl)^2 \propto J^2 l^5$. The power dissipated in the coils, P_d, is proportional to $J^2 l^3$. Thus as l decreases the power output decreases more rapidly than the dissipated power, so that the efficiency must tend to zero.

The calculation of the maximum power output per unit volume of the motor requires rather more subtle reasoning. The normal constraint is that the current density shall not exceed a given value J_m, irrespective of the dimensions. This value is set by the requirement that the heat generated inside the coil windings must be conducted to the surface without the internal temperature becoming unduly high. When the dimensions are small, however, it is not so much the heat transport from inside to outside that counts but rather the heat dissipation at the surface. For a given maximum temperature difference from the ambient air, the surface dissipation is proportional to l^2. The generated heat ($\propto J^2 l^3$) must therefore not exceed a constant times l^2; thus, if J_m is the permissible current density, then $J_m^2 l^3 \propto l^2$, or $J_m^2 \propto l^{-1}$. The maximum power output P_m (proportional to $J_m^2 l^5$, see above) is thus proportional to l^4, and the power per unit volume $P_m/l^3 \propto l$ decreases with the volume.

The reasoning given above does not apply to all types of electromagnetic motor; in particular the situation is more favourable for motors with permanent magnets. This does not alter the general conclusion that problems of heat dissipation are always encountered as a result of the generation of heat in the current conductors.

The losses in the electrostatic motor (ohmic losses in the electrodes and dielectric losses) are less severe and do not limit the size of the motor in so definite a manner. This is the basis for the assumption that ultimately miniaturization gives the advantage to the electrostatic motor. Of course, there are other limits set to the process of scaling down. For example, a gap width of 10 μm would seem to be the lower limit with present techniques. Whether the electrostatic motor would already have the advantage at this limit is a question that we leave open.

Arbitrary periodic variations of voltage and capacitance

In the preceding discussion we have for simplicity referred to a square-wave voltage and a capacitance variation of triangular form. Other voltage waveforms, such as a sinusoidal alternating voltage, possibly rectified or superimposed on a d.c. voltage, are more readily available than a square-wave voltage and are often more efficient for operating the motor. Moreover the capacitance variation is not normally triangular.

If we examine the problem in more general terms, we find that the motor can also run at speeds other than the synchronous speed. This can be shown as follows. We expand $V(t)$ and $C(\Theta)$ into a Fourier series:

$$V(t) = V_0 + V_1 \cos \omega_s t + \\ + V_2 \cos(2\omega_s t - \beta_2) + \ldots, \quad (8)$$

$$C(\Theta) = C_0 + C_1 \cos \Theta + \\ + C_2 \cos(2\Theta - \gamma_2) + \ldots. \quad (9)$$

(The phase angle of the fundamental harmonic in (8) and (9) can always be equated with zero by means of a suitable choice of the origin on the t-axis and the Θ-axis.) Assume now that the motor is running uniformly but not necessarily with the synchronous angular velocity ω_s/N. We then have:

$$\Theta = h\omega_s t - \alpha. \quad \ldots\ldots \quad (10)$$

For the synchronous speed we have $h = 1$. Eq. (10)

implies that the angular velocity ω_m of the rotor is given by:

$$\omega_m = h\omega_s/N. \quad \ldots\ldots \quad (11)$$

Substituting (8), (9) and (10) in expression (1) we find for P_m a sum of terms of the type:

$$\cos(k\omega_s t - \beta_k)\cos(l\omega_s t - \beta_l)\sin(qh\omega_s t - \delta_q),$$

with $k,l = 0, 1, 2, \ldots$ and $q = 1, 2, \ldots$. These products can be converted into sums of trigonometric functions. The arguments for the terms of the series for P_m thus obtained are sums and differences of the original arguments. When averaged over time all except the *constant term* yield the value zero. The constant term is composed of terms with "zero" frequency which arise for

$$k\omega_s \pm l\omega_s \pm qh\omega_s = 0,$$

that is when

$$h = \pm\frac{k \pm l}{q}. \quad \ldots\ldots \quad (12)$$

This implies that the motor can generally give an average power output at the angular velocities given by (11) with h substituted from (12). The velocities possible in a particular case depend on the terms occurring in the series (8) and (9). The amounts of power that can be delivered at the various possible velocities will, in general, be quite different from each other.

Let us now take as an example the simple case of a motor with a sinusoidal capacitance variation, connected to a sinusoidal alternating voltage. We have:

$$V = V_1\cos\omega_s t, \qquad C = C_0 + C_1\cos\Theta.$$

This corresponds to $k = l = 1$ and $q = 1$. The possible values for h are thus ± 2. This means that the motor can only run at twice the synchronous speed (which is not surprising since the frequency of the variation of V^2 is $2f_s$). The power output is:

$$\begin{aligned} P_m &= \tfrac{1}{2}V^2\mathrm{d}C/\mathrm{d}t \\ &= -\tfrac{1}{2}C_1V_1^2\cos^2\omega_s t \times 2\omega_s\sin(2\omega_s t - \alpha) \\ &= -\tfrac{1}{2}\omega_s C_1V_1^2(1+\cos 2\omega_s t)\times \\ &\qquad \times(\sin 2\omega_s t\cos\alpha - \cos 2\omega_s t\sin\alpha). \end{aligned}$$

In working out the product the only term that contributes to the average is the term with $\cos^2 2\omega_s t$, and the average of this is $\frac{1}{2}$. The average power is thus:

$$\overline{P}_m = \tfrac{1}{4}\omega_s C_1V_1^2\sin\alpha,$$

and the maximum average power is:

$$\overline{P}_{max} = \tfrac{1}{4}\omega_s C_1V_1^2.$$

Table I gives possible speeds and the corresponding maximum average power outputs in various cases. The first three examples are for a motor with sinusoidal capacitance variation connected to a) a sinusoidal a.c. voltage (the example worked out above), b) a sinusoidal a.c. voltage superimposed on a d.c. voltage, c) a half-wave rectified sinusoidal a.c. voltage. Example d) is for a motor with an arbitrary capacitance variation connected to a sinusoidal a.c. voltage. From (a) and (b) in Table I it can be seen that half the original speed is possible when a d.c. voltage is superimposed on the original sinusoidal a.c. voltage. *Fig. 11* shows a large demonstration model with which this and other features of the synchronous motor can be demonstrated.

From case (d) we can draw some conclusions which could be of practical importance. In the first place it can be seen that for a motor supplied with a pure a.c. voltage and which runs at the speed appropriate to this (twice the synchronous speed) *only the first harmonic* of $C(\Theta)$ is of importance. The power at this speed is proportional to C_1 and independent of the other coef-

Table I. Possible speeds and corresponding power outputs for various forms of capacitance variation $C(\Theta)$ and supply voltage $V(t)$.

	$V(t)$	$C(\Theta)$	h	$\overline{P}_{max}$
a)	$V_1\cos\omega_s t$	$C_0 + C_1\cos\Theta$	2	$\frac{1}{4}\omega_s C_1V_1^2$
b)	$V_0 + V_1\cos\omega_s t$	$C_0 + C_1\cos\Theta$	1	$\frac{1}{2}\omega_s C_1V_0V_1$
			2	$\frac{1}{4}\omega_s C_1V_1^2$
c)	$\frac{V_1}{\pi}(1 + \frac{1}{2}\pi\sin\omega_s t - \frac{2}{3}\cos 2\omega_s t + \ldots)$	$C_0 + C_1\cos\Theta$	1	$\frac{1}{4\pi}\omega_s C_1V_1^2$
	(half-wave rectified sine curve)		2	$\left(\frac{2}{3\pi^2} + \frac{1}{16}\right)\omega_s C_1V_1^2$
			3,4, ...	...
d)	$V_1\cos\omega_s t$	$\sum_q C_q\cos(q\Theta - \gamma_q)$	2	$\frac{1}{4}\omega_s C_1V_1^2$
			2/2	$\frac{1}{4}\omega_s C_2V_1^2$
			2/3	$\frac{1}{4}\omega_s C_3V_1^2$
			...	...
			2/q	$\frac{1}{4}\omega_s C_qV_1^2$

ficients. Given these supply and speed conditions it is clear that we require a design for the electrodes that will result in a capacitance variation as nearly sinusoidal as possible (which gives us case (a) again).

In the second place it can be seen that *subharmonic* speeds are possible with $h = 2/2, 2/3, 2/4, 2/5, \ldots$.

Capacitance variation; design of the motor

It follows from the theory given above that the main objective in the design of a synchronous electric motor must be to obtain the greatest possible variation of the capacitance. The capacitance variation results from the changes in the amount of overlap of the condenser

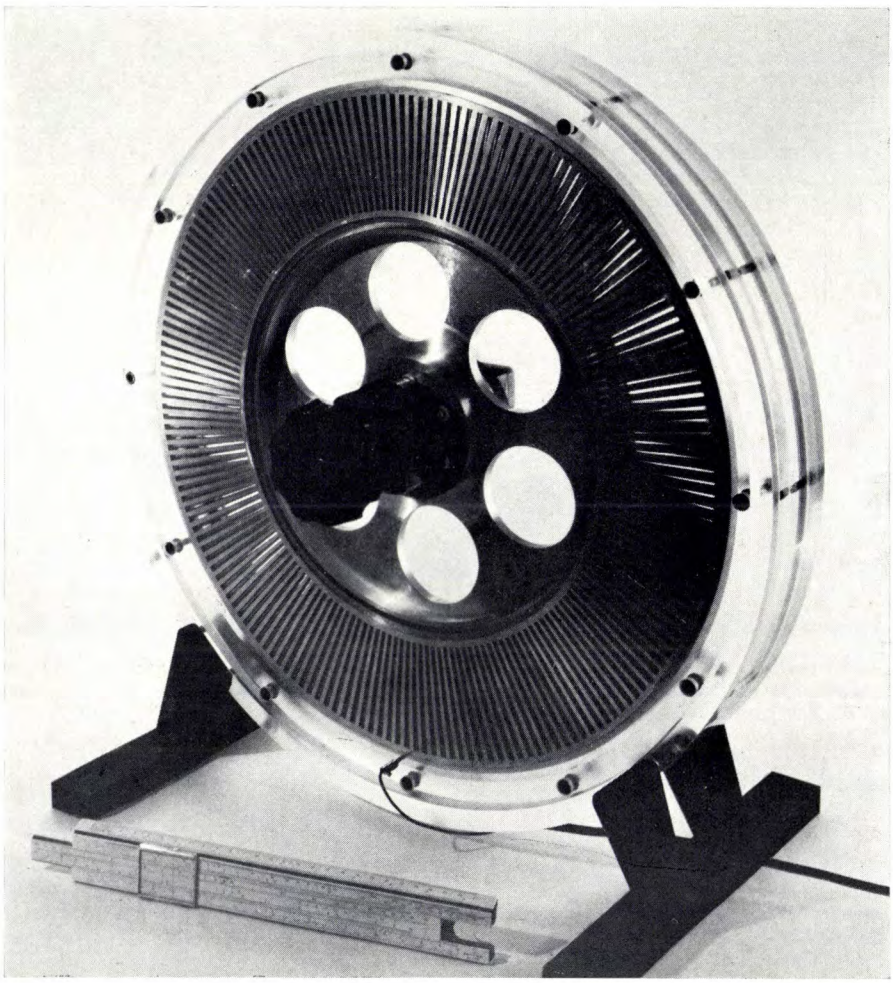


Fig. 11. This large, slowly running synchronous motor was built for demonstration purposes. It is made of methacrylate (except for electrodes and shafts) and almost all the working parts are visible. The number of electrodes (on both sides) is 180. The motor is rated for 1000 V, 50 Hz a.c. and normally runs at 5/9 rev/s. The speed can be halved if a d.c. voltage is superimposed on the a.c. voltage (see Table I, a and b). The motor is started manually. By using a stroboscope the movement of the rotor can be observed in the period just after starting, and when the load changes the resultant phase shift of the rotor with respect to voltage can be seen.

We have been able to demonstrate this with the clock motor shown in fig. 21, which can run at 1/3 and 1/6 of the normal speed. With the usual forms of electrode, however, the coefficients C_2, C_3, ... are always very small compared with C_1 (see also page 188). The corresponding powers are therefore relatively small, so that normally the use of subharmonic speeds does not seem to be an attractive proposition.

plates of the rotor and stator, which form one or more pairs of coaxial surfaces of revolution. The simplest surfaces of revolution that can be used for this purpose are the flat disc and the cylinder. The first leads to the type of "variable capacitor" shown in fig. 1, the second to the cylindrical type in fig. 2.

A large variation of capacitance calls for a large maximum capacitance, which implies a large capacitor

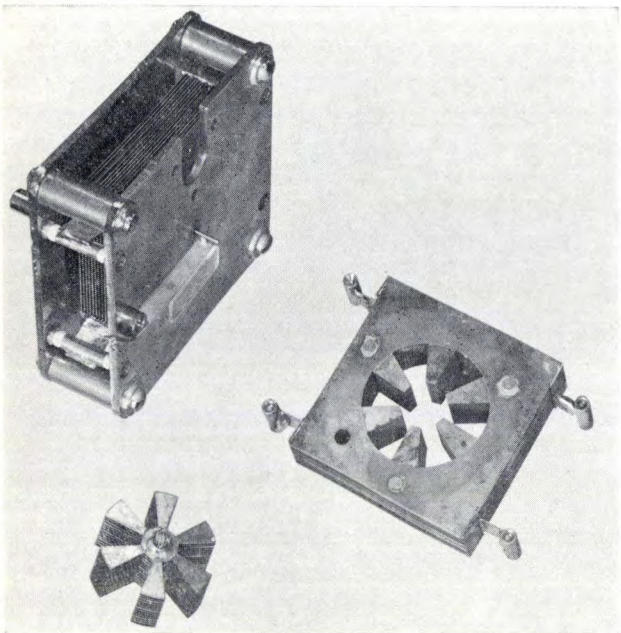


Fig. 12. A synchronous electrostatic motor with flat electrodes. The design is based on the variable capacitor used in radio engineering. The motor was made at the Philips variable-capacitor factory in Tilburg. There are 15 plates in the stator and 14 in the rotor each 300 $\pm$ 3 μm thick. The stator and rotor plates are separated by spacers and are all soldered simultaneously in a furnace. The stator plates are fixed to support bars, and the rotor plates to the shaft. After soldering the spacers are removed. The distance between rotor and stator blades is 150 μm.

The flat-plate motor in *fig. 12* was made in the Philips variable-capacitor factory at Tilburg for our experiments. Further details are given in the caption.

Because of the simplicity of manufacture and its ruggedness the cylindrical type seems the more attractive at present. In the example shown in *fig. 13* both the rotor and the stator consist of a hollow cylinder

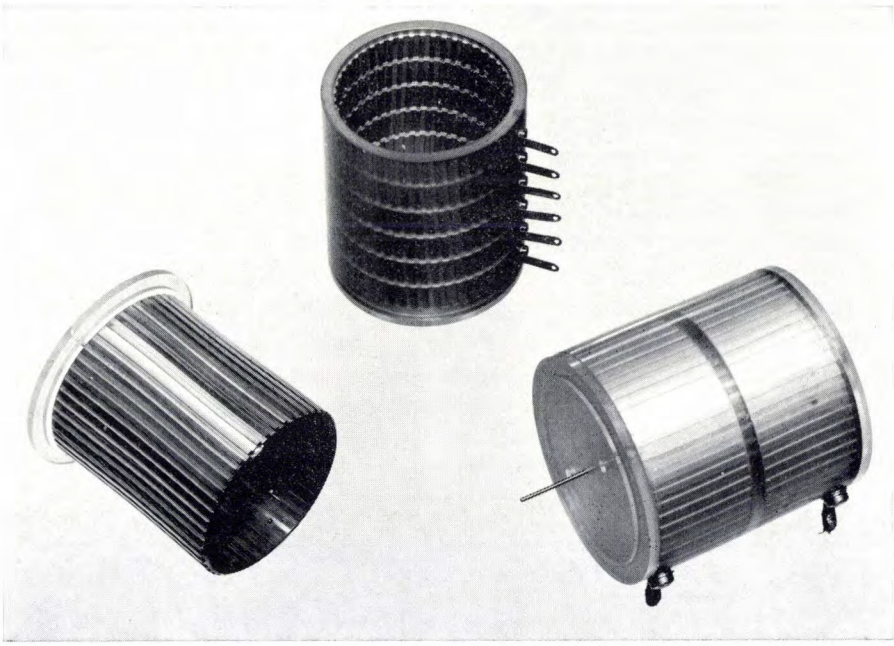


Fig. 13. A cylindrical synchronous electrostatic motor of methacrylate. The grooved surfaces are covered with a vapour-deposited film of aluminium.

surface with small spacings between the plates. To achieve this in a limited volume we have to make a large number of thin plates at small distances apart. The possibilities here are ultimately limited by the strength and stiffness of the plates, by their flatness and by the play in the rotor shaft.

We have designed and built motors both of the flat type and the cylindrical type, and some examples can be seen in figs. 12, 13 and 14.

of methacrylate in which grooves have been milled. The grooved surfaces are coated with an aluminium layer, deposited by vacuum evaporation. The teeth form the electrodes. A design of this type can also be made in solid aluminium (see fig. 2). The use of a dielectric as basic material has the advantage that the motor can be made somewhat lighter; moreover a motor of a material like methacrylate is almost completely transparent, which is useful for demonstration

purposes. It would be simpler to deposit metal strips straight on to the cylinders without first having to cut grooves in the cylinder surface. If, however, the dielectric constant of the cylinder material is much greater than 1, the result will be a smaller capacitance variation and we have not been able to find a suitable dielectric with a dielectric constant of about 1.

For the cylindrical motors in *fig. 14* we have made N very large ($N = 100$). In spite of this large value for N the capacitance variation was made reasonably large by having several pairs of electrodes (instead of one) in the radial direction.

in parallel. The power can be restored to the original value by doubling the motor supply voltage (see eq. 1). As this only restores the field strength in each capacitor to its original value, it does not increase the danger of breakdown. This technique has been applied in the motor of fig. 14, and also in that of fig. 21, where the stator is divided into two along a plane perpendicular to the shaft.

Profile of the electrodes in the cylindrical motor

As the cylindrical motor is the simplest to manufacture we made a more detailed study of this type,

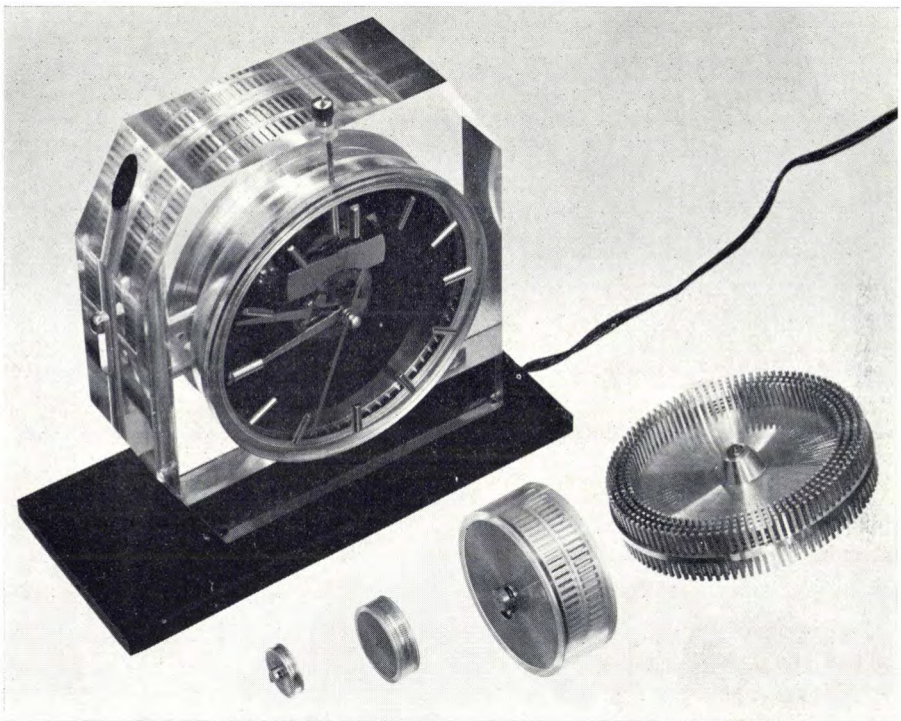


Fig. 14. A clock incorporating a cylindrical synchronous motor. The rotor (on the right in the photograph) consists of a disc with four rows of 100 electrodes on each side. The stator is formed by the bearing covers, each with five rows of 100 electrodes. The gap width in the axial direction is 0.1 mm. At 200 V, 1 rev/s, the maximum output is $P_{max} = 600$ μW. No slip-ring is used (see fig. 15). On the left of the rotor three smaller scale versions of the same motor design are shown.

A point that calls for attention in the construction of the motor is the slip-ring on the rotor shaft, which is necessary to connect the rotor to the voltage source. This is a possible source of trouble and increases the friction losses. It can be eliminated by dividing the stator into two mutually insulated halves and by connecting the terminals of the voltage source to these halves instead of to the whole stator and the rotor. This is illustrated in *fig. 15*. This method reduces the capacitance variation by a factor of four, because we now have two capacitors in series instead of the original two

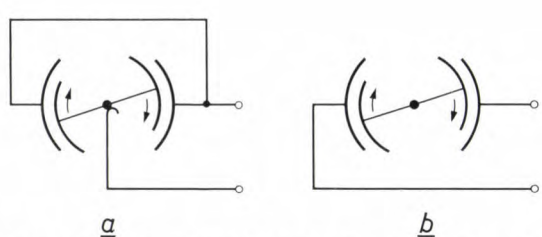


Fig. 15. A synchronous motor *a*) with slip-ring on the rotor shaft and *b*) without slip-ring. For the same dimensions the capacitance variation in (*b*) is a quarter of that in (*a*). However, for the same breakdown voltage per capacitor the motor (*b*) can be run at twice the supply voltage of (*a*), and then the power is the same in each case.

paying particular attention to the optimum electrode profile. Imagine a motor which is long in the axial direction and disregard the boundary effects that occur at the ends. This reduces the problem to two dimensions. Next we represent the cylinders by a planar structure consisting of two parallel straight rows of electrodes (*fig. 16*). What we are interested in is the capacitance c_t *per unit length in the axial direction for each pair of electrodes* and the variation of c_t that occurs when one row of electrodes is displaced with respect to the other. In particular we require the maximum variation Δc_t.

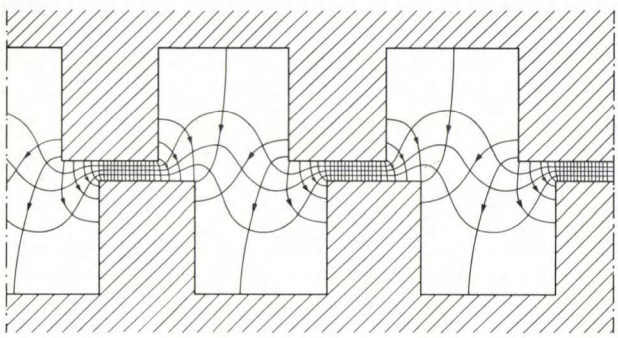


Fig. 16. Planar representation of the stator and rotor cylinder. The electric field between the electrodes varies periodically in the horizontal direction with the same periodicity as the electrode pattern. Lines of force and equipotential surfaces are shown schematically.

For an arbitrary position and with a given voltage difference between the electrodes the charge distribution on the electrodes, and hence the capacitance, is determined by the electric field distribution between the electrodes (see fig. 16). The boundary conditions for this field are a) that the electrodes form equipotential surfaces, each row at a given potential, and b) that the field repeats with the periodicity of the electrodes.

The variation of c_t was investigated using the resistance network shown in *fig. 17*. It consists of a regular square network of identical resistances fitted on two cylinders, the upper one of which can be rotated. By means of pneumatically operated contacts the two cylinders can be combined to form one network. To each cylinder a shorting wire is applied in the form of the electrodes to be studied, and scaled in size so that the circumference of the cylinder corresponds to the pitch of the electrodes. The resistance measured between these "electrodes" is then a direct measure of the required capacitance, and the capacitance variation can be determined by means of resistance measurements with the cylinders in different positions. The principle of this technique is that the current flow in the network — provided it is sufficiently fine — corresponds to the field lines between the electrodes, while the periodic boundary condition (b above) is automatically fulfilled due to the fact that the network, of a length corresponding to one electrode pitch, is closed in itself. The relation between the measured resistance R_t and the capacitance c_t is:

$$c_t = \varrho\varepsilon/R_t, \qquad \ldots\ldots \quad (13)$$

where ϱ is the resistance of a resistance element of the network (in our case 1200 $\Omega \pm 1\%$) and ε is the dielectric constant of the medium between the electrodes (ordinarily $\varepsilon \approx \varepsilon_0 = 8.86$ pF/m).

In this way we determined the capacitance variation for a large number of rectangular profiles. From a Fourier analysis of these results, and also from direct capacitance measurements on some motors we made, we found that a wide range of rectangular profiles lead to a capacitance variation which is a surprisingly good approximation to a pure sine wave. As long as the pitch t is not greater than 30 times the gap width s, about 90% of the maximum capacitance variation is due to the first harmonic C_1. The third harmonic C_3 is about 5% of C_1 (C_2 is even smaller, as was to be expected from considerations of symmetry).

Fig. 18 gives a more detailed example of the results obtained with the network. For a profile with rectangu-

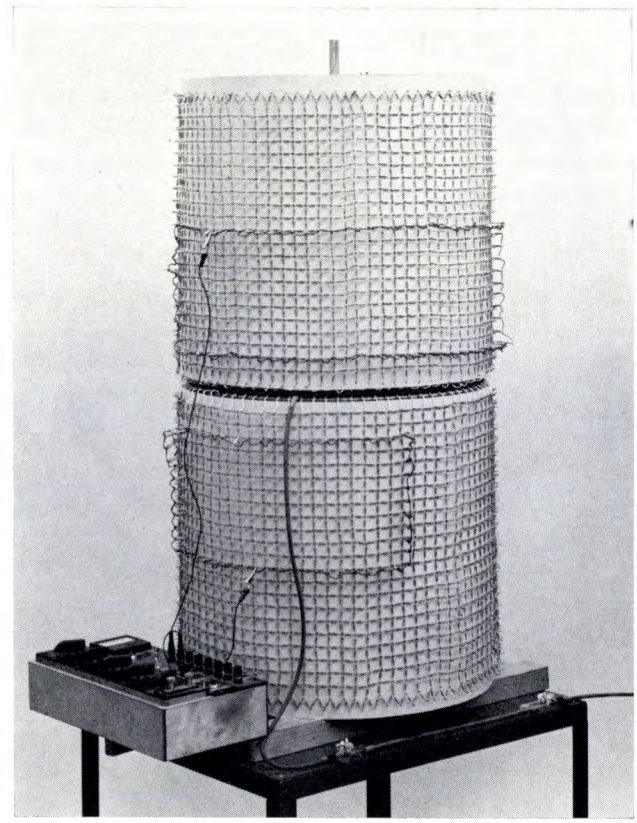


Fig. 17. The resistance network used for determining the capacitance variation for any given electrode profile.

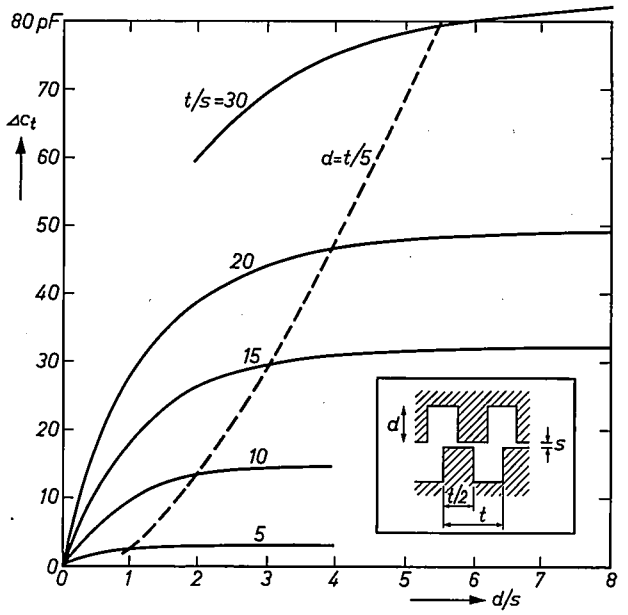


Fig. 18. The maximum capacitance variation of toothed electrodes (Δc_t) as a function of d/s for different values of t/s. The grooves have the same width as the teeth. The dashed curve corresponds to $d = t/5$.

lar teeth and grooves the maximum variation Δc_t is plotted as a function of d/s for various values of t/s (d being the groove depth, t the pitch length and s the gap width). This shows, for example, that where $t/s \leqq 30$ there is little point in making the grooves deeper than $5s$. For lower values of t the curves remain virtually level down to lower values of d, and it is seen that there is no point in making d greater than $t/5$. It can also be seen that for a given d and s and for a decrease in t, Δc_t decreases more rapidly than t. This is a quantitative demonstration of what was shown in fig. 9: the *total* capacitance variation of the motor ($\propto N\Delta c_t \propto \Delta c_t/t$) decreases when a given rotor surface is divided into an increasing number of electrodes, d and s being kept the same.

Another of the results we have derived with this method is that thin strips give a greater capacitance variation than teeth. A further conclusion is that it is advantageous to give the strips a width which is slightly smaller than the spaces between them. Finally we mention that the capacitance variation decreases if the strips are applied in or on a synthetic material. This is mainly caused by the increase of the minimum capacitance. The latter result was derived using a specially built resistance network in which different values of resistance were used for the "air region" and for the "dielectric region".

The losses of the synchronous electrostatic motor

Two loss mechanisms in the synchronous electrostatic motor have already been mentioned; the *ohmic losses* in the electrodes and in the electrode connections on charging and discharging the capacitor, and the *dielectric losses* in the insulating material of which the motor is made. Low ohmic loss is an inherent feature of the design of the motor, with its short electrodes connected in parallel. By choosing a suitable insulating material the dielectric losses can also be kept low.

At the low powers normally involved it is important to take account of the *friction losses in the bearings*. For the experiments we mainly used sapphires as bearings, for convenience in dismantling the test models. Another advantage of sapphires is that they act as insulators between rotor and stator. The coefficient of friction of a polished steel shaft running in sapphire bearings is about 0.1. Ball bearings, which are now available in very small sizes (down to 1 mm outside diameter), have ten times less friction, but they are difficult to mount and protect against dirt. Moreover they are about twenty times more expensive than sapphires.

Finally, the *air-friction losses* can constitute a significant proportion of the total power loss, as there are large surfaces close together in relative movement. We found from measurements that the air-friction losses with our profiled cylinders were no greater than those with smooth cylinders of the same radius r, length l and gap width s. The latter are proportional to $r^3 l \omega_m^2/s$. It follows that the losses can be minimized by using a long thin motor instead of a short thick one. The air-friction losses are only of importance in the faster-running motors. In small motors for timers and counters, with speeds of 1 to 2 rev/s, they can usually be ignored. Of course the air-friction losses can be avoided entirely by running the motor in a vacuum.

Starting the synchronous electrostatic motor

Many of the devices used for starting small synchronous electromagnetic motors are based on the presence of permanent magnets in the rotor. These are ruled out for the electrostatic case because there is no useful electrical analogue of the permanent magnet. Another method used for electromagnetic motors, and which can also be used for our purpose, is to use a hand-operated mechanism to give the rotor a speed close enough to the synchronous speed for the motor to start running in phase. Successful starting depends not only on the speed but also on the position of the rotor and the phase of the supply voltage at the moment the rotor and starting mechanism are decoupled. It may therefore be necessary to repeat the starting manipulation several times.

To calculate the spread in position and speed for which the motor will start, we must solve the equation of motion for the rotor and determine the conditions

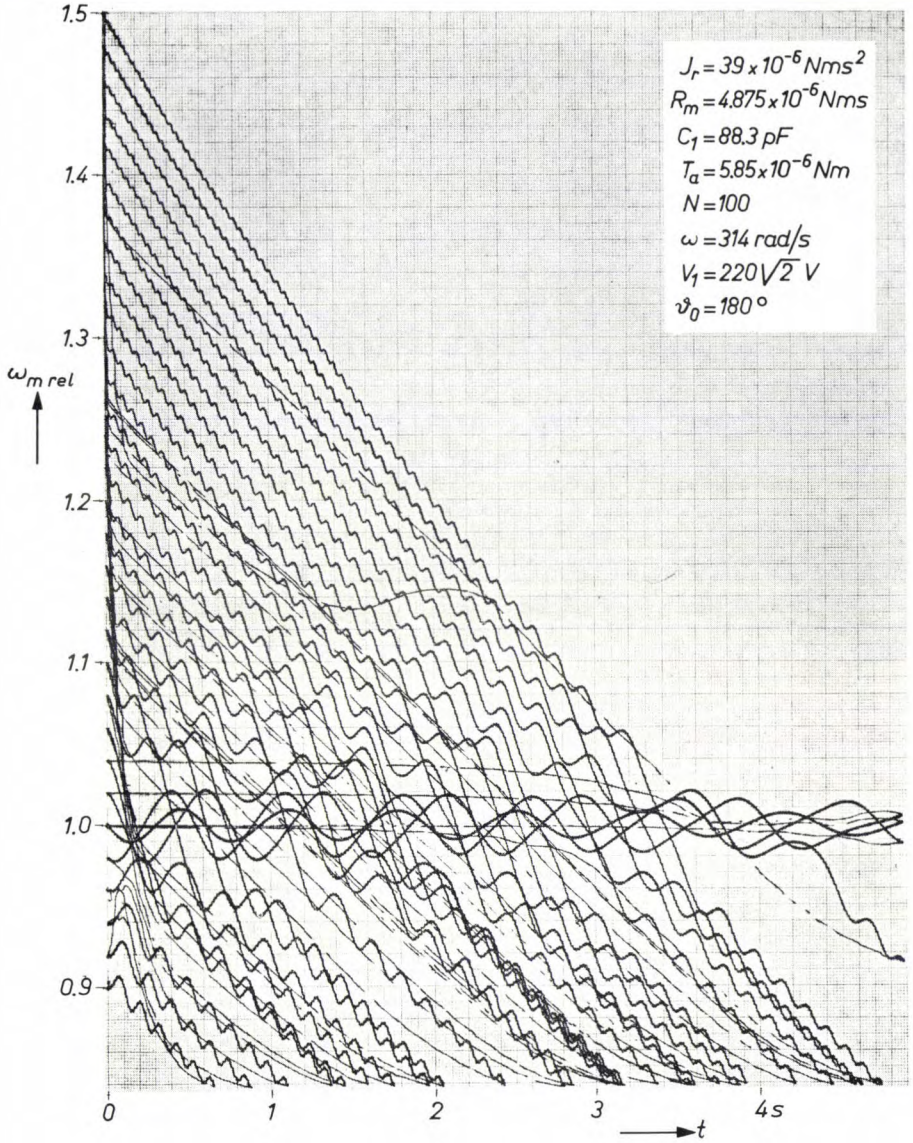


Fig. 19. Analysis of successful starting for a synchronous motor for the parameter values indicated. The results were obtained from the equation of motion of the rotor (14) by analogue computer. The graph shows the normalized speed of rotation $\omega_{m\,rel}$ as a function of time for a given initial value ϑ_0 of the rotor position, and for a large number of different initial values ω_0 of the angular velocity ω_m. Only three of the solutions given correspond to a successful start. $\omega_{m\,rel}$ is the angular velocity of the rotor divided by the normal angular velocity at the given supply voltage, i.e. $\omega_{m\,rel} = \omega_m(t)/(h\omega_s/N)$.

for the average rotor speed to equal the synchronous speed a few seconds after starting.

The equation of motion for the rotor is:

$$J_r\ddot{\vartheta} + R_m\dot{\vartheta} + \frac{\dot{\vartheta}}{|\dot{\vartheta}|}T_a - \tfrac{1}{2}V^2\,\mathrm{d}C/\mathrm{d}\vartheta = 0\,. \quad (14)$$

Here J_r is the moment of inertia of the rotor, R_m the coefficient of friction of air and T_a the counteracting torque of the driven mechanism. The last term is the motor torque (eq. 2). The electrical losses and the bearing friction are disregarded. Eq. (14) was solved by analogue computer for

$$V = V_1 \sin \omega_s t,$$

$$C = C_0 + C_1 \cos N\vartheta,$$

for a number of values of the parameters N, C_1, V_1, J_r, R_m, T_a and ω_s, and for a large number of initial values ϑ_0 and ω_0 of position and angular velocity. By choosing $V = V_1 \sin \omega_s t$ for the voltage we limited ourselves to those cases in which the phase of the supply voltage is zero at the moment of decoupling the starter mechanism ($t = 0$) and in which there is no reaction from the running motor on the voltage. *Fig. 19* shows the results obtained for a particular case and serves as an example of the variation of the angular velocity ω_m as a function of time for various initial values of angular velocity at a given initial position. In this example only three of the chosen initial speeds result in a successful start.

The values of ϑ_0 and ω_0 that lead to a good start are presented in *fig. 20* for two cases. The size of the cluster of small circles is a direct measure of the chance of a good start.

From figures like this we were able to conclude that the chance of a good start increases if the voltage V_1 or the capacitance variation C_1 is increased, and if the moment of inertia J_r is decreased. In starting experiments with various small motors, and in particular with

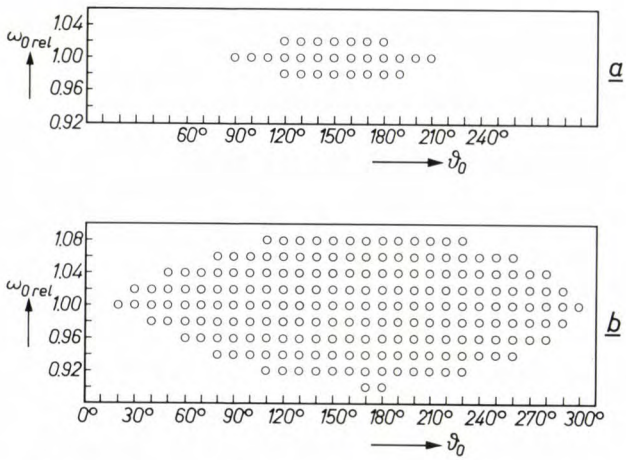


Fig. 20. The diagram shows those combinations of initial phase ϑ_0 and initial angular velocity ω_0 which the computer predicts will result in a satisfactory start. In (*a*) the parameters of the motor are as in fig. 19. In (*b*) the parameters are as for (*a*) but with $V_1 = 440/\sqrt{2}$ V (or $C_1 = 353.2$ pF). The ordinate shows $\omega_{0\,rel} = \omega_0/(h\omega_s/N)$, the initial angular velocity divided by the "normal" angular velocity $h\omega_s/N$.

starting synchronous motor. We have built such a combination which is shown in *fig. 21*.

The asynchronous electrostatic motor

The effect underlying the asynchronous electrostatic motor has already been mentioned in the introduction. A cylinder of a material which is slightly electrically conducting, or shows dielectric hysteresis, experiences a torque when placed in a rotating electric field. As previously stated this effect has been extensively studied in the past, but mainly for investigating the properties of materials and not for constructing a motor.

An electrostatic motor based on this principle requires a rotating electric field derived from polyphase voltages on a ring of electrodes (see figs. 4 and 5).

In our first experiments with such motors, with rotors of glass, quartz or methacrylate, it was repeatedly found that some time after the motor was connected to the supply it gradually slowed down and finally stopped. Blowing into it — on the assumption that dust had got in — caused it to start running again for a while. We concluded that it was not the dielectric hysteresis or the electrical conductivity of the cylinder material that caused the motor to turn, but a surface conductivity caused by moisture. On the basis of this idea we developed a motor in which the rotor consists

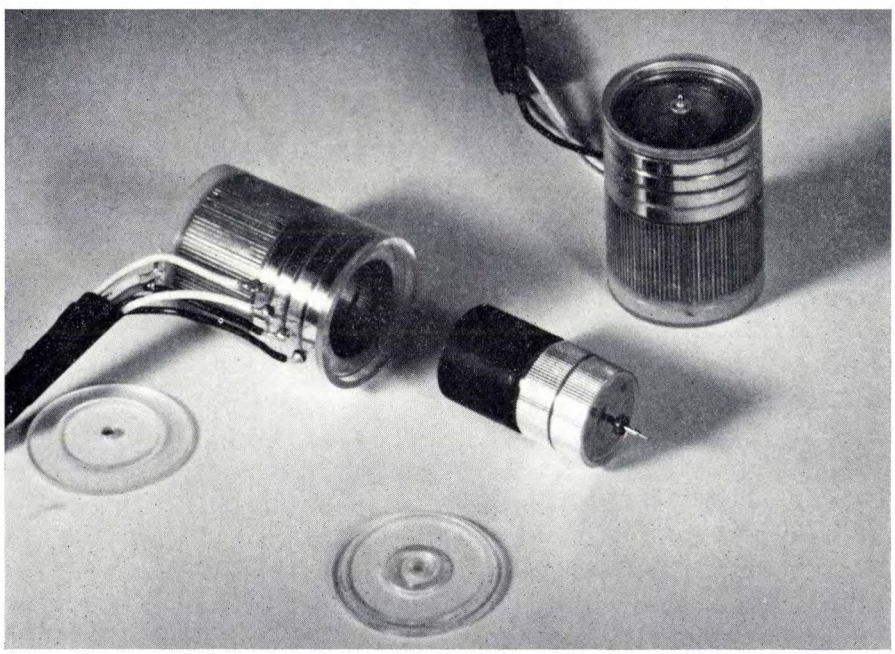


Fig. 21. A self-starting synchronous electrostatic motor, suitable for driving a clock mechanism. At 220 V, 50 Hz, the motor turns at a speed of 1 rev/s and gives a power output of about 60 μW. The synchronous motor consists of two parts, fed with voltages 90° out of phase. Both parts have the stator in two sections to avoid the use of a slip-ring (see fig. 15). Starting is based on the operation of the asynchronous part (rotor with black resistance layer, stator supplied with three-phase voltage). The diameter of the rotor is 15 mm.

the motor in fig. 14 the initial angular velocities were measured stroboscopically. The relative number of cases of a successful start at the various velocities were found to be in good agreement with the starting probability predicted by the computer.

Finally we note the possibility of combining a synchronous and asynchronous motor to obtain a *self-*

of a solid body of insulating material, covered with a thin layer of a slightly conductive material, and the stator is a ring of electrodes very close to the surface.

In retrospect it is clear that a conducting surface layer is to be preferred to a homogeneously conducting cylinder, for the charges caused by induction are mainly located at the surface anyway. Moreover it is much easier to obtain the required resistance reproducibly with a thin film (by vacuum evaporation) than with solid material (which would require a resistivity of about 10^8 Ωm).

Like the synchronous motor, the asynchronous motor can be realized in two forms, cylindrical and flat. A cylindrical motor can be seen in *fig. 22*, and a flat-disc type in *fig. 23*. *Fig. 24* shows a motor of a somewhat unusual type, a *linear* asynchronous motor. This may be regarded as a segment of a rotating-field motor bent straight.

Using a simplified model of this type of motor, C. Kooy of the Eindhoven Technical University and J. Ubbink of Philips Research Laboratories have calculated the torque to be expected on the stationary rotor and its dependence on the angular velocity of the rotating field and on the surface resistance of the rotor [8] [9].

These calculations show that the torque on the stationary rotor of a given motor has a maximum value for a particular angular velocity of the rotating field. This maximum torque is independent of the surface resistance R of the rotor surface, and occurs at a speed which is inversely proportional to R.

The reason why the torque on the stationary rotor is a maximum for a particular angular velocity of the rotating field — and hence for a particular supply voltage frequency — is as follows. Although low velocities give rise to high induction charges in the rotor surface, these charges can follow the rotating field so easily that there is no angular displacement between the field and the induction charges, and therefore the torque is zero. As the velocity increases the induction charges begin to lag behind, resulting in an angular displacement from the rotating field and hence in a torque. As the velocity increases still further, the angular displacement increases but the induction charges become steadily smaller so that the torque now decreases. It can be seen that the velocity at which the torque is maximum is greater for smaller values of R.

The greatest simplification in the calculations mentioned is that the voltage on the stator is represented by a sinusoidal wave travelling around the stator. The

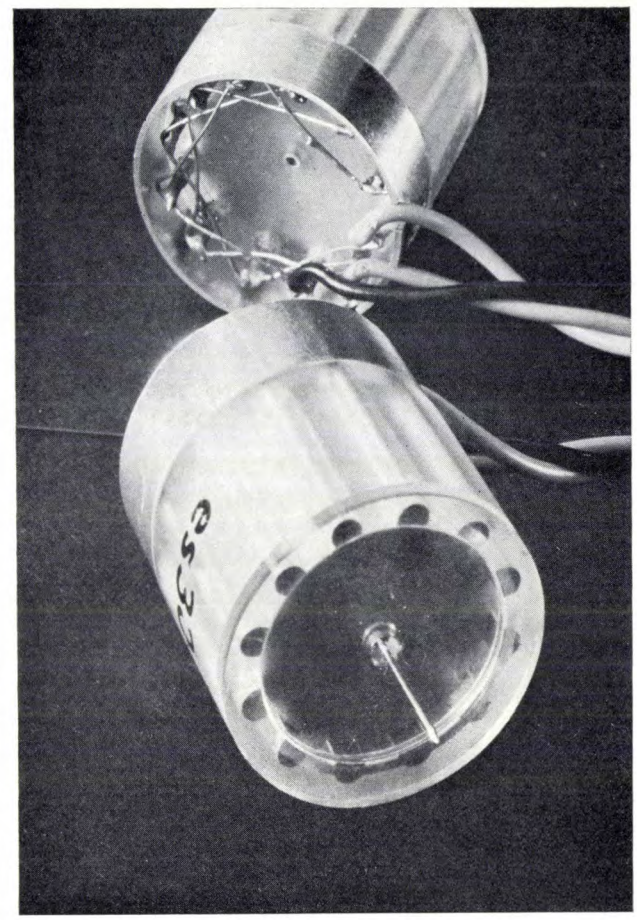


Fig. 22. A cylindrical asynchronous electrostatic motor, operating from a three-phase supply at 220 V, 50 Hz (the electrode connections can be seen in the mirror). There are 12 electrodes, i.e. four pole-pairs. The diameter of the rotor is 15 mm.

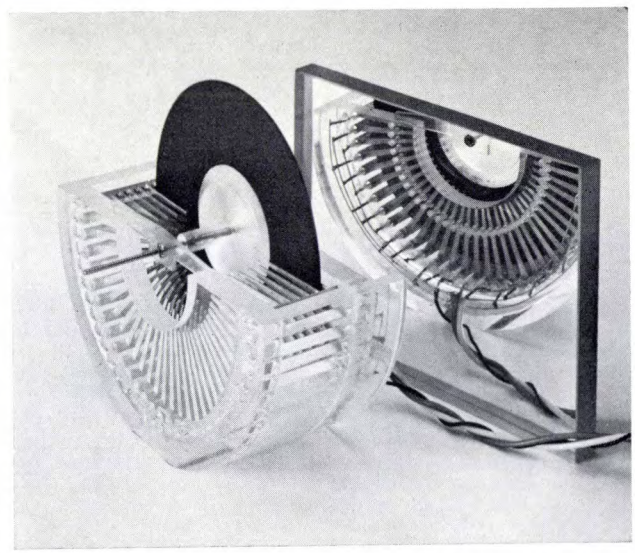


Fig. 23. An asynchronous disc-type motor operating from a three-phase supply, made for demonstration purposes. The rotor consists of five glass discs of 60 mm diameter, covered with a layer of resistive material; only one is shown mounted here. The rotor turns inside a ring of 30 comb-shaped electrodes. For operation it is not necessary for all five discs to be mounted, nor for the stator electrodes to cover the whole circumference. With a ten-disc motor of the same type (where 60 stator electrodes do in fact cover the whole circumference) a maximum torque of 200 μNm was obtained at 220 V, 50 Hz.

[8] C. Kooy, Torque on a resistive rotor in a quasi electrostatic rotating field, Appl. sci. Res. **20**, 161-172, 1969 (No. 2/3).

[9] J. Ubbink, Appl. sci. Res., to be published.

variation along the stator of the rotating field that is produced by a simple ring of electrodes and a three-phase voltage (as in our motors) is, of course, far from ideally sinusoidal. The higher harmonics of the voltage can make a substantial contribution to the stationary torque, which is not allowed for in the calculations. Nor is it immediately clear what amplitude should be assigned to the sinusoidal voltage in the theoretical calculation so that it approximates most closely to the real one.

In order to obtain an experimental check on the magnitude of the torque that can be attained with motors of this type, we made some tentative measurements on a small motor specially built for this purpose using interchangeable rotors to obtain different values of surface resistance. As it is difficult to measure the very small torque of such a motor when it is running, we confined our initial experiments to measurement of the torque on the stationary rotor as a function of the frequency of the supply voltage. It must be noted that because of the higher harmonics present in the rotating field, the maximum torque of a running motor may differ appreciably from the maximum torque on a stationary rotor.

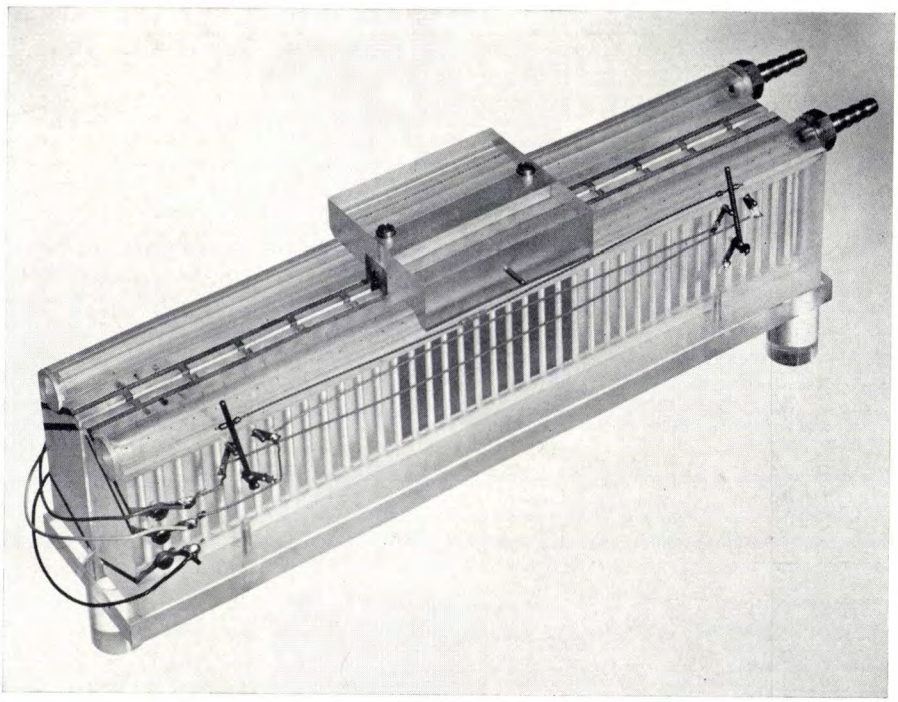
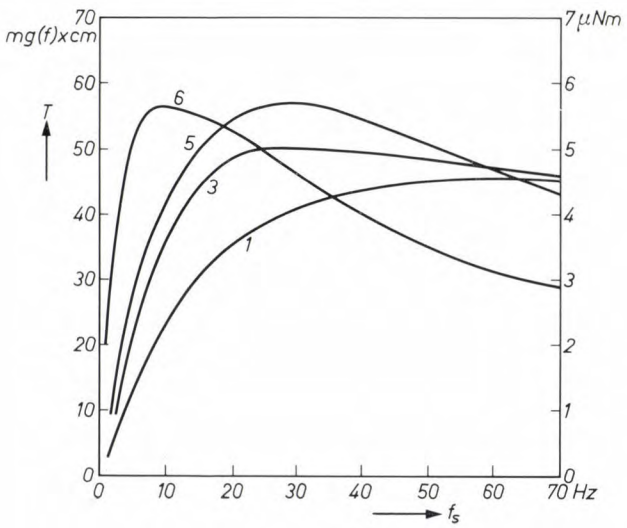


Fig. 24. A linear asynchronous motor. The "rotor" consists of a glass plate coated on both sides with a resistive material. It is mounted on a carriage which is supported by an air cushion and moves between two rows of stator electrodes. At each end the direction of movement of the field, and hence the movement of the carriage, is reversed by means of a switch operated by the carriage.



The experimental results agreed in broad outline with the calculations mentioned. For example the results shown in *fig. 25* confirm the predictions that when the rotor surface resistance R is varied there is no strong variation of the maximum stationary torque, and that the frequency at which the maximum occurs decreases as R increases. A quantitative comparison between the

Fig. 25. The torque T on the stationary rotor as a function of the frequency f_s of the supply voltage. These measurements were made on a cylindrical asynchronous motor with the following parameters: radius of rotor 15 mm, length of rotor 45 mm, gap width 55 μm, number of electrodes 120, number of phases 3, number of pole-pairs 40. The basic material of the rotor is polyphenylene oxide with $\varepsilon_r = 2.8$. The different curves were obtained using rotors with different surface resistances, the values being given in Table II.

results of experiment and the calculation is given in *Tables II* and *III*.

On the evidence of our experimental results and the calculations mentioned, we believe that it is possible to make asynchronous electrostatic motors with a volume of a few cubic centimetres and giving torques of practical interest (e.g. of a few μNm). In order to construct a practical model however, further investigations would be desirable, in particular into means of obtaining the most favourable form of rotating field.

The success of the experiments described here is largely due to the skill of H. Harhuis, C. J. Th. Potters and J. Th. Driessen of the workshops of Philips Research Laboratories, and to the assistance received from A. W. M. Bogers and Th. M. B. Schoenmakers.

Table II. Values of f_s and Rf_s for which the stationary torque reaches its maximum, as shown by fig. 25, for four rotors (i.e. four values of R). The theoretical value is obtained using the results of the calculations quoted [9].

rotor	R in $10^{10}\ \Omega$	$(f_s)_{opt}$ in Hz	$(Rf_s)_{opt}$ in $10^{13}\ \Omega s^{-1}$
1	6.5	60	0.3_9
3	9.2	30	0.2_8
5	13.8	30	0.4_1
6	25.6	9	0.2_3
theoretical value			0.5_4

Table III. Values of T_{max}/V_{rms}^2, obtained from measurements on the motor described in the caption to fig. 25, using one particular rotor (No. 5) for different values of the supply voltage (V_{rms} = r.m.s. voltage between two successive electrodes). The theoretical value is obtained from the formulae of [9], taking $V_0^2 = \frac{2}{3} V_{rms}^2$.

V_{rms} in V	T_{max} in μNm (1 μNm ≈ 10 mg-force × cm)	T_{max}/V_{rms}^2 in 10^{-10} Nm/V²
100	2.2	2.2
150	5.2	2.3
170	6.8	2.3
theoretical value		4.0

Summary. The operation of the electrostatic motor is based on the action of electric fields on charges. The principle has been known for a long time but, unlike the electromagnetic motor, the electrostatic motor has not found any application. The reason is that the power of the electrostatic motor is limited by the electrical breakdown strength, and is very much smaller under normal and comparable conditions than the power of the electromagnetic motor. Perhaps electrostatic motors might be able to compete with electromagnetic motors if operated in a high vacuum and in cases where a very small motor is required. The characteristics of this type of motor have been studied at Philips Research Laboratories in Eindhoven and the problems of construction investigated.

The synchronous electrostatic motor is basically a variable capacitor whose capacitance varies periodically as it is rotated, and which is connected to a square-wave voltage. A variety of small motors have been built, in particular the flat-plate type with vane-shaped electrodes, and the cylindrical type with electrodes in the form of teeth or corrugations. In the latter case the effect of the electrode profile on the capacitance variation was studied. An asynchronous type of electrostatic motor can also be built. This is based on the action of a rotating electric field on a rotor made of a material of low electrical conductivity (a poor insulator). Small motors of this type have been built of both the flat-disc and the cylinder type, the rotor consisting of an insulating material covered with a resistive layer. Experimental measurements of the stationary torque on the rotor were found to be in reasonable agreement with the values calculated on the basis of a simple model.

High-precision spark machining

C. van Osenbruggen

In the last ten or twenty years spark erosion has proved of real value for machining of material in various situations where the material is too hard or the workpiece too complicated for ordinary machining. It has also been found that spark machining is more efficient for many processes where cutting tools were previously used. The research and development work described in this article has shown how to achieve very accurate and relatively fast machining down to very small dimensions, and consequently a much wider range of applications.

Spark machining: method and application

One of the new workshop techniques in which the material is machined in a non-conventional way is the technique known as spark erosion or machining (or EDM, for "electro-discharge machining"). The fundamental principle of the process is the removal of small particles of material from two electrodes as the result of a spark discharge [1]. The method has gained its wide acceptance largely through the work of B. R. and N. I. Lazarenko [2]. Its principle is illustrated in *fig. 1.*

The electrodes are mounted close to one another, separated by the *spark gap*. The work is immersed in an insulating liquid, the *dielectric fluid.* Discharges of short duration are generated in rapid succession between the electrodes. To ensure regularity of the sparking process the length of the spark gap — which is the main factor determining the breakdown voltage — has to be constant. In most spark-erosion machines this is achieved by using the breakdown voltage to control the displacement of the "tool" electrode in such a way that deviations from the desired value of this voltage are compensated.

In the remainder of the article we shall refer to the tool electrode simply as the "electrode" and we shall call the other electrode the "workpiece".

A characteristic feature of spark erosion is that all electrical conductors and all non-conductors with a conductive surface coating (in some cases this can be formed in the discharge process) can be machined irrespective of their mechanical properties.

The chief application of spark-erosion machines is for machining very hard materials that conventional cutting tools cannot handle. Since there is no mechanical contact between the work and the electrode in spark erosion, there is no need for the electrode material to be particularly strong or hard. The various applications of the method are based on the fact that a negative and slightly larger "copy" of the electrode is machined in the workpiece, so that holes of any desired profile can be drilled; it is, in fact, possible to make a number of profiles, gaps or holes simultaneously. (This can also be done in ultrasonic machining, which, though less

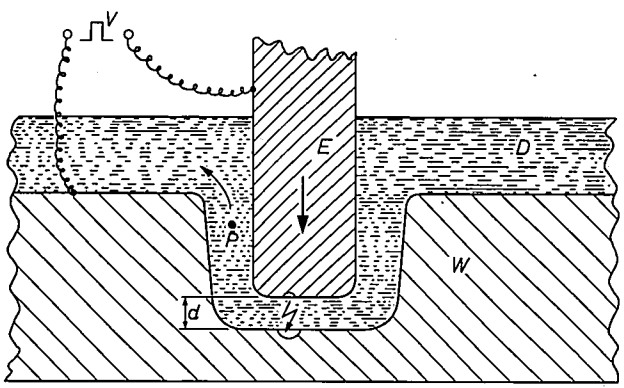


Fig. 1. The principle of spark erosion. An electrode E and a conductive workpiece W are held a short distance apart, the spark gap between them being filled with the dielectric fluid D. If the distance d between two points on either side of the gap is small enough, a voltage pulse V applied between the electrode and the work will result in a spark discharge, causing material to be removed from both E and W. A special mechanism ensures that the spark distance is held constant by advancing the electrode E. Particles P removed by sparking have to be flushed away.

C. van Osenbruggen is with Philips Research Laboratories, Eindhoven.

[1] G. Bredig, Z. angew. Chemie **11**, 951, 1898.
T. Svedberg, Arkiv för Kemi, Mineralogi och Geologi **2**, Paper No. 40, 1906.

[2] B. R. and N. I. Lazarenko, Physical foundation of the electric spark machining of metals, Vestnik Akademii Nauk SSSR **29**, No. 6, 1959.

accurate, is nevertheless of great significance for machining materials such as glass, ceramics and semiconductors.)

The main incentive for the research and development work on spark erosion at Philips Research Laboratories came from production departments, and this work has been chiefly concerned with equipment for machining very small workpieces much more accurately than has previously been possible: we refer to this as *high-precision spark machining*. Several machines with different accuracies and machining rates will be described in this article.

The mechanism of the removal of the material in normal spark erosion is chiefly a thermal process [3]. Two phases in the process can be distinguished (for simplicity we assume that the electrodes are of metal):

1) Heating of the metal to the melting point at the point of discharge.
2) Removal of the molten metal particles.

The heating phase can be explained as follows. In the conditions prevailing during normal spark erosion a breakdown occurs as a result of an avalanche of field-emission electrons leaving the cathode, which ionize the dielectric fluid and thus produce new electrons. The process is assisted by "bridges" of particles in the fluid moving towards points with the highest field strengths [4].

When the breakdown has started, a gaseous conducting channel forms and this channel is enlarged by the supply of energy from the spark generator. One advantage of using a dielectric fluid is that the width of the discharge channel is kept small, so that a high energy density can be obtained. (Another advantage of the fluid is that eroded particles can be flushed away from the spark gap more easily.) As a result of the voltage drop that appears at the two electrodes, the electrical energy transported by the channel is converted into heat, and this process can give high temperatures.

There is an optimum discharge time for a particular energy per discharge. If the discharge time exceeds this optimum, the power delivered is lower. The metal then takes longer to reach its melting point and in the meantime some of the energy is lost by thermal conduction into the metal. If, on the other hand, the discharge time is shortened and the discharge energy remains the same (so that the power is increased), material will vaporize and part of the energy will consequently be lost as latent heat.

The original difficulties with spark-erosion machining were the relatively low machining rate and the rapid wear of the tool electrode (the ratio between the amounts of material removed from the two electrodes depends chiefly on the energy dissipated in each of them). These difficulties were overcome by improving the power supply for the spark discharge and by using special electrode materials. It is very important here to find the best combination of electrode and workpiece material [5] [6].

Problems of high-precision spark machining

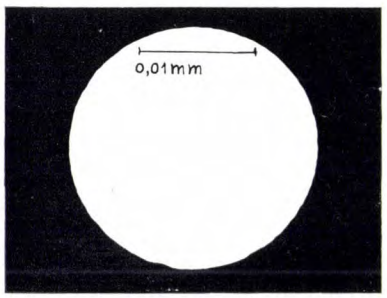
The nature of the processes which can be performed by high-precision spark machining is illustrated by a few examples in *fig. 2*. One of the technique's first applications was in making diaphragms for electron microscopes, which requires very accurate burr-free holes of 20 μm diameter to be made in 20 μm sheet platinum. Various mechanical methods had been tried previously for making these holes, but none of them had given satisfactory results. A hole produced by high-precision spark machining is illustrated in fig. 2*a*. The irregularities at the edge of the hole are smaller than 0.2 μm, the out-of-roundness less than 0.4 μm. Another example of the technique's unprecedented capabilities is the needle point shown in fig. 2*b*, in which a hole of 50 μm diameter has been sparked. The eye of the needle is also shown for comparison.

Machining to such exceptional accuracy demands a very careful choice of the conditions in which the spark-machining process is carried out. The electrode has to be very accurately shaped and an accurate control mechanism has to be employed to maintain the narrow spark gap (about 1 μm). The eroded particles have to be smaller than this since otherwise they cannot be removed from the spark gap without causing a short-circuit. Moreover, the smoothness required of the machined surface sets a limit on the dimensions of the craters formed in the sparking. Our investigations have shown that for these reasons low-energy discharges have to be used for high-precision spark machining, with a discharge time which is so short that only the *spark* stage of the discharge is utilized, the *arc* stage which follows it generally being suppressed. While energies ranging from 0.0003 to 2.0 J and discharge times of 2 to 2000 μs are customary for normal spark machining, the energy required for high-precision spark machining is about 10^{-8} J and the discharge time about 0.03 μs. Despite the large differences in discharge energy and discharge time, it has now been found possible to attain an *efficiency* (in terms of quantity of material removed per unit of energy) which is just as high as that of normal spark machining [7], while electrode wear can also be kept within very acceptable limits.

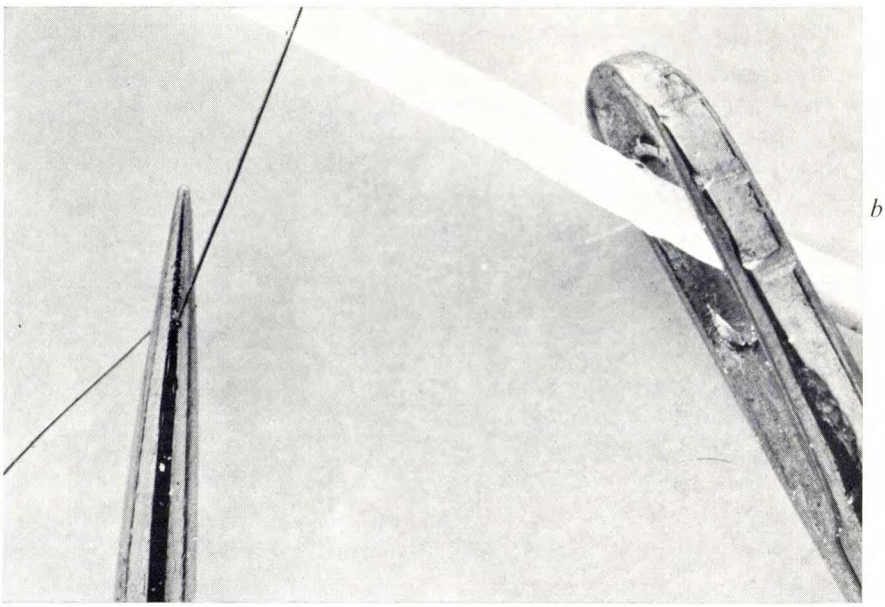
The *rate* of machining is a rather different matter. It might appear difficult to perform a given operation quickly enough at the same relatively high efficiency, since in high-precision spark machining only a very small quantity of material is removed at each discharge.

However, a relatively high machining rate can be attained by employing a high repetition rate. This means that certain special requirements also have to be met in generating the spark; we shall return to this later.

For further development of the high-precision spark-machining process it was important to obtain a clearer insight into the process under the "micro" conditions referred to above, so that it would be possible, for instance, to calculate the desired discharge energy and discharge time for a particular workpiece material and surface roughness. A series of experiments were therefore carried out with the first high-precision spark machines to become available. Some of the results of these experiments provided data for calculations which showed that the thermal model which applies for normal spark erosion is also approximately correct for high-precision spark machining (see Appendix).



a



b

Fig. 2. *a*) Electron-microscope image of a diaphragm hole in 20 μm sheet platinum made in a few seconds by spark erosion. *b*) Steel needle with a hole spark-eroded in it using an 0.04 mm diameter tungsten pin as the electrode. The hole diameter is 0.055 mm at the top and 0.050 mm at the bottom. The process took 1 minute, and the electrode wear was 15%. The photograph also shows the threaded eye of the needle for comparison.

Dielectric fluid and spark generator

The repetition rate has an upper limit, which is determined by the speed with which impurities such as eroded particles and the decomposition products of the dielectric fluid can be removed, e.g. by forced circulation of the liquid through a filter. During this process the spark gap must be flushed thoroughly. The reason for this is that the presence of conducting particles affects the breakdown field strength of the spark liquid and also causes discharges at the wrong places, so that the eroded gap becomes tapered. On the other hand, a certain amount of contamination by very small particles (< 0.01 μm) is necessary for rapid formation of the spark once the breakdown voltage has been reached [4].

A relatively heavy hydrocarbon is frequently employed as the dielectric fluid. The choice of fluid is rather limited, since the liquid must satisfy a number of requirements (it must have low viscosity, a low vapour pressure so that it does not ignite, it must not be poisonous and the liquid and its decomposition products must be chemically neutral).

At a very early stage in our tests, however, we found that when *de-ionized water* was used as the dielectric fluid, higher machining rates were obtained. This is because water dipoles directed into the electric field accelerate the formation of the spark (10 ns compared with 2000 ns under similar conditions in odourless kerosene) and the breakdown field strength is quickly restored after the discharge (within 50 ns), so that higher repetition rates are possible. Moreover, higher discharge densities are attainable with water. A further advantage is that spark discharges in water do not give solid decomposition products. If it is desired to use a closed water circuit, in addition to filters for removal of eroded particles, ion exchangers can also be

[3] N. Mironoff, Einführung in das Studium der Elektroerosion, publ. Microtecnic-Scriptar, Lausanne 1968.
[4] J. A. Kok, Electrical breakdown of insulating liquids, Philips Technical Library, Eindhoven 1961.
[5] P. E. Berghausen and H. D. Brettschneider, Electro-discharge machining program, 15 Nov. 1960 - 20 July 1963, Clearinghouse, Washington 1964.
[6] H. E. de Bruyn, Annals C.I.R.P. **16**, 183, 1968 (No. 2).
[7] H. Opitz, Funkenerosive Bearbeitung, Westdeutscher Verlag, Köln 1963.

incorporated to keep the conductivity sufficiently low (about 10^{-5} $\Omega^{-1}cm^{-1}$).

Depending on special requirements which the operation may have to satisfy (e.g. more emphasis on accuracy than on speed or *vice versa*), either type of liquid may be used in the machines that we have developed. From the hydrocarbons that could be used we have chosen odourless kerosene with a boiling range from 210 to 260 °C, which is sufficiently safe and has a reasonably low viscosity.

The spark generator can generally be formed from a simple *RC* circuit, whose capacitor discharges across the spark gap. With this circuit, however, there is a voltage across the spark gap for a large part of the spark cycle; when water is used there will be electrochemical side effects because of the residual conductivity that is always present (even after de-ionization). The resultant etching affects the dimensions of the work. When water is used as the dielectric fluid, a pulse generator must therefore be used which can be set to give short rectangular pulses so that no voltage appears between the electrode and the workpiece in the interval between two pulses.

Because of the limited resistance of the water in the spark gap, surface area A of the workpiece at sparking distance from the electrode cannot be arbitrarily large (a point which is usually of importance only for normal spark machining). Let us assume that the generator is a current source delivering a current of 40 A. In a spark cycle it must be possible for the voltage across the spark gap to increase from 0 V to the breakdown voltage, e.g. 300 V. To achieve this, the resistance in parallel with the discharge path must therefore be at least $R_{min} = 7.5\ \Omega$. If the resistivity ϱ is held at 10^5 Ωcm by ion exchangers, the surface area in question must not exceed

$$A_{max} = \varrho \frac{d}{R_{min}} = 13.3\ cm^2,$$

if the spark distance d is equal to 10^{-3} cm. The "capacitor" formed by the work and the electrode (see *fig. 3*) has a capacitance $C = 0.094\ \mu F$, so that with the maximum permissible workpiece surface area the discharge energy cannot be less than $E_{min} = \frac{1}{2}CV^2 = 4.2 \times 10^{-3}$J. Since the roughness of the machined surface depends on the energy, this high minimum value indicates that spark machining of workpieces with large electrode surfaces and with de-ionized water as the dielectric fluid is a relatively coarse process. For precision machining, a surface area A smaller than the maximum will therefore have to be chosen.

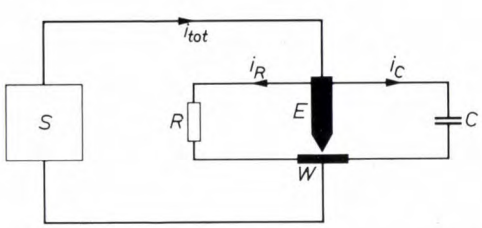


Fig. 3. Equivalent diagram of a spark-discharge circuit. i_R is the leakage current through the dielectric fluid, S a current source whose output current i_{tot} is equal to 40 A for periods of 2 μs at a time. i_C is the charging current of the capacitor C formed by the opposing surfaces of the electrode E and workpiece W.

An example of an operation with a high permissible spark frequency and hence a high machining rate is the spark-erosion process which we employ on silicon semiconductor elements with alternating layers of *N*-type and *P*-type material. A typical process is that of cutting grooves around transistor elements. A special feature of this case is the fact that the workpiece does not conduct in the normal way because the *P* and *N* layers pass current in opposite directions. By employing alternating polarity (see *fig. 4*), however, every current barrier can be crossed because half of the pulses are then in the correct direction. The breakdown occurs about 30 ns after the voltage has passed through zero and the time between two discharges is about 150 ns. The dielectric fluid in this case is de-ionized water. For comparison we may note that in hydrocarbons the probability of a discharge occurring within a 2000 ns pulse cycle is only 10%.

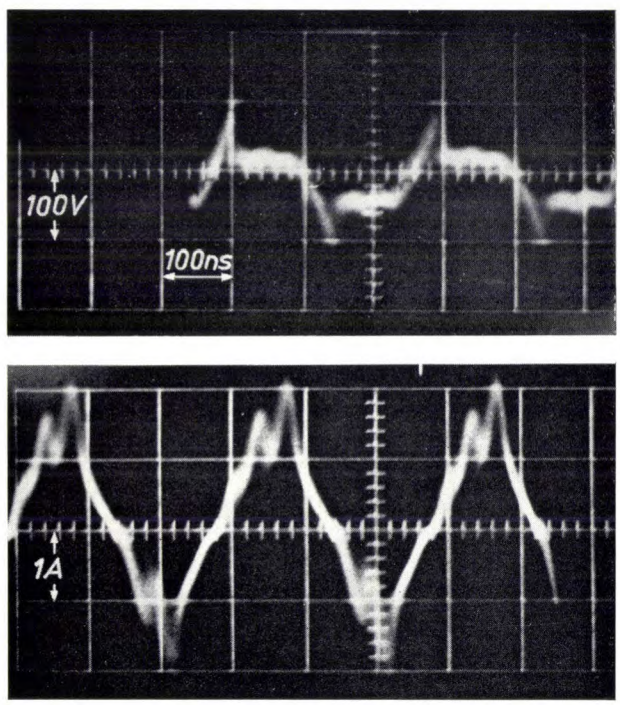


Fig. 4. Variation of voltage (upper curve) and current (lower curve) as a function of time during spark machining of silicon semiconductor elements. Capacitor discharges of alternating polarity are employed here. The dielectric fluid is de-ionized water.

The electrode

The properties required of the electrode material depend partly on the material of which the workpiece is made. We shall assume that when the discharge occurs the electrode surface is heated to no higher than its melting point T_0 and that as yet no material is removed. In this process the temperature at the surface of the workpiece T_w is required to be as high as possible. The ratio T_w/T_0 must therefore be large. With the aid of equation (1) in the Appendix it can be calculated that for a given workpiece material $pT_0/\sqrt{(\lambda\sigma c)_{el}}$ must then be large (λ is the coefficient of thermal conductivity, σ the density and c the specific heat of the electrode material; p is the ratio between the power dissipated in the workpiece and the power dissipated in the electrode). The minimum value which the material constant $T_0/\sqrt{(\lambda\sigma c)_{el}}$ should have depends on the workpiece material.

The power ratio is $p = V_w/V_{el}$, i.e. it is equal to the ratio of the effective voltage drop at the workpiece to that at the surface of the electrode after the discharge has become stable (after 10 ns). The effective voltage drop at the cathode (V_c in *fig. 5*) in a discharge is independent of the current density and decreases with

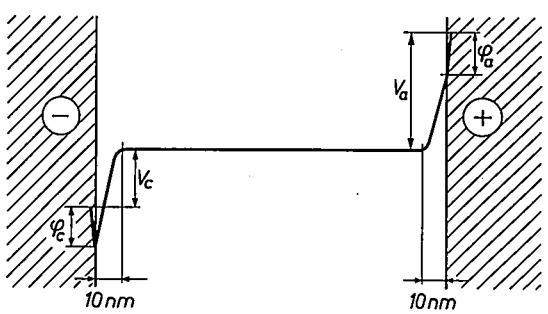


Fig. 5. Voltage drops at the electrode surfaces for an electrical discharge (schematic). φ_c is the work function of the cathode material, φ_a the work function of the anode material. V_c is the effective voltage drop at the cathode, V_a the effective voltage drop at the anode.

increasing work function (φ_c) of the electrode material. However, according to H. Schierholt [8] the effective voltage drop at the anode (V_a) is proportional to the current density and increases with the work function of the anode (φ_a). Since the diameter of the discharge path increases and thus the current density decreases during discharges of relatively long duration, p then has its greatest value if the anode is used as the electrode. In high-precision spark machining, however, discharges are short and current densities are high, so that V_a is large; the cathode is made the electrode and it is made from a material with the highest possible work function.

In semiconductors, unlike materials which are good conductors, the heat generated at the point of incidence of the discharge as a result of electrical resistance has a considerable effect on the temperature distribution in the material and the "entry resistance" makes an appreciable extra contribution to the voltage drop in the surface region. The voltage drop in a silicon surface connected as an anode is typically 400 times greater than in copper. This means that in machining silicon a) a high voltage is necessary; b) the temperature distribution is chiefly determined by the Joule heating; and c) the polarity has little effect on electrode wear, so that there is no objection to employing alternating polarity (as in the process described above; see fig. 4).

Some machines that have been developed

The results of earlier experiments, which had shown that low-energy discharges of short duration have to be used for high-precision spark machining, were made use of in developing a number of machines for this application at Philips Research Laboratories. Two main types of machine were developed: high-precision spark machines proper, in which the emphasis is on machining accuracy, and high-speed spark machines, which can give high machining rates. In addition to these we shall also discuss a number of machines which have been derived from the above types and adapted for special applications. All the machines are fitted with a very sensitive control system for displacement of the electrode. This system, which is controlled by the actual spark process, keeps the width of the spark gap accurately constant. By this means and by choosing the pulse shape and frequency that are most suitable for each case an excellent balance can be achieved between accuracy and machining rate.

Between them, the machines can tackle a very wide range of operations, from making 0.005 mm diameter holes in metal foil 0.05 mm thick to cutting holes with a cross-section of 50 mm² in tungsten carbide or hardened-steel punch plates. Some of these larger machining operations can in fact be performed by spark-erosion machines at present available commercially, but our machines compare very favourably because of their short machining times and efficient design.

A small grinding machine that we developed for making the sometimes extremely small tool electrodes has proved indispensible: with this grinder a tungsten rod of 1 mm diameter can be ground down to a 4 μm diameter electrode in about twenty minutes.

The high-precision spark machine

The high-precision spark machine, which is shown in the schematic diagram of *fig. 6* and the photograph of

[8] H. Schierholt, Thesis, Aachen 1964.

fig. 7, is built on a fixed column *1* and base. A second column (*2*) which moves in a straight high-precision guideway is fitted to the main column. Column *2* contains two flexible diaphragm springs *3* and *3'* joined by a tube *6*. The electrode holder *4* is fastened to the lower diaphragm spring. The diaphragm springs are used to give frictionless vertical movement of the electrode holder. Inside the tube there is a metal strip *5* with its lower end fastened to diaphragm spring *3'* and its top end rigidly fastened to column *2*; it is tensioned by a stiff spring *7* which pushes diaphragm spring *3'* downwards. If a current is passed through the strip, the Joule heating makes the strip expand and the electrode holder is moved downwards (but not, of course, far enough to cause a permanent extension of the strip). The current through the strip is controlled as a function of the breakdown voltage in such a way that this voltage and hence the spark gap are kept constant. A cooling jacket *9* protects the sensitive mechanical parts from the heat generated in the strip. The electrode is given a very small vibratory movement by means of an electromagnetic system that applies a longitudinal oscillation to strip *5*. This vibratory movement, which can be varied in frequency and amplitude, helps to give good flushing of the spark gap.

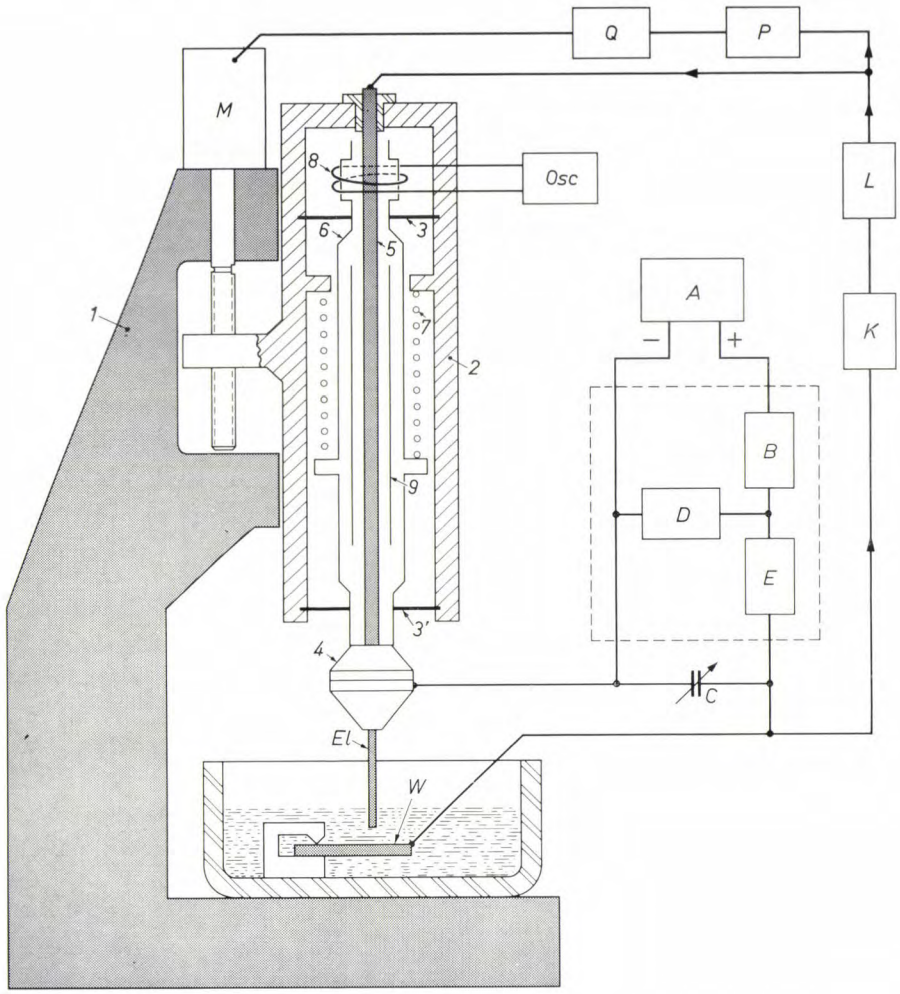


Fig. 6. Schematic cross-section of the high-precision spark machine developed at Philips Research Laboratories in Eindhoven, and a block diagram of the *RC* spark generator and the control system for positioning the electrode. *1* fixed column with base. *2* movable column, in which, between diaphragm springs *3* and *3'*, there is a tube *6* which is pressed down by a spring *7* so that the metal strip *5* is put in tension. *4* electrode holder with electrode *El*. *W* workpiece. *8* vibrator system. *9* cooling jacket. *M* stepping motor. *A* d.c. voltage source. The current stabilizer and limiter *B*, the voltage stabilizer *D* and the delay line *E* together form the charging circuit. *C* capacitor. The required signal is applied to the displacement system of the electrode holder by spark-voltage detector *K* via amplifier *L*. *P* voltage-dependent pulse generator which supplies the control unit *Q*.

An electrode displacement of 0.8 mm can be obtained with the diaphragm suspension. This displacement may be sufficient for very fine machining, but larger displace-

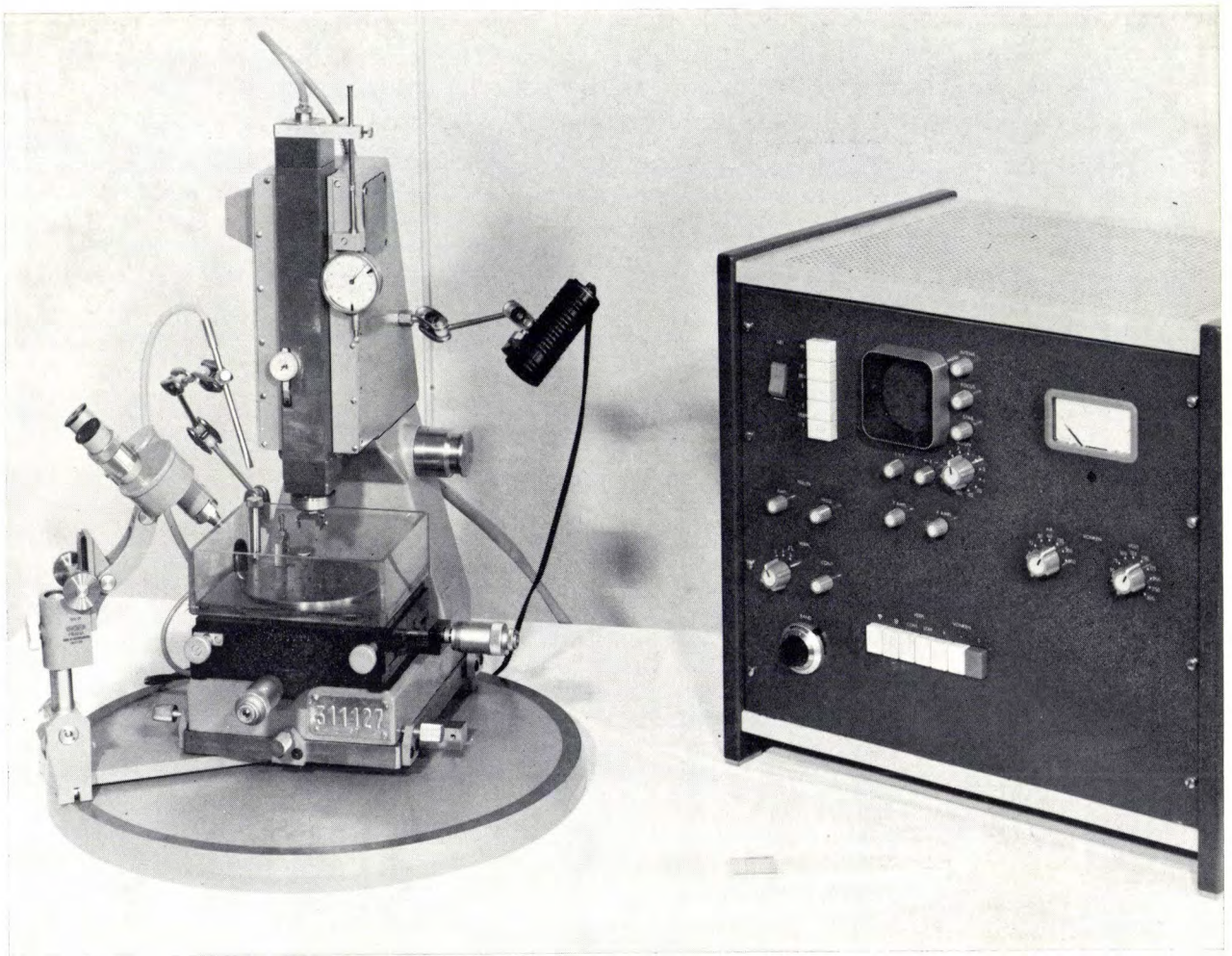


Fig. 7. The high-precision spark machine. The unit on the right contains some of the control equipment.

ments are often necessary. These are provided by a combination of the metal-strip control and a second control system in which the whole of column *2* is displaced vertically with respect to column *1* by means of a stepping motor *M*. The displacement range is thus increased to 10 cm, the length of the guideway.

The dielectric fluid used in this machine is odourless kerosene. It is pumped from a reservoir via a filter housing to the workpiece mount and from there back to the reservoir. There is a second drainage hole at the centre of the work-table; if a workpiece with a preformed hole is placed above this drain-hole, the contaminated liquid is more easily removed from the spark gap.

The spark generator is an *RC* circuit in which the capacitor is discharged across the spark gap. This spark generator has to satisfy the following requirements:

a) The open-circuit spark voltage (i.e. the maximum voltage the capacitor can be charged to) must be stabilized.

b) The capacitor charging current must be constant and adjustable.

c) The charging current must be limited if there is a short-circuit or arcing between the workpiece and the electrode.

The block diagram of the circuit used is shown in fig. 6. A d.c. source *A* charges the variable capacitor *C* via a current stabilizer and limiter *B* and a delay line *E*. *D* is a voltage stabilizer. The delay line prevents stray capacitance from causing discharges across the spark gap in addition to the discharge of *C*.

The spark voltage is adjustable from 30 to 300 V, the capacitor charging current from 6 to 600 mA. Since the stray inductance and capacitance of *C* have been kept small, the discharge time is only 0.03 μs. The minimum discharge energy which can be adjusted by selecting the voltage and capacitance is 10^{-8} J, which will give a surface roughness as low as 0.06 μm (the maximum height of irregularities). The maximum energy which can be obtained is 1.5×10^{-3} J, which gives a surface roughness of 3.4 μm (maximum height). *Fig. 8* shows the charge and discharge characteristics of the circuit. The frequency of the discharges, which is determined by the capacitance of *C* and the resistance in the system (and

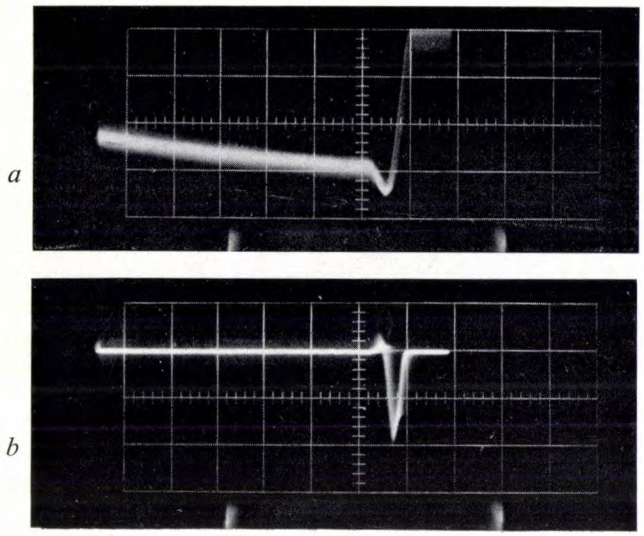


Fig. 8. Oscillogram of a very-low-energy *RC* spark cycle. *a*) Voltage between workpiece and electrode. *b*) Current pulse of spark discharge. $C = 70$ pF, spark energy $\frac{1}{2}CV^2 = 0.13$ μJ.

therefore by the conditions in the spark gap), can be varied from 10^3 to 10^5 Hz under favourable conditions.

The high-precision spark machine is also fitted with a very accurate cross-feed table for setting up and locating the work, and a microscope or a binocular magnifying glass for observing the position. *Fig. 9* shows some examples of work produced on this machine.

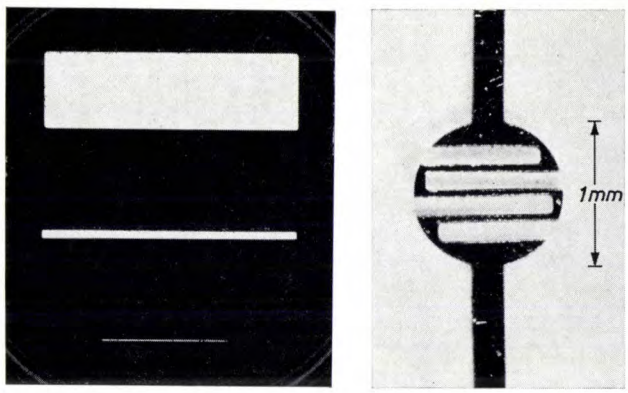


Fig. 9. Examples of machining by the high-precision spark machine.
a) Three slots with widths of 3, 0.3 and 0.03 mm in 0.1 mm thick nickel sheet, for which spark energies of 360, 10 and 0.15 μJ, respectively, were used. The machining time was about ½ hour for each slot, the wear 3-5%.
b) Resistance element spark-machined in 5 minutes with a tungsten-copper electrode in a 0.03 mm thick plate. Spark energy 17 μJ. Wear 10%.

The high-speed spark machine

The high-speed spark machine (*fig. 10*), which does not look very different from the high-precision spark machine just described, has been adapted to permit de-ionized water to be used as the dielectric fluid, so that high discharge frequencies and hence high machining rates can be achieved. On the other hand, the machine is less suited to applications where the very highest accuracy is required.

To enable the advantages of de-ionized water to be exploited to the full, the spark is not generated by an *RC* circuit but by a pulse generator giving well-defined current pulses.

The high-speed spark machine has the following ranges of adjustment:

discharge current	2- 40 A
discharge time	0.2- 2.0 μs
repetition rate	2-250 kHz
discharge energy	0.006-2 mJ (up to 10 mJ with extra discharge capacitor for preliminary machining).

The drive and guide system for the electrode holder in this system are housed in a "spark unit" which is mounted rigidly on a column (similar to *1* in fig. 6). In one particular version the spark unit is readily interchangeable, which is a valuable facility in production

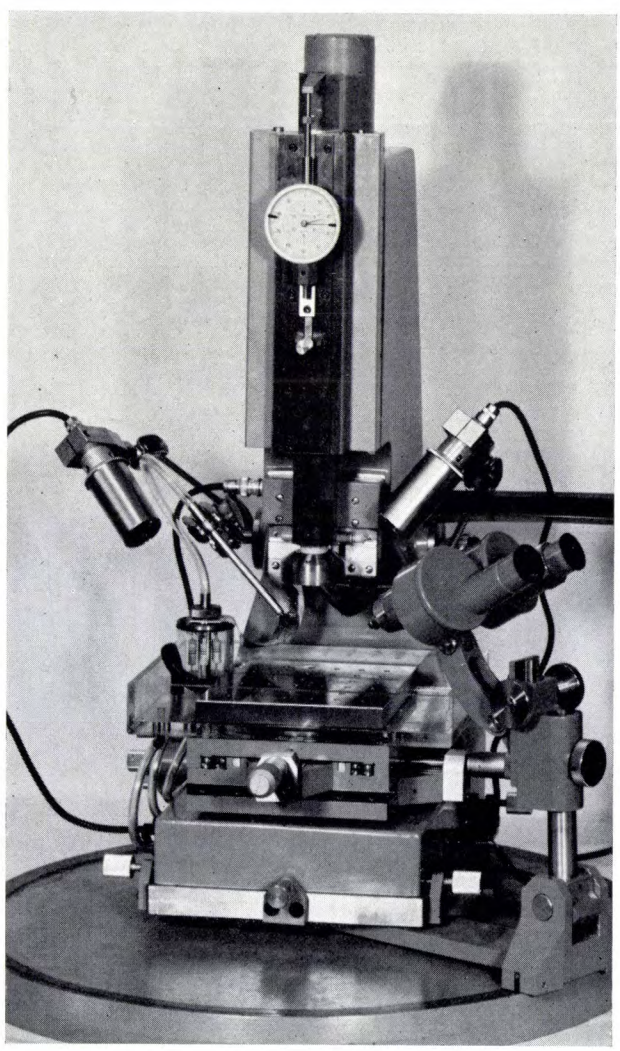


Fig. 10. High-speed spark machine.

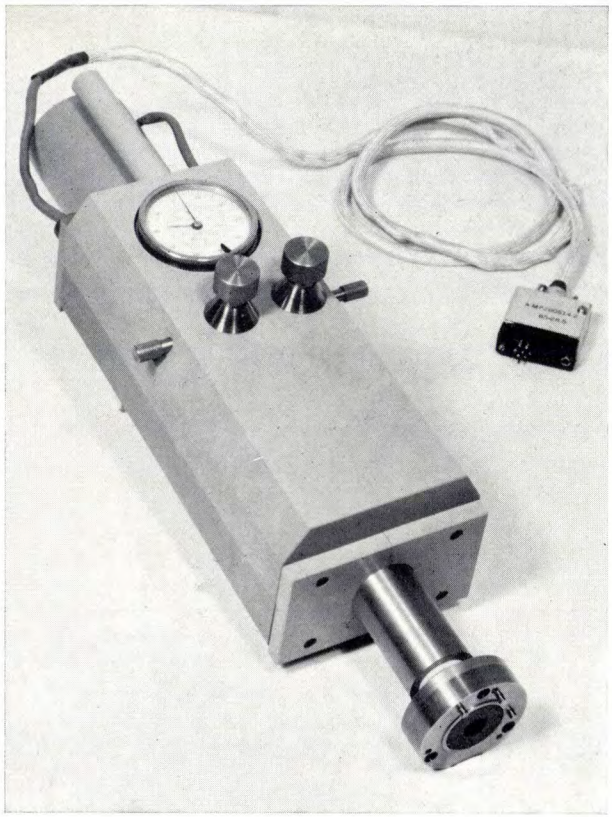


Fig. 11. Spark unit for the high-speed spark machine. This unit is easily interchangeable so that the machine can be quickly adapted to various types of operation.

work. *Fig. 11* shows a photograph of one of these spark units. The electrode in this unit is moved up and down on a threaded shaft driven by a stepping motor. After manual location of the electrode above the work, the stepping motor is automatically controlled by the breakdown voltage during the spark-machining process. The discontinuous displacement of the electrode is 1.25 μm for every step of the motor; it was found experimentally that to permit this the spark gap has to be at least 2 μm. The smallest hole that can then be made (diameter 30 μm) depends on the minimum spark energy (6 μJ). A thinner electrode would be overloaded at this energy, and the wear would be abnormally high.

The mechanical dielectric fluid is circulated in a closed system that includes filter and ion exchangers; a great deal of care has been taken to ensure that the forced flushing of the spark gap is efficient.

An example of a workpiece made with this machine has already been shown in fig. 2*b*. Several more can be seen in *fig. 12*.

A semi-automatic modification of the two machines described above is used for quantity production of diaphragms for electron-optical systems. With this machine it only takes six seconds to spark a 50 μm hole in a sheet of nickel-chromium steel 50 μm thick

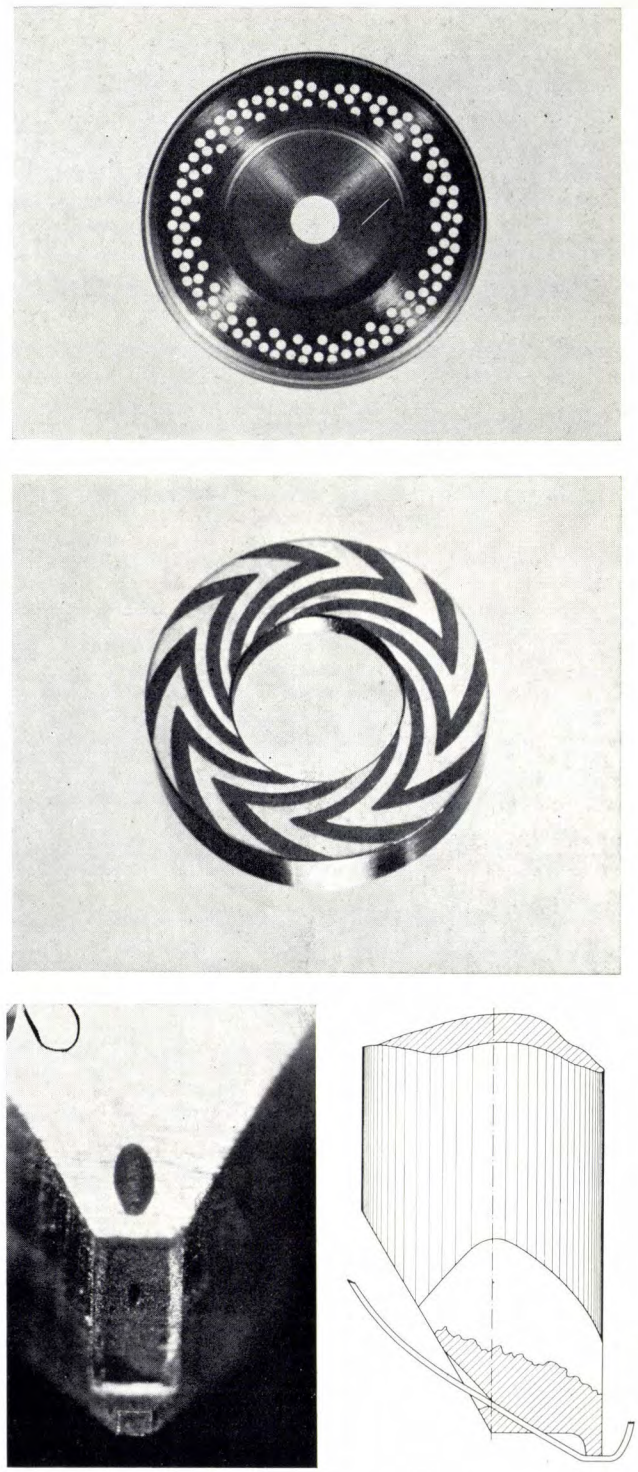


Fig. 12. Examples of machining with the high-speed spark machine. *a*) Shaver head with 120 holes made simultaneously in 0.08 mm thick chromium steel. Electrode: 120 tungsten-thorium pins (diameter 0.49 mm) in one holder. The holes have a diameter of 0.52 mm. Spark energy 2 mJ. Machining time ½ min. Electrode wear 8%. *b*) Spiral-groove bearing [9], external diameter 12 mm, profile depth 0.03 mm. The electrode was a copper plate with a negative profile. Spark energy 0.25 mJ. Machining time 15 s. Electrode wear 30%. *c*) Steel "welding bit" for ultrasonic welding of 50 μm diameter wire. The duct and tapered insert hole through which the wire is led were spark-machined in 3 minutes with a tungsten electrode (60 μm diameter). Spark energy 10 μJ. Electrode wear 15%.

[9] E. A. Muijderman, Philips tech. Rev. **25**, 253, 1963/64.

(the accuracy is 1 μm and the spread in the diameter is no more than 2 μm in a run of 500 holes). The operator merely has to position the diaphragm plate with one movement of the hand and press a starting button; the movement of the electrode is then completely automatic.

A contouring machine

The high-speed machine is also used as a contouring machine (*fig. 13*) in which the work is automatically displaced in the horizontal plane to correspond with a line on a drawing, the electrode again being movable in the vertical direction. This method is very useful when only one or two copies of a workpiece of very complicated shape have to be made or for an experimental optimization of the dimensions of a provisional design of workpiece: in both of these cases it would be too costly and time-consuming to make special electrodes of the required shape. The line is followed automatically by means of a digital scanning procedure that makes use of a system of four photodiodes placed side by side above the drawing (see *fig. 14*); the line on the drawing is magnified 10 times and projected on to the diode system. In the scanning process the light source, lens and photodiodes are moved in steps of $\pm \frac{1}{2}d$ (= 2.5 mm) at a time in the x- or y-direction by stepping motors. The accuracy with which the drawing is scanned is therefore 2.5/10 = 0.25 mm, which is also the minimum permissible line thickness. The system is displaced in such a way that the photodiodes are always kept above the line. The steps of the detection system are transmitted via two other stepping motors and with a reduction by a factor of 50 or 500 (adjustable) to the cross-feed table of the spark-erosion machine. The drawing is therefore made 50 or 500 times true size. In preparing the drawing an allowance also has to be made for the thickness of the electrode. The total inaccuracy introduced by the line-tracking system is then ± 5 μm at 50× and ± 0.5 μm at 500×. The repetition frequency of the scanning cycles can be set manually or controlled by the breakdown voltage during the spark-machining process.

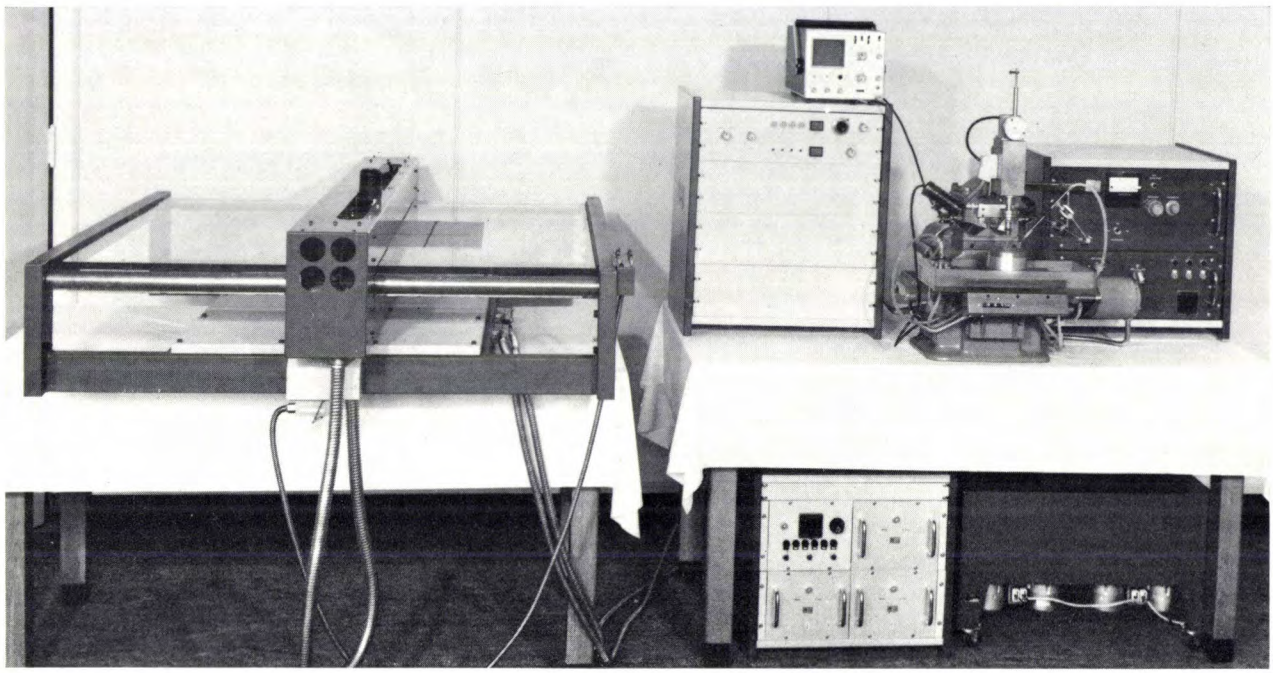


Fig. 13. Contouring machine with automatic line-tracking system. A greatly enlarged contour drawing of the workpiece (e.g. 500 times full size) is placed on the table on the left; the contours of the drawing are traced out automatically and transferred to the control of the x-y carriage on which the workpiece is mounted. The associated electronic equipment is located under the high-speed spark machine and to the left of it; the unit on the right is the power supply.

Fig. 15 shows a steel tuning-fork 20 mm long and 2 mm thick which was made by this method. The electrode was a tungsten rod 0.5 mm in diameter. A number of workpieces (10 for example) can be made simultaneously with the contouring machine just described, because the spark generator has several outputs. The control of the vertical electrode feed motion has been adapted to this special application.

Spark erosion with piezoelectric spark-gap control

A special method for spark-gap control has also been developed in our laboratory and this has enabled us to

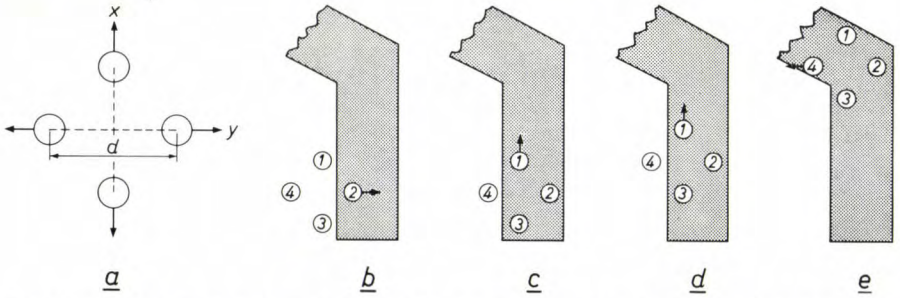


Fig. 14. Line-tracking system of the contouring machine. The line on the drawing being scanned, which is reproduced 10× magnified at the detection system, is shown dark.
a) The diode array. Distance *d* is 5 mm. The arrows indicate the direction of movement associated with the individual diodes in the detection system. During the scanning cycle the conduction of the diodes is measured in a particular sequence. If the diode being scanned is found to be receiving no light, an associated bistable circuit (flip-flop) changes over and the detection system is moved on one step in the corresponding direction.
b) An assumed initial position: diode *1* is scanned first. *1* receives light, so no displacement takes place. The next diode to be scanned (proceeding clockwise) is diode *2*; this is dark and the detection system is moved on one step to the right.
c) Next situation: a fresh cycle starts. A memory circuit ensures that this always starts with the *preceding* diode (in this case, *1*). This time diode *1* is dark, so the system makes one step upwards.
d) The next cycle begins with *4*: no step. Now clockwise to *1*: one step upwards, and so on.
e) In this position *4* is again scanned first. This time it is found to be receiving *no* light, so the system moves one step to the left, thus following the bend in the line.

The drawing is tracked clockwise in this way along its left-hand outer edge. It may happen that a particular place on the light-dark boundary is occupied by a diode which, because its sensitivity is different, gives an opposite signal to that of another diode which occupied the same spot previously. A memory store ensures that this illogical signal sequence (which would cause the system to change sign) is impossible.

make a "minidevice" which is particularly well suited to certain surface-processing operations. These include the profiling of workpiece surfaces (e.g. perforating a thin coating to a particular pattern) and cutting very fine grooves in thin conducting layers on an insulator. (If the method described above were used for this, the electrode could get bent by being pressed against the insulator after the layer to be removed has been sparked away.)

The spark-gap control device consists simply of a strip of piezoelectric material coated on both sides with metal foil. This element is clamped at one end (see *fig. 16*) and the free end carries a very light electrode holder which is in electrical contact with the lower foil.

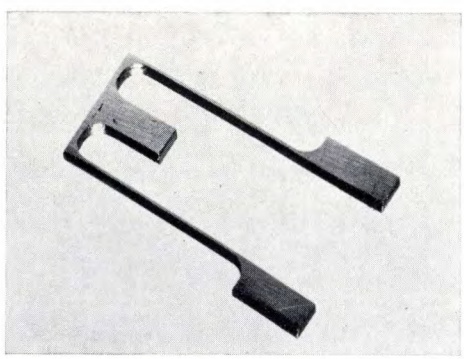


Fig. 15. Steel tuning-fork 20 mm long and 2 mm thick, made with the contouring machine. The electrode was a tungsten rod 0.5 mm across. 10 tuning forks can be made simultaneously in 4-5 hours.

The element, which is used as a vibrator, can be positioned above the work by means of a cross-feed table (*fig. 17a*).

The workpiece is connected to the other foil of the vibrator via an inductor L and a resistor R. The vibrator, whose capacitance is represented by C_1, is connected to a current source A.

For machining, the tip of the electrode (a thin metal wire) is located 0.1-0.2 mm above the surface of the workpiece. The current source delivers a constant charging current, causing the voltage across the vibrator (i.e. across capacitor C_1) to increase linearly. The operation of the element causes electrode *2* to move towards the workpiece until a discharge occurs. Careful selection of L and R ensures that the discharge is quenched when the voltage of the variable capacitor C_2 ($\ll C_1$) falls to a low value. The short relaxation time of the circuit comprising R, L, C_1 and C_2 enables C_2 to reach the breakdown voltage quickly again. This is followed by a new discharge, and so on. After a series of discharges, however, the voltage across the vibrator has dropped to such an extent (and the distance between

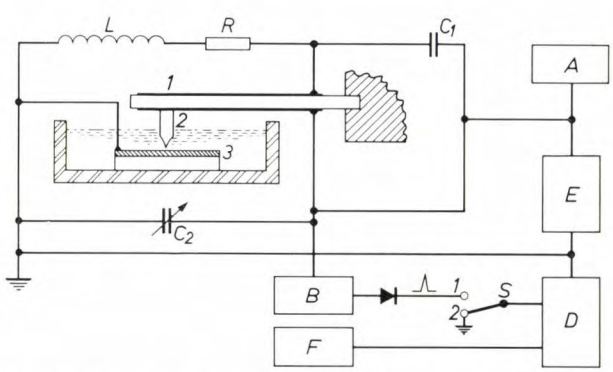


Fig. 16. Piezoelectric spark-gap control. *1* piezoelectric vibrator with electrode holder *2* above workpiece *3*. Capacitor C_1 represents the capacitance of the vibrator. L inductor, R resistor. C_1 is charged from the current source A. C_2 small variable capacitor. B is a differentiating circuit network which (if switch S is in the upper position) ensures that the voltage drop after a discharge changes over the bistable circuit D. This makes the blocking circuit E conduct, so that the discharge circuit is short-circuited and a second discharge cannot occur. A reset circuit F is connected to a second input of D; when a switch included in this circuit is operated, the next single discharge can take place.

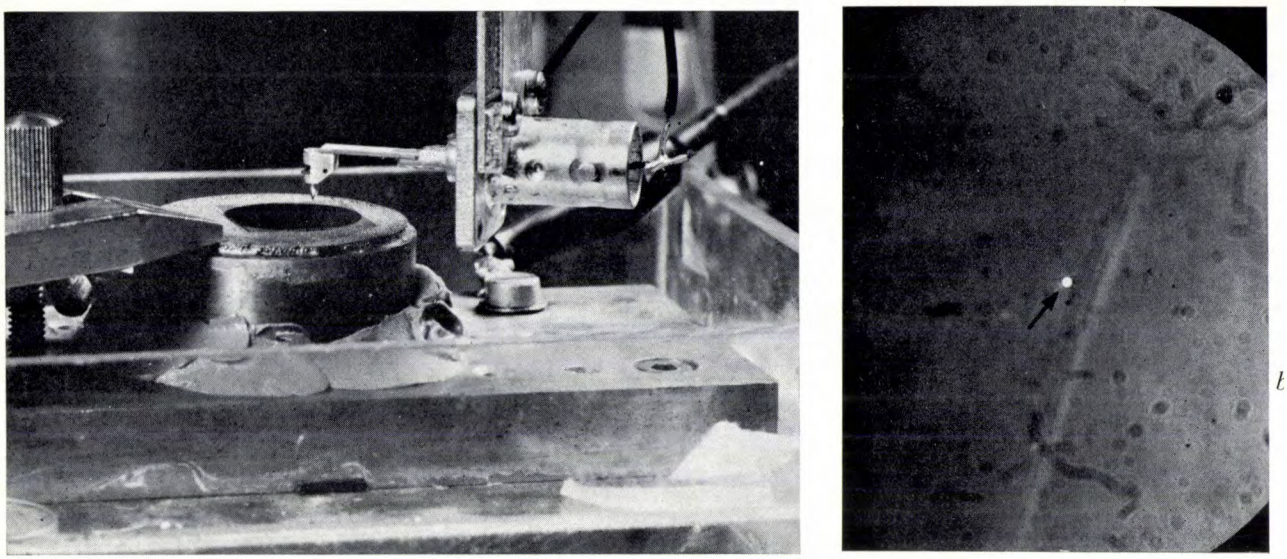


Fig. 17. *a*) Piezoelectric vibrator as an electrode holder. *b*) 3 μm hole made in a thin metal coating with a single discharge.

the electrode and the workpiece has consequently become so great again) that discharge is no longer possible. C_1 is then charged up again until the next series of discharges starts. (Where less accuracy is required, L, R and C_2 can be omitted; the discharge energy will then be determined solely by the capacitance C_1 of the vibrator.)

If the work is displaced in the horizontal plane during the sparking process just described, very fine and accurate lines can be machined in the surface. The system reacts swiftly to irregularities in the surface of the workpiece so that they do not affect the length of the spark gap.

Special provision has been made for the machining of very small holes in evaporated films or thin foil 0.3 to 5 μm thick with single discharges. The circuit diagram is shown in fig. 16; the differential of the voltage drop after the first discharge is used to block subsequent discharges by means of a logic circuit. The holes produced in this way (see fig. 17*b*) are always of good quality, the size depends on the energy of the discharge.

The accuracy and speed of the high-precision spark methods described above, coupled with their ready adaptability for particular machining problems down to very small dimensions, should ensure a rapid expansion in their field of application.

Appendix: A thermal model

The change of temperature with depth in the metal has been calculated for normal electro-discharge machining by various workers using a thermal model [10] [11]. In a theoretical investigation into this process at the Philips Mechanical Engineering Works, J. H. Leemreis has calculated both this temperature distribution and the optimum duration of the spark discharge and the efficiency, all as functions of the fraction of the discharge energy dissipated in the electrode.

To obtain a theoretical insight into the *high-precision* spark-machining process and to test this process against the thermal model referred to above, the first requisite was experimental data. These were obtained with the following arrangement.

The workpiece and the electrode were both steel pins 0.3 mm in diameter, the spark distance was kept at 1-2 μm, and de-ionized water was used as the dielectric fluid. The spark generator was a pulse shaper giving square-wave pulses of current: the peak current was 1 A, the pulse time was adjustable between 20 and 100 ns and the repetition rate could be varied up to 50 kHz. *Fig. 18* shows the behaviour of the voltage and current during a single

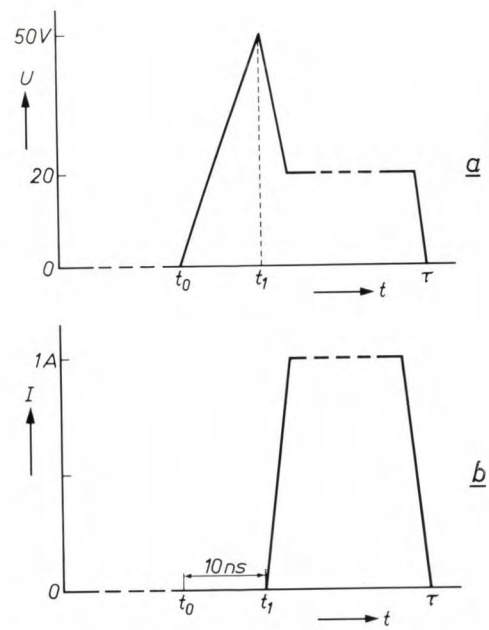


Fig. 18. Schematic representation of the discharges employed in a test rig for investigating spark erosion. *a*) The voltage U as a function of time. *b*) The current I as a function of time. The discharge begins about 10 ns after the start of the pulse; the discharge time is variable between 20 and 100 ns.

pulse. A discharge started 10 ns after the beginning of the pulse, followed by a voltage drop within 10 ns from the breakdown voltage (50 V) to a value of 20 V which remained constant for the rest of the discharge. The power during the discharge was constant at 20 W.

It was now important to find out how much of the power was dissipated in the workpiece and how much in the electrode, and also how the conversion of electric energy to heat varied as a function of place and time. Since the wear ratio S (= electrode wear/workpiece wear) in the particular equipment is a good indicator of the division of power between the workpiece and the electrode (since the shape of the two electrodes and the material of which they were made, which might affect this division of power, were identical), S was first measured as a function of the pulse length (*fig. 19a*). In the area under consideration the wear ratio is found to increase with the discharge time.

The amount of power dissipated in the workpiece can be calculated from S as follows:

$$P_w = (1 - S)P, \quad \ldots\ldots\ldots \quad (1)$$

where P is the (constant) discharge power (20 W).

To calculate the effect of a discharge the power density also has to be known. Although the power is constant during the discharge, the cross-sectional area of the discharge path increases with time. It can, however, be assumed that this area is the same for both electrode surfaces, and that it is equal to the cross-sectional area of the craters that are formed. We therefore measured the crater diameter as a function of the energy of the pulse by employing trains of pulses of different durations. The cross-sectional area of the discharge path at any moment during the discharge can be derived from the resultant curve (fig. 19*b*), so that (with the aid of equation 1) the change in the amount of power dissipated in the workpiece per unit of surface area during discharge can be calculated (fig. 19*c*).

The heat developed at the surface is conducted into the workpiece and melts or vaporizes the metal down to a certain depth. An important piece of information derived from the experiments is that the diameter of the very small craters is considerably larger than their depth ($d/h \geqq 10$). This means that the model we should use for this heat conduction problem is not that of a point source of heat but rather that of a flat heat source of infinite extension, from which heat is conveyed in a direction perpendicular to the metal surface [10] [11].

It is possible to use this model and the curve of fig. 19*c* to calculate the temperature as a function of place and time [12]. This involves a great deal of computation, however, and since the main thing we are interested in is whether or not the surface of the workpiece is raised to its melting point under the conditions used for the sparking process, we can apply the following simplification. After 30 ns (a normal pulse duration) the power is 1.4 W/μm². If we now assume that the heat flow at the surface ($x = 0$) has been equal to this value (F_0) during the entire discharge between 10 and 30 ns, then the temperature should increase as a function of time:

$$T(t)_{x=0} = \frac{2F_0}{\lambda}\sqrt{\frac{at}{\pi}}, \quad \ldots\ldots \quad (2)$$

where λ is the thermal conductivity of the metal and $a = \lambda/\sigma c$, the temperature equalization coefficient (σ is the density, c the specific heat). For a discharge time of $t = 20$ ns, T is found to be about 17 000 °C. This is in fact far higher than the boiling point T_b (3000 °C), so that we are compelled to correct the specific heat c for the latent heats of fusion m and vaporization v and make a fresh calculation. This corrected specific heat is then approximately equal to:

$$c' = c + \frac{m + v}{T_b} \; (= 2510 \text{ J/kg °C}) \quad (3)$$

Recalculation of the temperature for $x = 0$ after a 30 ns pulse yields $T = 7200$ °C, a value which is also considerably higher than the boiling point.

Our next step consists in calculating the depth to which the metal vaporizes. For a constant heat flow F_0 we have [13]:

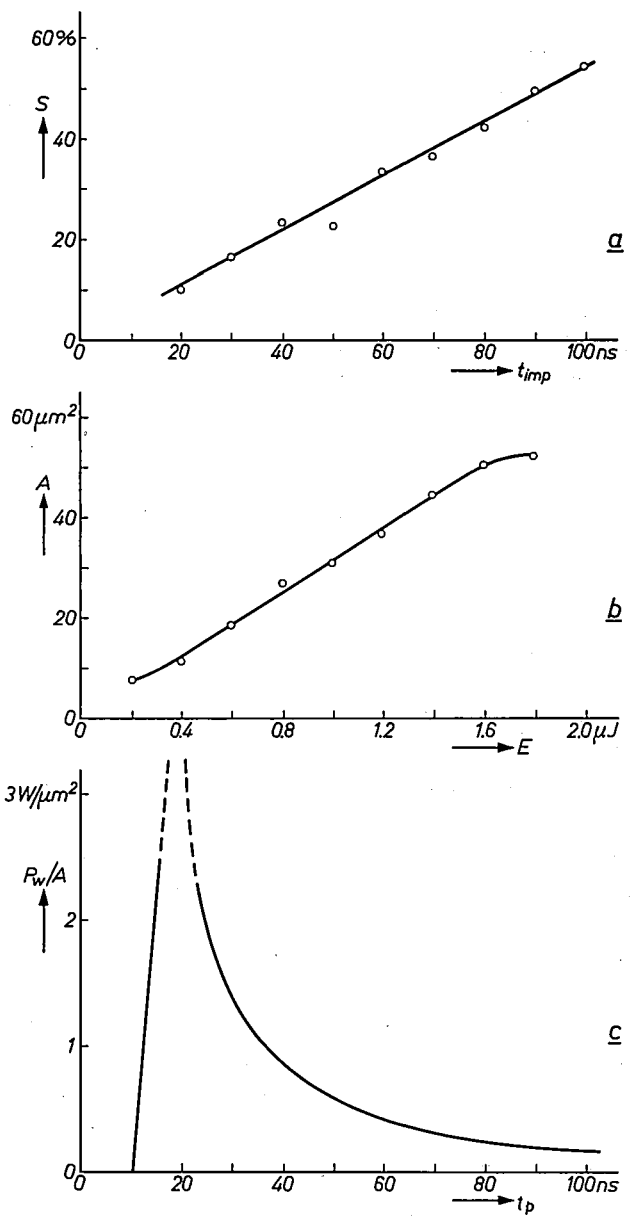


Fig. 19. *a*) Wear ratio S as a function of the pulse duration t_p for two steel electrodes using the discharge waveform shown in fig. 18. The dielectric fluid was de-ionized water. The electrode from which more material was removed (the "workpiece") in this experiment was the electrode connected as the anode. *b*) Cross-sectional area A of the discharge path as a function of the energy E per discharge. *c*) The power dissipated in the anode per unit surface area as a function of the pulse duration.

[10] A. S. Zingerman, Soviet Phys. Solid State **1**, 255, 1959.
[11] J. H. Leemreis, private communication.
[12] H. S. Carslaw and J. C. Jaeger, Conduction of heat in solids, Clarendon Press, Oxford 1959, pp. 75-76.
[13] The function in square brackets (the function ierfc, which is important in error theory) is tabulated in [12] for the argument $x/\sqrt{4a't} = u$.

$$T(t,x) = \frac{2F_0}{\lambda}\sqrt{\frac{a't}{\pi}}\left[\exp\left(-\frac{x^2}{4a't}\right) - \frac{x}{\sqrt{4a't}}\int_{x/\sqrt{4a't}}^{\infty} e^{-\xi^2}\,d\xi\right], \quad (4)$$

where a' is the corrected heat equalization coefficient $\lambda/\sigma c'$. If we take $T(t,x) = T_b = 3000$ °C and $t = 20$ ns (the discharge time), it follows that $x = 0.17$ μm.

There will, of course, be a molten zone underneath the vaporization zone. To calculate the depth to which the material has been melted we can use standard thermodynamic data to give a figure of 70% for the fraction of the heat flow involved in the vaporization, so that 30% remains for melting (and raising to the melting point). We therefore introduce a heat flow $F_0' = 0.3\,F_0$. The specific heat also has to be corrected again, but this time only for the latent heat of fusion: $c'' = c + m/T = 653$ kJ/g °C, so that $a'' = \lambda/\sigma c'' = 8.8\times10^{-6}$ m²/s. If we substitute a'' for a' in eq. (4) and the melting point (1500 °C) for T, we find that as a result of the discharge the metal has been melted to a depth $x = 0.39$ μm (as we have already seen 0.17 μm of this is subsequently vaporized).

We have taken a heat source of indefinite extent in our model; in fact, of course, the discharge path has a limited diameter. Our calculations are based upon a pulse duration of 30 ns (see above) and we have assumed that the power was always 1.4 W/μm² during the discharge time of 20 ns (see fig. 19*c*). During that time an energy of 20 W × 20 ns = 0.4 μJ was transferred, and fig. 19*b* shows that the cross-sectional area of the discharge path was then 12 μm², which means that the diameter d is equal to 3.8 μm. Assuming that the crater is concave, the situation is as shown in *fig. 20*.

It is known from the literature on conventional spark machining [14] that not only vaporized material is removed from the surface of the workpiece during the discharge but particles of molten material as well. There are various suggestions as to the nature of the forces involved in the removal of the liquid particles, including: 1) electrodynamic forces which are effective down to a certain depth in the material during the discharge, 2) mechanical forces which arise mainly after the current flow has ceased, as a result of the sudden drop of pressure in the discharge path (comparable to the cavitation forces that occur in a liquid subjected to ultrasonic vibrations, causing bubbles of liquid to implode).

The volume of the removed metal can be calculated from $V = \frac{1}{6}\pi h(h^2 + 3r^2)$, where h is the depth and r the radius of the crater. If the vaporized metal alone were removed, this relation would show that the efficiency of the process was

$$9.65\times10^{-10}/0.4\times10^{-6} = 2.4\times10^{-3}\ \text{mm}^3/\text{J}.$$

If the molten metal is also removed, the total efficiency becomes

$$22.4\times10^{-10}/0.4\times10^{-6} = 5.6\times10^{-3}\ \text{mm}^3/\text{J}.$$

These calculations can be compared with the results obtained in actual practice for a diaphragm hole of 50 μm diameter spark-eroded in 6 seconds in a 50 μm thick steel plate. The repetition rate was 15 kHz, but only 80% of the pulses gave discharges, so that a total of 72 000 discharges with an energy of 0.4 μJ was required. The wear ratio was 8%, so that an energy of 0.37 μJ per discharge was dissipated in the workpiece. The efficiency was therefore

$$\pi\times0.025^2\times0.050/0.37\times10^{-6}\times72\,000 = 3.7\times10^{-3}\ \text{mm}^3/\text{J}.$$

This shows that there is good agreement between the calculation and the experimental result. There is therefore every reason to assume that high-precision spark machining is in fact a thermal process which corresponds approximately with the model considered (and is not, for instance, based on mechanical explosive forces). Although the model itself is only fairly approximate, it can be concluded that part of the molten metal is also removed in addition to the vaporized metal.

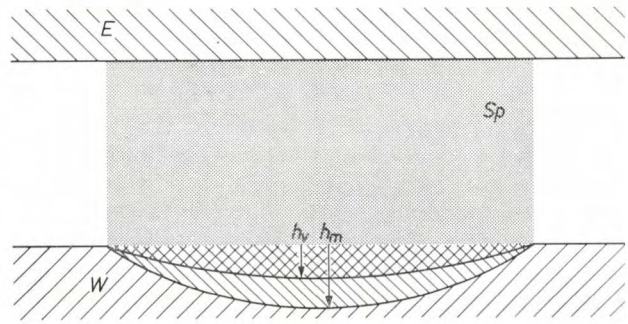


Fig. 20. The calculated material heating as a result of a discharge. *E* is the electrode, *W* the workpiece, *Sp* the discharge path. h_v is the depth to which the material vaporizes, h_m the depth to which it is melted.

Summary. The process known as spark machining depends on the interaction of a large number of electrical parameters and properties of materials. The work described in this article has made it possible to use this process for precision work and to achieve a relatively high machining rate. This has been accomplished by using very short discharges of low energy and a high repetition rate. The use of de-ionized water as the dielectric fluid is found in many cases to offer considerable advantages. A number of spark machines are described in which the principles discovered in this work have been applied: a high-precision spark machine with which dimensional tolerances as low as 0.5 μm can be achieved, and a high-speed spark machine that will give tolerances down to about 1.5 μm and machining rates up to 5 mm³ per minute. Descriptions are given of a special version for semi-automatic spark erosion of holes in diaphragm plates for electron-optical systems, a contouring machine with a photoelectric line-tracking system, and a machine with piezoelectric spark-gap regulation for making very accurate slots and very small holes e.g. in thin metal coatings. An Appendix gives a calculation of the energy dissipation, made with the aid of a heat-flow model; good agreement is found to exist between calculations and the results found in practice.

[14] B. A. Krasyuk (editor), Electrospark machining of metals, Consultants Bureau, New York 1965, part 3 (translated from Russian).

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

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.

C. S. Aitchison & J. C. Williams: Active reactance compensation of parametric amplifiers.
Electronics Letters **5**, 139-140, 1969 (No. 7). *M*

D. Andrew: The development of sputter-ion pumps.
Proc. 4th Int. Vacuum Congress, Manchester 1968, p. 325-331. *M*

D. Andrew, D. R. Sethna & G. F. Weston: Inert-gas pumping in a magnetron pump.
Proc. 4th Int. Vacuum Congress, Manchester 1968, p. 337-340. *M*

G. Arlt & P. Quadflieg: Electronic displacement in tellurium by mechanical strain.
Phys. Stat. sol. **32**, 687-689, 1969 (No. 2). *A*

K. H. Beckmann & R. Memming: Photoexcitation and luminescence in redox processes on gallium phosphide electrodes.
J. Electrochem. Soc. **116**, 368-373, 1969 (No. 3). *H*

J. van den Berg & H. J. R. Perdijk (Philips Electronic Components and Materials Division, Eindhoven): Barium carbonate deposits on valve components.
Vacuum **19**, 122, 1969 (No. 3).

G. Blasse: Energy transfer in oxidic phosphors.
Philips Res. Repts. **24**, 131-144, 1969 (No. 2). *E*

G. Blasse: Crystallographic and fluorescence data of calcium yttrium borate ($CaYBO_4$).
J. inorg. nucl. Chem. **31**, 1519-1521, 1969 (No. 5). *E*

G. Blasse & A. Bril: Luminescence of europium-activated strontium yttrium oxide (SrY_2O_4-Eu).
J. inorg. nucl. Chem. **31**, 1521-1523, 1969 (No. 5). *E*

A. J. Bosman & C. Crevecoeur: Electrical conduction in Li-doped CoO.
J. Phys. Chem. Solids **30**, 1151-1160, 1969 (No. 5). *E*

P. C. Brandon: Artifacts in intracellular enzyme distribution caused by alkaline homogenates.
Plant Physiol. **44**, 461-462, 1969 (No. 3). *E*

P. B. Braun, J. Hornstra & J. I. Leenhouts: Automated crystal-structure determination by Patterson search using a known part of the molecule.
Philips Res. Repts. **24**, 85-118, 1969 (No. 2). *E*

J. van den Broek: Optical lattice vibrations and dielectric constant of tetragonal lead monoxide.
Philips Res. Repts. **24**, 119-130, 1969 (No. 2). *E*

E. Bruninx & J. Crombeen: Thick target neutron yields and neutron spectra produced by 20 MeV helium-3 ions, 14 MeV protons and 7.5 MeV deuterons on a beryllium target.
Int. J. appl. Rad. Isot. **20**, 255-264, 1969 (No. 4). *E*

K. H. J. Buschow: The magnetic properties of the compound Dy_3Al_2.
Physics Letters **29A**, 12-13, 1969 (No. 1). *E*

K. H. J. Buschow, H. W. de Wijn (Natuurkundig Laboratorium der Universiteit van Amsterdam) **& A. M. van Diepen** (N. L. Univ. Amst.): Magnetic susceptibilities of rare-earth-indium compounds: RIn_3.
J. chem. Phys. **50**, 137-141, 1969 (No. 1). *E*

H. B. G. Casimir: G. Holst and his ideas on industrial research.
Philips Res. Repts. **24**, 161-167, 1969 (No. 3). *E*

T. Chisholm: An analysis of thermal and space-charge spreading of electron beams from guns of the Pierce type.
Proc. IEEE **57**, 357-358, 1969 (No. 3). *M*

W. A. Crossley, R. W. Cooper, J. L. Page & R. P. van Stapele: Faraday rotation in rare-earth iron garnets.
J. appl. Phys. **40**, 1497-1498, 1969 (No. 3). *E, M*

P. J. Daniel, R. A. Ford & J. M. Schofield: Transistor fabrication using diffusion barriers produced by electron beams.
Electronics Letters **5**, 169-170, 1969 (No. 8). *M*

J. C. Decroly, Y. Genin & J. M. Mouchart: La détermination des orbites de transfert.
Rev. MBLE **11**, 145-158, 1968 (No. 4). *B*

J. A. W. van der Does de Bye: Comment on the application of configuration coordinates to the red emission of gallium phosphide.
J. Phys. Chem. Solids **30**, 1293, 1969 (No. 5). *E*

J. A. W. van der Does de Bye & R. C. Peters: Preparation and properties of epitaxial gallium phosphide grown by HCl-gas transport.
Philips Res. Repts. **24**, 210-230, 1969 (No. 3). *E*

G. Domburg & B. Lopes Cardozo (Institute for Perception Research, Eindhoven): Dropout. An investigation into its effects on recorded music.
Tape Recorder **11**, 312-313, 1969 (No. 8).

C. Z. van Doorn: Demonstration of electrostatic generator principles.
Amer. J. Phys. **37**, 225, 1969 (No. 2). *E*

D. G. Elson: An analogue equipment for use in production planning.
Operat. Res. Quart. **19**, 175-184, 1968 (No. 2). *M*

N. V. Franssen: Geluidsveldversterking.
Nederlands Akoestisch Genootschap Publ. No. 14, 54-58, 1969. *E*

S. Garbe & G. Frank: Efficient photoemission from GaAs epitaxial layers.
Solid State Comm. **7**, 615-617, 1969 (No. 8). *A*

J.-M. Goethals: A polynomial approach to linear codes.
Philips Res. Repts. **24**, 145-159, 1969 (No. 2). *B*

C. A. A. J. Greebe & P. A. van Dalen: Ultrasonic amplification by minority charge carriers in piezoelectric semiconductors.
Philips Res. Repts. **24**, 168-200, 1969 (No. 3). *E*

G. Groh: Holographie in der Schwingungstechnik.
VDI-Berichte No. 135, 145-153, 1969. *H*

G. E. G. Hardeman & G. B. Gerritsen: Magnetic resonance in 6H SiC.
Mat. Res. Bull. **4**, S 261-272, 1969 (SiC issue). *E*

J. 't Hart (Institute for Perception Research, Eindhoven): Fonetische steunpunten.
Nieuwe Taalgids **62**, 168-174, 1969 (No. 3).

J. Hasker: Transverse-velocity distribution for space-charge-limited beam current in an electron gun with rotational symmetry.
Philips Res. Repts. **24**, 231-240, 1969 (No. 3). *E*

E. E. Havinga, M. H. van Maaren & H. Damsma: Superconductivity of compounds having the Si_2U_3-type structure.
Physics Letters **29A**, 109-110, 1969 (No. 3). *E*

K. Herff & E. Roeder: An apparatus for measuring the thickness of thin layers.
Pract. Metallogr. **5**, 557-566, 1968 (No. 10). *A*

A. E. Hilling & S. K. Salmon: Intermodulation in common-emitter transistor amplifiers.
Electronic Engng. **40**, 360-364, 1968 (No. 485). *M*

E. P. Honig & J. H. Th. Hengst: Points of zero charge of inorganic precipitates.
J. Colloid Interface Sci. **29**, 510-520, 1969 (No. 3). *E*

E. P. Honig & J. H. Th. Hengst: Temperature dependence of the point of zero charge of AgI.
J. Colloid Interface Sci. **30**, 109-110, 1969 (No. 1). *E*

F. N. Hooge: $1/f$ noise is no surface effect.
Physics Letters **29A**, 139-140, 1969 (No. 3). *E*

P. J. W. Jochems: Geïntegreerde schakelingen, 1. De planaire technologie.
Natuur en Techniek **37**, 173-181, 1969 (No. 5). *E*

M. de Jong & M. G. J. Knapen (Philips Electronic Components and Materials Division, Eindhoven): Cutting of thin carbon films with a CO_2-gas laser, II. Experiments on cylindrical bodies.
Microelectronics and Reliability **8**, 89-90, 1969 (No. 2).

H. Jonker, C. J. Dippel, H. J. Houtman, C. J. G. F. Janssen & L. K. H. van Beek: Physical development recording systems, I. General survey and photochemical principles.
Photogr. Sci. Engng. **13**, 1-8, 1969 (No. 1). *E*

Y. Kamp: Synthesis of a multivariable 2-port using parallel and series stubs.
Electronics Letters **4**, 474-475, 1968 (No. 22). *B*

Y. Kamp & J. Neirynck: Maximally flat approximation for noncommensurate cascaded transmission lines.
Electronics Letters **4**, 477-478, 1968 (No. 22). *B*

W. Kebschull: Die Reflexion trockner und feuchter Straßenbeläge.
Thesis, Berlin 1968. *A*

W. F. Knippenberg & G. Verspui: The influence of impurities on the growth of silicon carbide crystals grown by gas-phase reactions.
Mat. Res. Bull. **4**, S 33-44, 1969 (SiC issue). *E*

W. F. Knippenberg & G. Verspui: The influence of impurities on the growth of silicon carbide crystals by recrystallization.
Mat. Res. Bull. **4**, S 45-55, 1969 (SiC issue). *E*

J. W. L. Köhler: The Stirling refrigeration cycle in cryogenic technology.
Advancement of Science, March 1969, 261-268. *E*

D. Lamport: Image tubes shaped by computer.
New Scientist **43**, 10 July 1969, feature section, p. 21, 23, 24 (No. 657). *M*

J. Loeckx: Mechanical construction of bounded-context parsers for Chomsky 0-type languages.
Thesis, Louvain 1968. *B*

S. R. Longley: Experimental 4-port *E*-plane junction circulators.
IEEE Trans. **MTT-15**, 378-380, 1967 (No. 6). *M*

F. K. Lotgering: Magnetic and electrical properties of $Co_{1-x}Cu_xRh_2S_4$.
J. Phys. Chem. Solids **30**, 1429-1434, 1969 (No. 6). *E*

J. R. Mansell: Video-frequency light modulator having a wide angular aperture.
Proc. IEE **116**, 691-698, 1969 (No. 5). *M*

R. Memming & G. Neumann: Electrochemical reduction and hydrogen evolution on germanium electrodes.
J. electroanal. Chem. **21**, 295-305, 1969 (No. 2). *H*

R. F. Mitchell & M. Redwood (Queen Mary College, University of London): The generation of sound by non-uniform piezoelectric materials.
Ultrasonics **7**, 123-124, 1969 (No. 2). *M*

J. Neirynck & Ph. van Bastelaer: A class of filters with optimum transient behaviour.
Rev. MBLE **11**, 159-166, 1968 (No. 4). *B*

A. K. Niessen & C. H. Weijsenfeld: Anisotropic pinning and guided motion of vortices in type-II superconductors.
J. appl. Phys. **40**, 384-393, 1969 (No. 1). *E*

S. G. Nooteboom (Institute for Perception Research, Eindhoven): Hardop lezen als een vorm van continu taalgebruik.
Forum der Letteren **10**, 95-106, 1969.

A. van Oostrom: Field electron emission from alloys.
Proc. 4th Int. Vacuum Congress, Manchester 1968, p. 275-279. *E*

D. H. Paul: Use of the complex variable and the bilinear transformation in circuit analysis.
Electronic Engng. **40**, 378-382, 1968 (No. 485). *M*

D. H. Paul: Chebyshev polynomials.
Electronic Engng. **40**, 475, 1968 (No. 486). *M*

K. F. Poole (Philips International Institute, Eindhoven) **& A. Venema:** An investigation of whisker formation and growth associated with electrical breakdown in vacuum.
Proc. 4th Int. Vacuum Congress, Manchester 1968, p. 271-274. *E*

D. Pruis: Using computers for drawing (including perspective drawing) and manufacturing special products. Automation in the drawing office, Proc. Spring Conf. London 1969, 7 pp. *E*

J.-J. Robillard: Sur l'utilisation de la photodésorption de l'oxyde de zinc pour le couplage des sels de diazonium dans les procédés de reproduction diazo.
C.R. Acad. Sci. Paris **267C**, 1279-1282, 1968 (No. 20). *L*

W. Schilz: Magneto-acoustic investigation of the Fermi surface and spin splitting of the Landau levels in *p*-type and *n*-type PbTe.
J. Phys. Chem. Solids **30**, 893-901, 1969 (No. 4). *H*

E. Schwartz: Rise time and pulse length limitation for a source with parasitic capacitance and related problems.
IEEE Trans. **CT-16**, 96-98, 1969 (No. 1). *A*

J. M. Shannon: Simple expressions for the effect of substrate doping on the gain of m.o.s.t.s.
Electronics Letters **5**, 181-182, 1969 (No. 9). *M*

J. G. Siekman: Cutting of thin carbon films with a CO_2-gas laser, I. Theoretical considerations and experiments on flat plates.
Microelectronics and Reliability **8**, 87-89, 1969 (No. 2). *E*

L. A. Æ. Sluyterman, J. Wijdenes & B. G. Wolthers (Laboratorium voor Structuurchemie, Groningen): The non-essentiality of the N-terminal amino group of papain.
Biochim. biophys. Acta **178**, 392-393, 1969 (No. 2). *E*

F. L. H. M. Stumpers: Doel en structuur van de Internationale Wetenschappelijke Radio Unie.
Ingenieur **81**, ET 80-82, 1969 (No. 20). *E*

T. L. Tansley: Characteristics of GaAs based heterojunction photodetectors.
Proc. 2nd Int. Symp. on GaAs, Dallas, Texas, 1968, p. 222-229. *M*

R. W. Teale (Physics Department, Sheffield University), **D. W. Temple** (Phys. Dept., Sheffield Univ.), **U. Enz & R. F. Pearson:** Infra-red sensitive ferrimagnetics.
J. appl. Phys. **40**, 1435-1441, 1969 (No. 3). *E, M*

H. J. L. Trap: Elektronenleitende Gläser, ihre Eigenschaften und Anwendungsmöglichkeiten in der Elektronik.
Sprechsaal für Keramik, Glas, Email, Silikate **101**, 1103-1111, 1114, 1968 (No. 24). *E*

A. Venema: Recent trends in vacuum studies for electron tubes.
Proc. 4th Int. Vacuum Congress, Manchester 1968, p. 267-270. *E*

G. Verspui: Nieuwe vormen van koolstof.
Natuur en Techniek **37**, 113-123, 1969 (No. 4). *E*

J. Vlietstra: De APT-post-processor. Een rekenmachineprogramma voor het verwerken van APT-werkstukprogramma's door het APT-computersysteem.
Ingenieur **81**, W 33-39, 1969 (No. 7). *E*

W. L. Wanmaker & J. W. ter Vrugt (Philips Lighting Division, Eindhoven): Luminescence of some gallium-containing compounds.
Philips Res. Repts. **24**, 201-209, 1969 (No. 3).

J. D. Wasscher: Electrical transport phenomena in MnTe, an antiferromagnetic semiconductor.
Thesis, Eindhoven 1969. *E*

F. W. Willmott: Pyrolysis-gas chromatography of polyolefins.
J. chromatogr. Sci. **7**, 101-108, 1969 (No. 2). *M*

P. Wodon: Langages de programmation pour le calcul symbolique.
Rev. MBLE **11**, 167-178, 1968 (No. 4). *B*

Contents of Philips Telecommunication Review **28**, No. 2, 1969:

Th. J. Nieland: 300 kW shortwave transmitter type FB 010 (p. 49-58).
A. Boesveld, J. C. de Granje, J. Vernooy & H. van der Ploeg: Automatic transmission measuring equipment for international telephone lines (A.T.M.) (p. 59-87).
A. Tarassoff & C. Ribeyre: Attenuation peaks on open-wire lines and the advantages of using phantom circuits (p. 89-119).

Contents of Philips Telecommunication Review **28**, No. 3, 1969:

F. J. Schramel: The DS 714 computer system used as a message switcher (p. 125-134).
H. van Kampen: The type DS 714 computer-based message and data switching system (p. 135-146).
C. H. Luitwieler: A computer interface for the ELDO launcher inertial guidance system (p. 148-154).
J. K. A. Poppe: PE-6 negistor system 8TR 321/10 (p. 155-157).

Contents of Mullard Technical Communications **10**, No. 99, 1969:

R. J. Hatt, A. E. Jackets & D. B. Jarvis: Four-phase logic circuits using integrated m-o-s transistors (p. 266-276).
J. Merrett: Dynamic braking of d.c. shunt motors (p. 277-284).

Contents of Mullard Technical Communications **10**, No. 100, 1969:

K. Wilson: Gunn effect devices and their applications (p. 286-293).
A. Ciuciura: Spark-gaps for protection of radio receivers (p. 294-295).
D. Gent & J. Lammers: Line timebase and e.h.t. trebler for colour tubes (p. 296-301).
W. Ebbinge & D. C. de Ruiter: Transistorised water level sensor (p. 302-304).

Contents of Valvo Berichte **15**, No. 1, 1969:

F. Weitzsch: Zur Frage der Farbkorrekturen bei Farbfernseh-Empfängern (p. 1-19).
J. Koch: Berechnung der Wirbelstromverluste in der Wicklung von Schalenkernen mit Luftspalt (p. 20-31).
D. Hellwege: Messung der Wirbelstromverluste in der Wicklung von Schalenkernen mit Luftspalt (p. 32-40).

Contents of Valvo Berichte **15**, No. 2, 1969:

A. Petersen: Wirbelstromkupplungen und Dämpfungssysteme mit Permanentmagneten (p. 41-57).
E. Drossel & G. Samow: Ferroxdure SP 130, ein anisotroper kunststoffgebundener Bariumferrit-Werkstoff (p. 58-63).
J. Koch, K. Ruschmeyer & H. Schwang: Temperaturabhängigkeit der magnetischen Induktion von Ferroxdure (p. 64-77).

PHILIPS TECHNICAL REVIEW

VOLUME 30, 1969, No. 8/9/10

The various laboratories that contribute to the research effort of the Philips Group of Companies include the Laboratoires d'Electronique et de Physique Appliquée, which are situated near Paris. Their organization differs to some extent from those of the other cooperating laboratories, their research programme shows striking individualities, and their buildings are also highly individual in their architecture.

This issue of our journal is entirely devoted to the work of these French laboratories. As in similar issues in the past, devoted to the laboratories in Germany and England, an introductory article by the Director of the Laboratories sketches their history and gives a broad review of the work accomplished and in progress. This is followed by a series of articles, written by members of the Laboratories, on special topics of current research. The editing of these articles has largely been carried out by members of the staff of the Laboratories.

The Laboratoires d'Electronique et de Physique Appliquée at Limeil-Brévannes near Paris

Origin — Function — Activities

C. Ducot

Origin

In 1948/49 a number of laboratories interested in advanced techniques related to electronics and applied physics were brought together under the direction of Professor G.-A. Boutry in premises at 23 Rue du Retrait in Paris.

The first nucleus of what was to become the Laboratoires d'Electronique et de Physique Appliquée consisted of an Electron Tube and Vacuum Technique division, a Special Circuit division, mainly concerned with research on television, an Optics division and a small workshop, and these were soon joined by a Microwave Circuit division and a team of theoreticians. This modest undertaking was supported financially by several French companies of the Philips Concern, which also provided most of the general services required.

In 1950 it was felt that the time had come to unite these laboratories in a single research enterprise with its own more substantial general services, including administrative departments, stores, a design office and a mechanical workshop. This reorganization also permitted an integration of the research topics and a coordination of the auxiliary services required by the individual laboratories.

C. Ducot, ancien élève de l'Ecole Polytechnique, Director of the Laboratoires d'Electronique et de Physique Appliquée, Limeil-Brévannes (Val-de-Marne), France.

Thus on 4th September 1950 the Société Anonyme "Laboratoires d'Electronique et de Physique Appliquée" was formed, with a total staff of about 100 people. The shares of this Company — popularly abbreviated to "LEP" — are now held by the Société La Radiotechnique, the Société Télécommunications Radioélectriques et Téléphoniques, the Compagnie Française Philips and its subsidiaries Hyperélec, Philips Industrie and Philips Matériel Electronique Professionel.

The organization of LEP has always been rather unusual since its earliest days. In other comparable research laboratories the management usually consists entirely of representatives of the participating companies; but the Board of Management of LEP differs in that it includes a strong university representation. This assures good communications with the universities and at the same time gives an orientation towards the fundamental problems of physics.

Function

The function of the Laboratories is determined to a large extent by the nature of their organization. Before looking more closely into this function, we should perhaps consider what we understand by research.

In the past — and here we mean the period before the scientific and technological explosion of the last fifty years or so — a clear distinction could be made

between pure research and applied research. Pure research was intended to increase human knowledge and was generally found in the university, whereas applied research was intended to advance technology and was generally found in industry.

Today the situation is different. On one hand, the advancement of scientific knowledge now calls for facilities so elaborate and expensive that they can only be acquired by appealing to the economic motivations of industry or the political motivation of national prestige. On the other hand, technical progress now requires such a great total of knowledge that research which is apparently unconnected with any practical objective may well provide the quickest path to the resources of nature.

In such a situation it is becoming more and more difficult to distinguish any kind of boundary between pure and applied research, and the role of an industrial research organization such as LEP can no longer be located at some point of a graduated scale with "theoretical purity" at one end and "practical usefulness" at the other. More appropriately, the criterion for defining this role should be derived from the stages of evolution between initial research and industrial production. These stages are:

a) "*Basic*" *research* which, although its direct object is the increase of knowledge in a certain field, is already oriented by some indication of possible future applications. This stage extends as far as the demonstration that a given effect is capable of producing a given result.

b) *Product-oriented research*, which aims at obtaining a component, device or piece of apparatus that fulfils prescribed functions. This stage normally ends with samples or an experimental device which have not yet been subjected to all environmental conditions, reliability tests, etc.

c) "*Development*", which aims at designing a component, device or equipment fulfilling the same functions in a defined environment and with guaranteed reliability. This stage ends with a prototype.

d) The study and implementation of *production methods* for making products equivalent to the prototype as economically as possible.

Of course, these stages are not always so distinct as this scheme would suggest. Nevertheless it can be seen that in none of them is the research completely "pure", and that the theoretical disciplines predominate in the first stage but progressively give way in the later stages to more technical activities.

It is probably fair to say that the functional purpose of an industrial research laboratory is to be found in the first and second of these four stages. There are exceptions: for example, the study of a new product that derives from a particular technique already well established in industry will obviously be carried out in the development laboratory attached to the appropriate factory. With completely new techniques, on the other hand, it may be necessary to begin pilot production in the research laboratory if the factory does not yet have the necessary equipment and the staff to operate it. And in the end, the technical requirements of quantity production may even affect some of the objectives at the first stage, fundamental research itself.

In accordance with both the general description given above and the organizational environment of LEP, the work undertaken there can be divided into three main categories with different kinds of support:

a) Basic long-term research (over a minimum period of 5 years), which is not necessarily directed towards any specified product but because of its expected impact on future applications may be backed by large associated Companies or by a main Product Division.

b) Research coming within the general programme of the Laboratory but already of more direct interest to one or other of its associates with a specific end in view and, of course, supported by that associate. The work involved covers a shorter period (1 to 3 years). Work of this type is in particular done for R.T.C. La Radiotechnique-Compelec on components. Work on equipment and systems is supported chiefly by the other associates, which now include a division of the Société d'Etudes et de Réalisations Nucléaires (S.O.D.E.R.N.), specializing for example in nuclear or space techniques that are not suitable for quantity production.

c) Advanced studies within the competence of the Laboratory, commissioned and subsidized by State organizations such as C.E.A. (Commissariat à l'Energie Atomique), S.E.F.T. (Section d'Etudes et Fabrications des Télécommunications), C.N.E.S. (Centre National d'Etudes Spatiales), L.R.B.A. (Laboratoire de Recherches Balistiques et Aérodynamiques), and O.R.T.F. (Office de Radiodiffusion-Télévision Française), with a view to well-defined applications. This work is arranged by contracts, usually for periods of 1 to 2 years, which are often renewed if the first results show sufficient promise. In this category we should also mention the support received from certain organizations set up to promote research, such as D.G.R.S.T. (Délégation Générale à la Recherche Scientifique et Technique), D.R.M.E. (Direction des Recherches et Moyens d'Essais) and the Délégation à l'Informatique.

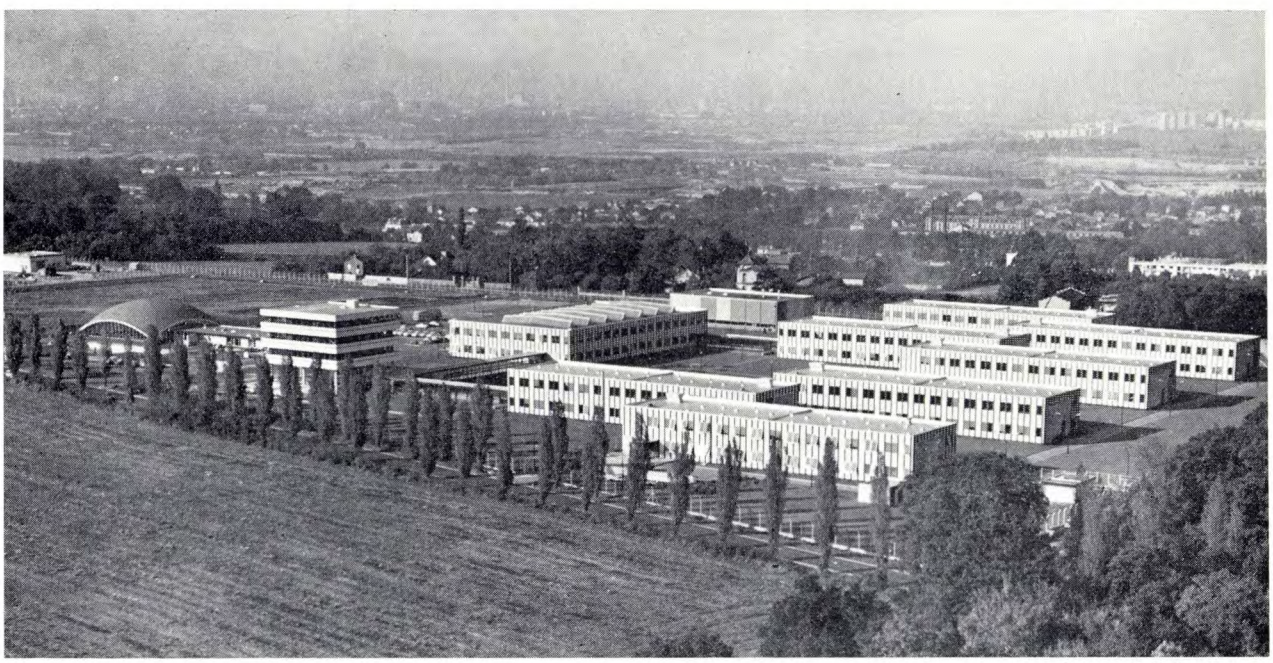
The scientific and technical specialization of the Laboratories was determined at the outset by the particular capabilities of the research teams and by the general trends which were then becoming apparent in electronics and applied physics in industry. This

meant that the field where the Laboratories could best contribute lay in the border regions where optics and electronics meet, and the first main topics that were most appropriate to the activities of LEP were photoelectricity, television, microwave transmission and infra-red systems.

This special interest in the fields where optics and electronics merge still applies today, although the terms of reference have naturally become broader and the changes that these subjects have undergone in the last twenty years have made their mark: the advent of semiconductors, microminiaturization and integrated circuits, the increasing use of computers, quantum electronics, non-linear optics, the conquest of space, and the like.

and 1965 at 3 Avenue Descartes, Limeil-Brévannes (near Paris), a site of 11 hectares (25 acres). All departments of LEP have been accommodated at this site since April 1965. The total floor area occupied is 15 000 m^2, about half of this being used for general purposes, covered ways, lecture rooms, reception and conference rooms, and the other half for laboratories and offices. A view of the buildings can be seen in *fig. 1*, a plan is given in *fig. 2*, and a description of the facilities will be found in reference [1] [*].

It can be seen from the illustrations that the work of the Laboratories is carried out in a group of buildings forming a complex. In the design of the buildings and the internal arrangements such as the distribution of gases, water, etc., every possible opportunity for standardization was taken, to allow a high degree of flexibility for the inevitable future modifications in the accommodation for staff and equipment. However, a few special variations were necessary, for example in the floor coverings and air conditioning systems, to suit the different kinds of operations carried out in the different buildings.



Photograph F. Dengremont

Fig. 1. The buildings of the Laboratoires d'Electronique et de Physique Appliquée at Limeil-Brévannes.

The Laboratories today

The Laboratories today employ a staff of about 300. A quarter of these are science and engineering graduates, half are technicians and the others are general administrative personnel. The scientific and technical staff are divided among five divisions: Electron Tubes and Vacuum Techniques, Chemistry, Semiconductors and Optics, General Electronics, Theoretical Physics and Applied Mathematics.

The considerable growth in the numbers of the staff made it necessary to transfer the Laboratories from Rue du Retrait to larger and more suitable premises.

The present laboratories were built between 1963

While the general function and field of work of the Laboratories remain substantially the same as at the start, the directions taken by the current research have to take into account two trends that are apparent in

[*] *Literature references are listed at the end of the article.*

the electronics and applied physics of today:
a) The previously exclusive domains of vacuum electronics are fast yielding to solid-state effects and solid-state devices.
b) Previously unconnected research topics give rise to cross-fertilization and sometimes lead to projects in which there are contributions from two or more technologies.

This does not mean that vacuum-tube electronics is undergoing a technical eclipse. Although semiconductors are encroaching on its traditional preserves, there is no shortage of new applications for the vacuum tube. Examples are the industrial applications of neutron tubes, image pick-up tubes, image converters or intensifiers, electronic shutter tubes for high-speed photography, and high-performance oscilloscope tubes.

With regard to outside contacts, relations with the large national organizations mentioned earlier have been intensified and extended. This is particularly relevant for some of the activities of great topical interest such as space research. In this field new relations have been established with European organizations such as ELDO (European Space Vehicle Launcher Development Organization), and ESRO (European Space Research Organization). Such contacts and cooperative arrangements with large organizations are usually established either by LEP alone or in connection with its associated Companies, depending on the relative amount of true research involved in a project.

On the other hand, excellent relations with the Philips Research Laboratories in Eindhoven and other associated laboratories in various European countries give LEP access to numerous technologies whose development is not directly a part of the research activities at LEP. One particularly valuable aspect of these relations is the systematic exchange of personnel between these laboratories. There are of course cases where an open exchange of information is ruled out by the confidential nature, for one reason or another, of the work. It is relatively seldom, however, that any such obstacle is placed in the way of the exchanges relating to the demonstration or interpretation of fundamental physical effects — and these are probably the most fruitful contacts of all.

Finally, in keeping with the motives that led to their foundation, our Laboratories have always had a close contact with the science faculties of the universities, and in recent years such contacts have also been extended to other organizations. Thus, apart from the long-standing cooperation with the Conservatoire National des Arts et Métiers and with various schools of technology, joint research projects have been undertaken with the Faculté des Sciences of Lille, the Institut Pédagogique National, the Faculty of Medicine of Strasbourg University, to mention only a few.

Some results obtained in the various fields of research

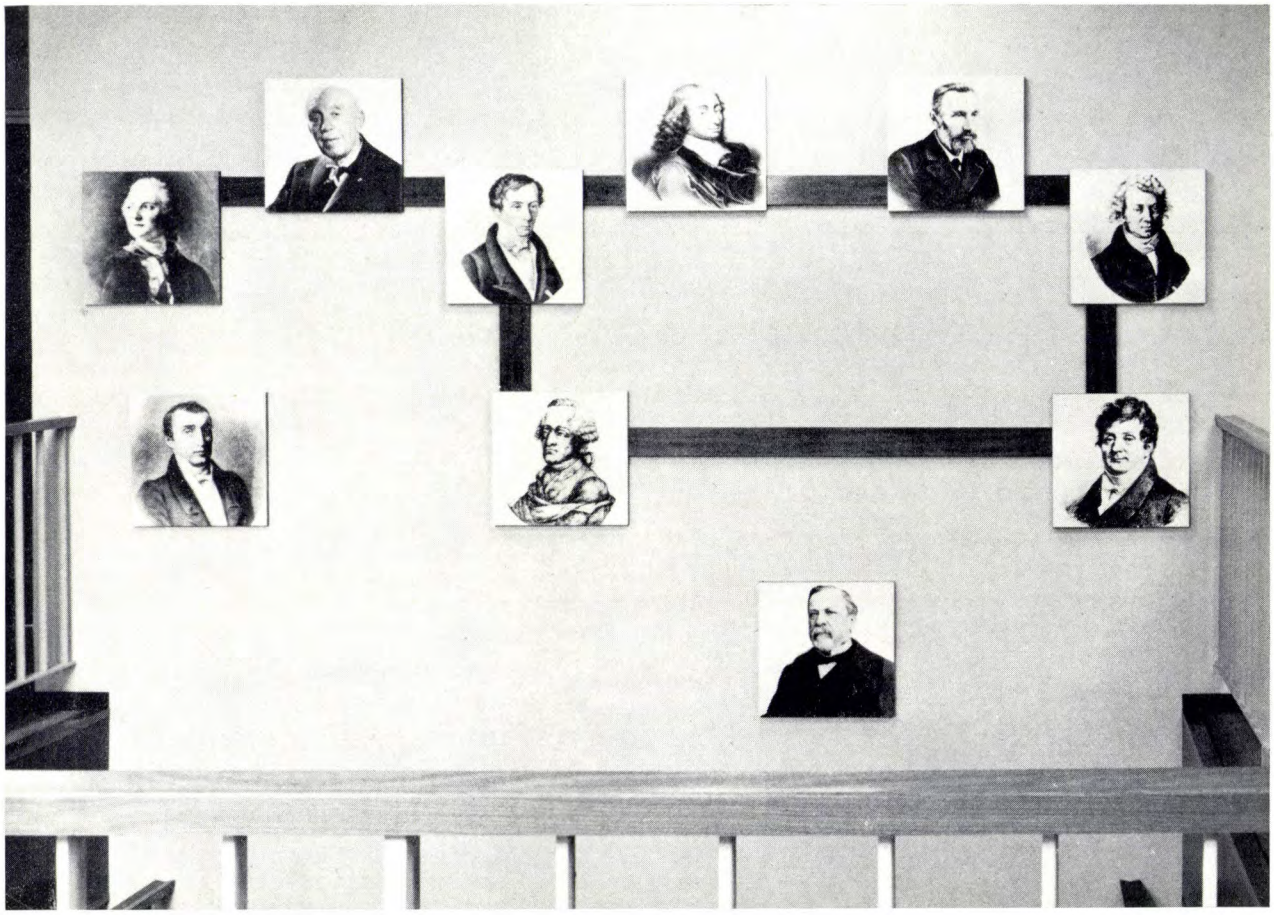
It is not easy to set down the full range of LEP activities, past and present, in a logical sequence, since the various paths branch, cross, or merge at many points — which in itself is a desirable situation. The schematic representation is therefore more like an interlacing pattern than a line.

On this account, the order in which we shall describe some of our work and the results obtained is rather arbitrary. Obviously, there is not the space here to give a full account or even a summary of all the work done by the Laboratories. What we have done is to pick out a few representative items in the hope that they will illustrate the interconnection between the subjects described in more detail in the articles which follow.

1) Work on the *photoelectric effect in vacuum tubes* and on *photomultipliers* in particular began not long after the foundation of LEP, and has continued to grow ever since. In the initial work, new techniques were examined and put into practical use, and at the same time studies were made of the physical principles involved in the design and operation of the tubes. The present work has two main objectives: the achievement of even better values for the sensitivity, gain, speed of response and geometrical dimensions for existing types of tube, and the design of completely new types of tube.

This research work is followed by development and production at associated establishments (R.T.C., Hyperélec) of many tubes which are of particular importance for applications in experimental and nuclear physics. These tubes have already assured LEP and the associated Companies of a strong position in the European and American photomultiplier markets. In the list of references at the end of the article there are several relating to the results of our work in this field: first-generation photomultipliers [2], design and characteristics of modern photomultipliers [3], fast-response photomultipliers for experimental physics [4], and the study of the photoemissive properties of alkaline layers [5].

The two recent developments in this field are the "transfer technique" and the introduction of new types of electron multiplier into photoelectric devices. Dealing first with the transfer technique, this is a method for activating photocathodes and assembling the tube under vacuum. The new types of electron multiplier which are used with the photoelectric devices are 1) a secondary-emission multiplier using bundles of extremely thin channels giving more or less "point-by-point" electron multiplication (our work on this



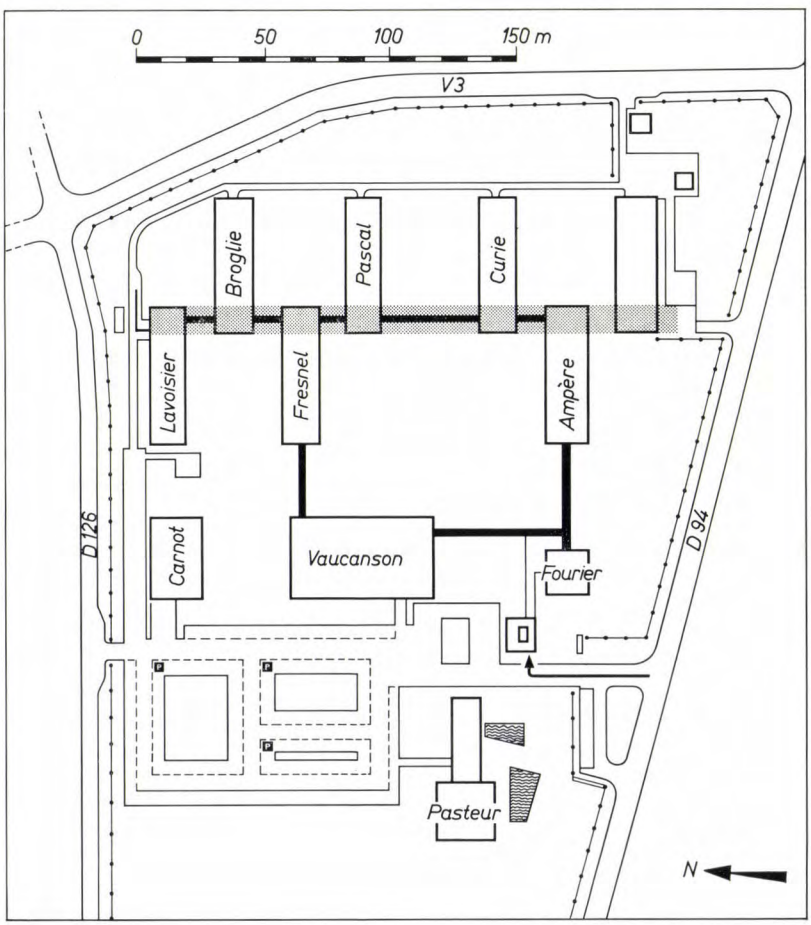


Fig. 2. *On the right:* Plan of the Laboratoires d'Electronique et de Physique Appliquée at Limeil-Brévannes. The buildings housing the main divisions of the Laboratories were given names of famous French scientists to reflect to some extent the nature of the activities pursued in each building (Lavoisier evidently symbolizes the chemical department, Vaucanson the engineering department and workshop, etc., while Carnot and Pasteur seemed not inappropriate names for the boiler-house and the building housing the medical, welfare and canteen services). Most of the buildings are connected by a covered passage-way. Underneath the passage-way between Lavoisier and Ampère there is a subterranean gallery (indicated by the shading) which provides an unobstructed space nearly 200 metres long for optical experiments.
Above: A decorative feature of the wall of the staircase in the Fourier building, which houses the management, accounting department, library and lecture hall, is a schematic plan of the whole premises with portraits of the scientists substituted for the names of the different buildings.

device is carried out in close cooperation with our colleagues at Mullard Research Laboratories); and 2) a semiconductor multiplier, which in fact is identical with the newly developed detectors for β-rays (fast electrons): the multiplication in this case is due to the fact that each electron liberates a number of charge carriers that is substantially proportional to its energy of arrival.

A short description of these notable developments will be found in two articles in this issue [6][7].

These recent technical developments enable a whole range of new tubes to be produced: image converters, electronic shutter tubes for high-speed photography, image intensifiers for viewing at very low light levels or for the observation of extremely rapid phenomena, photoelectric counters using solid-state multipliers which give good discrimination between pulses involving different numbers of primary electrons, something which was hardly possible with conventional secondary-emission tubes.

2) The programme of work on *microwave tubes* at LEP has been carried out in two phases with quite different subjects: velocity-modulated power tubes and travelling-wave cathode-ray tubes.

During the years 1957-1962 multi-cavity klystrons were developed that were suitable for high powers and large bandwidths. The special feature of these tubes was that the beam diameter was made appreciably smaller than the diameter of the drift tubes. This reduced the possibility of current losses at intermediate junctions and also of "multipactor" effect. The methods of construction required the use of glass-to-metal seals up to powers of 2 kW and ceramic-to-metal seals at higher powers [8].

After 1962 we turned our attention to very wide-band cathode-ray tubes (d.c. to several gigahertz). In these tubes the signal is coupled to the electron beam by means of a travelling-wave delay line. The main application of such tubes is for the real-time display of very short non-recurrent signals, a facility which is particularly useful in studying the physics of the elementary particles. In applications such as this the usual sampling techniques frequently cannot be used, since it requires recurrent signals. The present state of this work is described in another article in this issue [9]; the tubes have been put into production by R.T.C.-Hyperélec.

3) *Television* has been the subject of almost continuous — though highly diversified — research at LEP right from the beginning.

The earlier work was mainly on colour television and was not only concerned with the actual systems but also extended to the underlying physical and physiological principles of colour television. The peak of these activities was reached in 1957, with a number of contributions to the international Colloquium on the physical problems of colour television held in that year which were published in Acta Electronica [10]; the "double message" system developed earlier has been described elsewhere [11].

Other television problems were then tackled by LEP, particular attention being paid to transmission and camera systems, and equipment for outside-broadcast work. We can quote some of the results achieved in this field:

a) A two-standard camera (1960) which scans the same scene simultaneously in 819 lines and 625 lines [12][13].

b) Various cameras for outside broadcasts, in which the latest advances in circuit technology have been applied to achieve ease of handling, versatility, independence in operation and good coordination of several cameras from one central unit [14]. Some of the special features are:

A UHF link between cameraman and central unit, so that no connecting cable is required.

An automatic intersynchronization system for the cameras, controlled from the central unit.

A directional aerial with a device for automatic training, eliminating echoes and unwanted multipath effects.

All these circuits have, of course, been progressively transistorized and miniaturized as the various developments in circuit techniques became available.

The successive versions of this equipment, developed in conjunction with O.R.T.F. and put into production by T.R.T., have been used for events which mark the stages in the development of outside-broadcast work, from the inaugural voyage of the ocean-going liner "France" in February 1962 up to the 24-hours race at Le Mans in September 1968.

4) There is also an important and varied activity at LEP on the subject of microwave beam links for *telecommunication*. This work is carried out for and in conjunction with T.R.T.

Between 1950 and 1957 the work was concerned with "line-of-sight" links (*fig. 3*). This work was initially based on the multi-reflex klystron developed by F. Coeterier at Philips Research Laboratories in Eindhoven [15][16]. At that time this was the only tube capable of delivering a continuous-wave output of 10-15 W at 3400 MHz with a good efficiency (more than 20%). The later work made use of triodes for 4000 MHz, and important contributions were made to a series of microwave television links developed by T.R.T. These links are now widely used in France and abroad [17] [18].

In the meantime work began in 1956 on tropospheric-scatter radio links, and various studies and systems made between 1958 and 1964 have been described [19].

At present the research on microwave beam links is

not so much concerned with their "systems" aspects as with the introduction of new semiconductor devices to take over functions formerly performed by thermionic valves.

Another activity that is connected with telecommunication is the work on parametric amplifiers (see the last article under [19]) including amplifiers with transverse magnetic controlling field [20].

Fig. 3. Between 1950 and 1957 a considerable amount of work on "line-of-sight" microwave beam links was done at LEP in cooperation with TRT. The link between Bordeaux and the Pic du Midi (in the Pyrenees), established in 1957 using a multi-reflex klystron, at the time represented a world record: a high-definition television broadcast was transmitted directly over a distance of 213 km. The photograph shows the aerial system of the Bordeaux transmitter, with a parabolic antenna at ground level and a rectangular passive reflector at the top of the mast.

5) The first study of the *conversion of radiant energy* into electrical energy by means of semiconductor devices was made between 1955 and 1961, with a series of investigations into silicon photovoltaic cells and batteries for terrestrial applications. The decision to use "*N* on *P*" structures was made right at the start of the work at a time when the "*P* on *N*" structures were still generally advocated. The industrial development and production was carried out at R.T.C.

A survey of this work has been published in Acta Electronica [21]. After 1963 the work was continued with a special study of silicon solar batteries for space applications, in which the requirements for efficiency, weight, dimensions and ambient conditions are of course rather exacting. Another programme related to space research was a study of cadmium telluride cells, which was made to give support to production activities. As a result of all this work, R.T.C. now have a range of solar batteries that will meet various special requirements for both terrestrial and space applications and have been chosen as the supplier for equipment such as pick-ups, horizon sensors and solar batteries for several space vehicles (the satellites D1C, D2, TD, ESRO II, the Eole project, etc.). The net efficiencies achieved are between 10% and 11% with silicon, taking the losses due to encapsulation, filtering, etc., into account.

6) The *detection of radiant energy* by means of solid-state devices is another subject on which a great deal of varied work has been done at LEP. This work has mainly been concerned with the photovoltaic effect in semiconductor junctions — i.e. in fact the same effect as energy conversion. It is clear, however, that the detection of weak signals, in which the *information* contained is the essential feature, requires the parameters to be optimized in quite a different way from that used in energy conversion.

Two characteristics of solid-state radiation detectors that have been specially studied at LEP are their speed of response and their infra-red sensitivity. The investigations into speed of response have been and continue to be made with silicon and germanium photodiodes. The response times, which were a few nanoseconds until a few years ago, have now been reduced by an order of magnitude, and efforts are being made at the same time to increase the gain by making use of the avalanche effect in the semiconductor.

The work on sensitivity has not been confined to the near infra-red, to which silicon and germanium are sensitive, but has also extended to longer wavelengths (4 and even 10 microns). At these wavelengths it is possible to work with the thermal radiation from bodies at the ordinary ambient temperature. More recently, the work has extended to wavelengths up to 15 microns, which are particularly suitable for the horizon-scanning devices that can be used to stabilize the attitude of satellites. At these long wavelengths it is often necessary to resort to bolometric types of detector, since the energy gap which a junction detector must possess in order to respond to a 15 μm photon is so small that the detector can only be used at extremely low temperatures, which are not desirable in space-vehicle applications. On the other hand, at about 3 or 4 microns the indium antimonide or arsenide junction detectors made in our Laboratories give satisfactory results at the temperature of liquid nitrogen, which is acceptable for many terrestrial applications, in particular for infra-red viewing [22].

7) The subject of *infra-red viewing* is one that links with the preceding one at many points and has been worked on at LEP for a number of years.

There are two main approaches:

a) "active" viewing, in which the scene is illuminated by invisible infra-red radiation from a special source to produce an image which is made visible by methods like those used in television;

b) "passive" viewing, in which the normal thermal radiation from an object is used to produce an image.

Two types of camera tube based on the vidicon principle have been developed for active viewing. One type has a photoconducting target of lead oxysulphide on a silicon substrate, the other has a target consisting of a mosaic of germanium junctions [23].

Electronic-scanning cameras using these tubes have been built, first at LEP and later at T.R.T. These cameras can be used for active viewing, and also of course for passive viewing of objects that are sufficiently hot (e.g. at temperatures higher than 150-180 °C).

For passive viewing we have developed equipment with point detectors for wavelengths of 4 or 10 microns and an opto-mechanical scanning system. This equipment can be used to display a "thermal map" of an object at normal temperatures (*fig. 4*). This technique, known as *thermography*, provides a new valuable means of investigation for medical and industrial purposes [24].

A device intermediate between the two-dimensional target and the point detector for infra-red viewing has also been studied at the Laboratories. This is a linear array of point cells (a "barrette"), where only the slow scan (the field scan) has to be provided by mechanical systems and the fast scan (the line scan) is provided by means of electronic switching.

In connection with these investigations, various types of infra-red objective lense have been designed. These are described in an article in this issue [25] and in the last article under [26].

A number of fairly detailed articles about infra-red

techniques have appeared in eight special issues of Acta Electronica [26] [27].

8) The work at LEP on *data processing* has produced the following results:

a) Desk calculators with new logic circuits [28], first with optical display (1958-1962), later with electromechanical printer (1962-1966) [29].

b) Fast electromechanical printers, reaching speeds of the order of 3000 lines of text per minute [30].

c) Plotters based on the same principles.

d) Teaching aids which can perform immediate statistical analysis and record the replies given by pupils to multiple-choice questions [31]. These aids help the teacher to be continuously aware of how well his pupils have understood the lesson.

9) The general work at LEP on *high-speed electronics* has resulted in the development of special tubes (such as the cathode-ray tubes already mentioned under (2) on page 218) and circuits for real-time broadband oscilloscopes [9], as well as cameras for high-speed photography [32].

Rapid progress is being made in this field of high-speed electronics: three or four years ago the best time resolution attainable was still in the nanosecond region, but today time resolutions of the order of a few tens of picoseconds can be obtained.

10) The general field including *quantum electronics*, *electro-optics* and *non-linear optics*, has been represented at LEP by studies of lasers [33] and of the modulation of light for the transmission of information [34].

The principal subject of investigation in this field, however, is the application of electro-optical methods of light modulation in the development of an optical image relay tube for projection. A brief account of this tube is given in this issue [35].

11) It is quite evident that all these researches in electronics and physics would not get very far without a solid backing of *physical* and *chemical work* on the materials used.

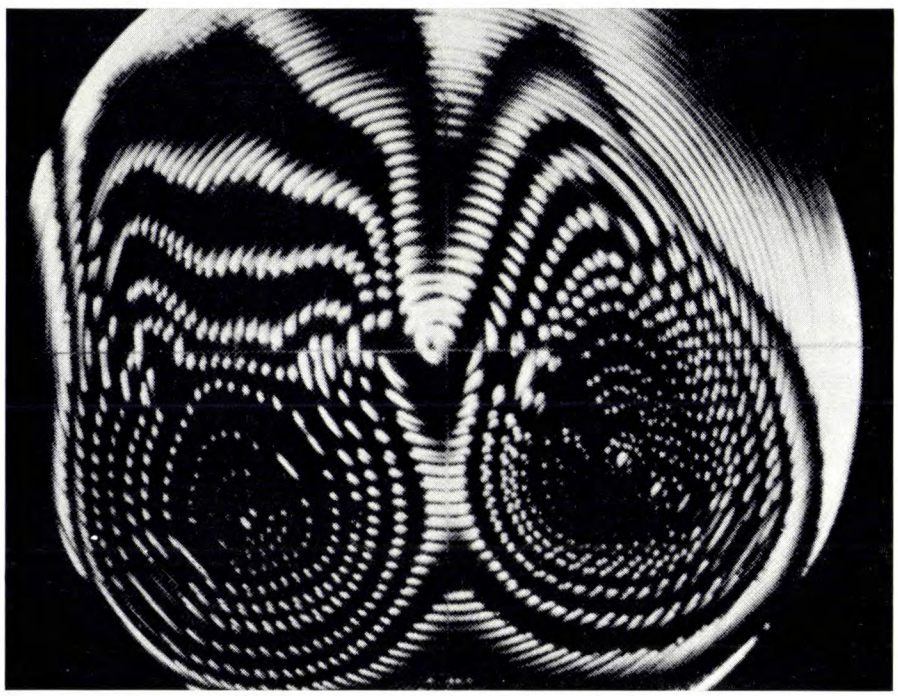


Fig. 4. "Thermal map" of female breasts produced by passive infra-red viewing in the 4 μm wavelength range. The temperature in the two peak regions is only 2 °C above the coldest parts. The image was produced by an opto-mechanical system involving spiral scanning. In recent equipment this is replaced by line scanning similar to normal television scanning [24].

Modern semiconductor and electro-optical techniques require the use of materials with far higher standards of purity and monocrystallinity than were generally required in earlier work. This means that the chemical division of a laboratory working in these new fields has not only to maintain its expertise in the more traditional methods, which are still required in providing a chemical service for the other departments, but must also carry out its own research into the techniques of purification, preparation, selection, characterization, etc., for the new materials. Our chemical division (*fig. 5*) is in fact playing this double role: it has made a noteworthy contribution to the realization of — and even to devising new ideas for — components using polycrystalline materials (camera tubes, for example [23] [36]),

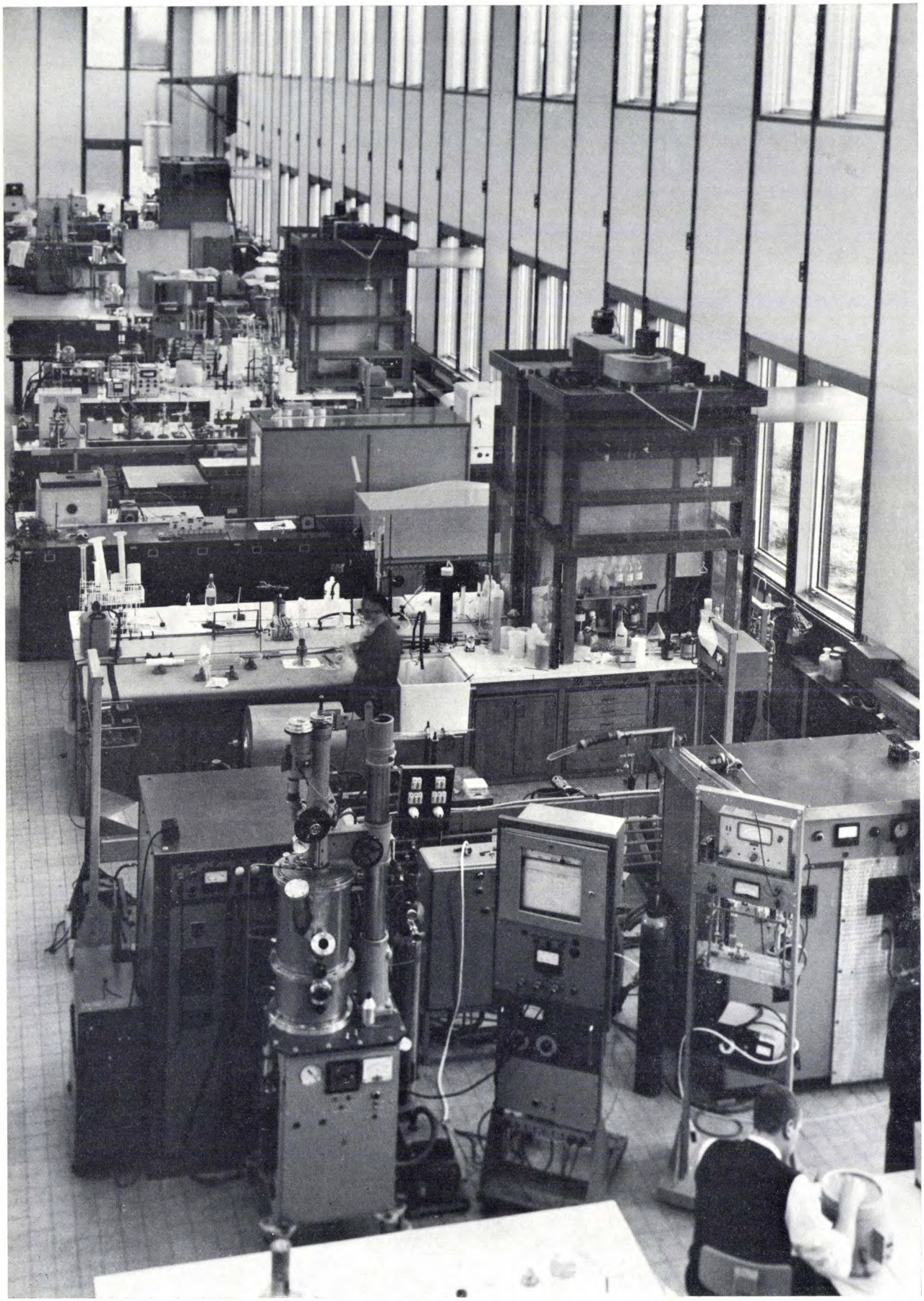


Fig. 5. Part of the Chemical Division in the Lavoisier building, with crystal-growing equipment etc.

and at the same time has devoted an ever increasing effort to single-crystal techniques, both for semiconductors and for optical and electro-optical materials. A number of older publications dealing with this work are listed in the Bibliography [33][37][38], and some recent results are described in articles in this issue [39][40].

12) Finally, we should mention several *technological studies* which qualify as research activities. These include various sealing techniques, high-precision machining, the mechanical treatment of fragile materials and ultra-high-vacuum technique. Three items in this issue relate to these activities [41][42][43].

Conclusions

As we said above, the material presented in this article should not be considered to give a complete or even a summarized account of the activities at LEP. The items selected are purely representative, and are intended to introduce the Laboratoires d'Electronique et de Physique Appliquée to the reader and to bring the work described in this issue into its proper context.

We hope too that what we have said will make it clear that the work of our Laboratories, like the work of any other research laboratory, is always located in the region where the already possible meets the not yet achieved; such work will always include an element of prediction or even inspired guesswork. At a time when techniques are evolving as rapidly as they are today, this situation for the research scientist is the origin of both his strength and his weakness, but if he is really worth his salt, it can only invigorate his enthusiasm and tenacity.

Bibliography

[1] Acta Electronica **9**, No. 2, 1965: Un ensemble moderne de laboratoires de recherches (conception et réalisation) (pp. 101-199).

[2] G. Piétri, Conception et réalisation de photomultiplicateurs, Acta Electronica **1**, No. 2, 35-46, 1956.

[3] G. Piétri and J. Nussli, Design and characteristics of present-day photomultipliers, Philips tech. Rev. **29**, 267-287, 1968 (No. 8/9).

[4] G. Piétri, Les photomultiplicateurs, instruments de physique expérimentale. Problèmes rencontrés dans leur conception, leur réalisation et leur mise en œuvre, Acta Electronica **5**, 7-30, 1961.

[5] G.-A. Boutry and H. Dormont, Some surface properties of pure alkali metals, page 225 of this issue.

[6] R. Legoux, The transfer technique, a new method for activating cathodes of phototubes; J. Nussli, Photomultipliers and the transfer technique, pages 234 and 236 of this issue.

[7] G. Eschard and R. Polaert, The production of electron-multiplier channel plates, page 252 of this issue.

[8] C. Zlotykamin, Réalisations de klystrons de puissance à large bande, Acta Electronica **4**, 47-57, 1960.

[9] G. Andrieux and C. Loty, A high-speed oscilloscope for real-time use, page 256 of this issue.

[10] Acta Electronica **2**, No. 1/2, 1957/58, *Coll. int. Problèmes physiques de la télévision en couleurs*: G.-A. Boutry and P. Billard, Visibilité des structures récurrentes dans les systèmes de télévision à analyse entrelacée (pp. 61-70). — P. Billard, Les seuils différentiels de luminance en télévision. Application à la recherche des caractéristiques de luminance optimum d'une chaîne de télévision (pp. 283-292). — J. Périlhou, Recherche de l'économie dans un système de télévision industrielle en couleurs (p. 363). — R. Genève, Présentation d'images de télévision en couleurs obtenues avec les systèmes N.T.S.C. et "Double Message" (pp. 371-377). — G.-A. Boutry, Colour television problems of to-day: an attempted survey (pp. 411-416).

[11] G.-A. Boutry and R. Genève, Etude du système de télévision en couleurs "double message", Onde électr. **37**, 337-357, 1957.

[12] P. Billard, L. Enselme, J.-M. Fournol, J. Maillard and M. Olifsson, Une caméra de télévision à double standard, Acta Electronica **4**, 289-314, 1960.

[13] R. Genève, Appareillages de prises de vues de télévision. Caméras à images multiples et équipements transistorisés, Acta Electronica **4**, 127-149, 1960.

[14] J. Cayzac, An outside-broadcast equipment with mobile television cameras, page 300 of this issue.

[15] F. Coeterier, The multireflection tube, a new oscillator for very short waves, Philips tech. Rev. **8**, 257-266, 1946.
F. Coeterier, The multi-reflex klystron as a transmitting valve in beam transmitters, Philips tech. Rev. **17**, 328-333, 1955/56.

[16] C. Ducot, Beam transmitters with double frequency modulation, Philips tech. Rev. **17**, 317-327, 1955/56.

Acta Electronica **1**, No. 1, 1956: C. Ducot, Les télécommunications par faisceaux hertziens (pp. 9-15). — G. Andrieux, Réflecteurs passifs pour faisceaux hertziens (pp. 35-46). — J. Cayzac, Procédé de contrôle automatique de fréquence en ondes centimétriques (pp. 47-53).

[17] G. Andrieux, Amplificateurs de puissance à triodes pour 4000 Mc/s, Onde électr. **37**, 777-780, 1957.

[18] Onde Electrique **39**, No. 384, 1959 (Communications au Congrès International "Circuits et antennes hyperfréquences", Paris 1957): R. Havot, Filtres de branchement pour faisceaux hertziens à 4000 MHz (pp. 180-185). — R. Dessert, Eléments de circuits hyperfréquence non réciproques utilisant des ferrites (pp. 215-221).

[19] Acta Electronica **9**, No. 3, 1965, *Systèmes de transmission à grande distance par ondes dirigées*: C. Ducot, Les télécommunications à grande distance par faisceaux dirigés. Quelques aspects de leur évolution (pp. 211-214). — C. Ducot, Problèmes de bruit liés à la transmission de signaux aux radiofréquences très élevées et aux fréquences optiques (pp. 215-226). — C. Ducot, Systèmes de modulation utilisables dans les liaisons hertziennes à grande distance (pp. 227-240). — R. Astor and J. Cayzac, Equipement pour liaison troposphérique utilisant la modulation d'amplitude à porteuse supprimée (pp. 241-254). — Y. Beurel and G. E. Zaaijer, Equipements UHF de réception à faible bruit pour transmissions à grande distance (pp. 255-268). — G. E. Zaaijer, Contribution à la théorie des amplificateurs paramétriques à résistance négative (pp. 269-291).

[20] G. Marie, Amplificateurs paramétriques basses fréquences à commande magnétique orthogonale, I and II, Acta Electronica **8**, 7-81 and 93-151, 1964.

[21] Acta Electronica **5**, No. 3, 1961, *Conversion directe de l'énergie solaire en énergie électrique*: F. Desvignes, La conversion directe de l'énergie solaire en énergie électrique. Fonctionnement et rendement des cellules solaires photovoltaïques (pp. 275-303). — C. Beauzée, Etude expérimentale des cellules solaires au silicium (pp. 305-329). — F. Desvignes and C. Beauzée, Aspects économiques de la fabrication et de l'utilisation des cellules solaires (pp. 379-391).

[22] F. Desvignes, J. Revuz and R. Zeida, Photoelectric solid-state devices and the perception of images in the infra-red, page 264 of this issue.

[23] M. Berth and J.-J. Brissot, Targets for infra-red television camera tubes, page 270 of this issue.

[24] M. Jatteau, Infra-red thermography equipment for medical applications, page 278 of this issue.

[25] C. Hily, Objective lenses for infra-red image formation, page 290 of this issue.

[26] Acta Electronica **5**, No. 4, 1961, *Cahier spécial "Infrarouge" No. 1*: G.-A. Boutry, Introduction (pp. 405-407). — H. Dormont and M. Auphan, Les phénomènes de bruit dans la détection du rayonnement infrarouge (pp. 427-458). — J.-J. Brissot and A. Dauguet, Les détecteurs actuels de rayonne-

ment infrarouge (pp. 459-516). — F. Desvignes, Analyse du fonctionnement des détecteurs de radiations optiques (pp. 517-592).
Acta Electronica 6, No. 1, 1962, *Cahier spécial "Infrarouge" No. 2*: P. Billard, Les matériaux utilisables dans les systèmes optiques travaillant dans l'infrarouge, I. Généralités (pp. 7-64).
Acta Electronica 6, No. 2, 1962, *Cahier spécial "Infrarouge" No. 3*: P. Billard and J. Cornillault, Les matériaux utilisables dans les systèmes optiques travaillant dans l'infrarouge, II. Caractéristiques (pp. 73-169).
Acta Electronica 7, No. 4, 1963, *Cahier spécial "Infrarouge" No. 4*: R. Genève, Télévision dans l'infrarouge (introduction) (pp. 291-293). — F. Desvignes, Problèmes et techniques de la Télévision en infrarouge (pp. 295-317). — J. Cornillault, Télévision infrarouge dans le domaine spectral situé au voisinage de 4 microns (pp. 319-360).
Acta Electronica 9, No. 1, 1965, *Cahier spécial "Infrarouge" No. 5*: J. Cayzac, Problèmes posés par la détection des faibles différences de température en télévision infrarouge (pp. 7-18). — C. Picot, Préamplificateurs pour appareils de prise de vues de télévision infrarouge (domaine spectral situé au voisinage de 4 microns) (pp. 19-34). — G. Marie, Préamplificateurs paramétriques adaptés à la détection de l'infrarouge (pp. 35-46). — M. C. Lefranc, Le refroidissement des détecteurs de rayonnement infrarouge (pp. 47-90).
Acta Electronica 11, No. 3, 1968, *Cahier spécial "Infrarouge" No. 6*: G.-A. Boutry, Avant-propos (p. 235). — P. Billard and C. Hily, Les systèmes optiques à champ étendu utilisables dans l'infrarouge (pp. 237-325).

[27] Acta Electronica 6, No. 4, 1962, *Les phénomènes d'accumulation dans les photodiodes (premier cahier)*: F. Desvignes, Introduction (pp. 287-290). — H. Dormont and M. Valton, Théorie des photodiodes fortement dopées à jonction abrupte (pp. 291-339). — M. Robert, Les phénomènes d'accumulation dans les photodiodes. Cas particulier du germanium (pp. 341-407).
Acta Electronica 10, No. 2, 1966, *Phénomènes d'accumulation dans les photodiodes (deuxième cahier)*: F. Desvignes, Introduction (pp. 121-122). — M. Berth and R. Petit, Tube analyseur d'images à mosaïque de germanium (pp. 123-135). — R. Petit, Problèmes d'optique électronique dans les tubes analyseurs d'images infrarouges (pp. 137-155). — M. Berth and C. Venger, Technologie des mosaïques de photodiodes (pp. 157-180). — P. Saget, Formation de mosaïques de jonctions de germanium *p-n* par épitaxie (pp. 181-194). — J. Revuz, Variation du facteur de contraste dans les cibles à mosaïque de jonctions (pp. 195-214).

[28] J. Borne and M. Audebert, Circuits de calcul utilisant des transistors et des tores magnétiques, Acta Electronica 5, 219-236, 1961.

[29] J. Borne, M. Audebert, M. Martin and R. Goudin, Une machine à calculer de bureau électronique, Acta Electronica 8, 203-239, 1964.

[30] J. Borne, High-speed printers for numerical data processing equipment, Philips tech. Rev. 29, 205-218, 1968 (No. 7).

[31] J. Borne, Dispositif électronique d'aide à l'enseignement, Onde électr. 47, 682-685, 1967.

[32] H. Bacchi and G. Eschard, Ultra-high-speed photography, page 241 of this issue.

[33] Acta Electronica 10, No. 1, 1966, *Les éléments pour lasers — technologie des monocristaux, spectres de fluorescence*: J.-J. Brissot, Préparation et propriétés des monocristaux destinés aux lasers (pp. 7-35). — A. Charles-Georges and A. Salmon, Taillage des monocristaux pour lasers (pp. 37-54). — Mme Lottin and J. Bonnefous, Les éléments lasers: contrôle des cristaux, finition optique et mesures (pp. 55-70). — J. Domain, Contribution à la réalisation d'un laser au néodyme (pp. 71-109).
Acta Electronica 10, No. 3, 1966, *Lasers utilisant des matériaux actifs solides*: R. Genève, Introduction (pp. 227-232). — G. Marie, P. Wurtz and R. Le Pape, Etude d'un laser au néodyme déclenché par un obturateur à effet Pockels (pp. 249-293). — R. Dessert, Lasers "solides" à fonctionnement continu (pp. 295-314). — A. Charles-Georges and A. Salmon, Usinage de cavités de pompage pour lasers "solides" (pp. 315-330).

[34] Acta Electronica 9, No. 4, 1965, *Utilisation du rayonnement lumineux comme véhicule d'information*: P. Billard, J. Donjon and G. Marie, Application de la modulation de lumière aux télécommunications (pp. 305-313). — J. Donjon and G. Marie, Modulateurs de lumière à large bande utilisant l'effet Pockels (pp. 315-385).

[35] G. Marie, Large-screen projection of television pictures with an optical-relay tube based on the Pockels effect, page 292 of this issue.

[36] G.-A. Boutry, J.-J. Brissot, R. Legoux, J. Périlhou and G. Piétri, Un tube convertisseur d'image pour l'infrarouge moyen "le Serval", Philips Res. Repts. 20, 684-706, 1965.

[37] J.-J. Brissot, La technologie des semi-conducteurs monocristallins, Acta Electronica 1, 187-200, 1956.

[38] J.-J. Brissot and H. Raynaud, Growing of single-crystal germanium in strips, Electrochem. Technol. 1, 304-307, 1963.

[39] J.-J. Brissot and A. Lemogne, Method for growing single crystals of cuprous chloride, page 261 of this issue.

[40] Single crystals for semiconductor devices and lasers, page 263 of this issue.

[41] R. Legoux, Transparent windows for a wide spectral range, page 299 of this issue.

[42] Bowl cutter machine, page 306 of this issue.
A method for small-diameter core drilling, page 307 of this issue.

[43] R. Evrard, Absolute micromanometers with diamagnetic levitation, page 231 of this issue.

Summary. Introductory article for this issue which is devoted entirely to LEP. After a brief account of the historical development of the Laboratories, which were set up about 20 years ago, a general account is given of the function of an industrial research laboratory. The article then looks more closely at the nature and function of LEP, and also at the organizational pattern within which the Laboratories operate. The field in which the Laboratories make most of their contributions lies in the border regions where electronics and optics meet, with particular attention being paid to photoelectricity (photomultipliers, photodetectors, solar batteries, etc.), television, microwave tubes and systems, infra-red viewing, data processing, fast electronics and electro-optics. A brief survey of representative items from recent work at LEP is included; an extensive bibliography gives a more complete picture of the work done by LEP during the last few years.

Some surface properties of pure alkali metals

G.-A. Boutry and H. Dormont

The development and improvement of photomultipliers has always been one of the more important activities at LEP. Certain characteristics such as the speed of response, the fluctuations in the amplitude of the output signal, the collection efficiency of the electron optics, etc. are determined by physical phenomena whose mathematical analysis is sufficiently advanced to make them amenable to calculation. However, this is not the case for the quantum efficiency and the spectral sensitivity of the photoemissive surfaces: the photoemissive effect, which is undoubtedly the longest-known quantum effect, is still not well understood nor completely mastered. The story of the discovery of the best photoemissive layers and their improvement is one of successive empirical experimentation. It is difficult to reproduce rather similar photoemissive layers by means of conventional vacuum technology. Although there is considerable literature on the subject, with many reports of measurements, the comparison and collation of these measurements yield confused results. In short, no coherent theory of the photoemissive effect appears as yet to exist. And if such a theory did exist, the existing results are not really adequate for experimental verification.

The advent of ultra-high-vacuum techniques has changed this situation. If we assume that the difficulties of reproducing surface properties are due solely to contamination by molecules of the residual atmosphere, then we might expect that a surface exposed to a residual pressure of 10^{-10} torr will become contaminated 10 000 times more slowly than a surface exposed to a residual pressure of 10^{-6} torr (as in ordinary thermionic valves). Experiment confirms this assumption. Thus when lithium is rapidly evaporated in an atmosphere with a residual pressure of 10^{-7} torr on to a substrate so arranged that the secondary emission of the resultant deposit can be measured immediately, the measured coefficient of secondary emission increases from about 0.8 immediately after evaporation to 1.5 in 15 minutes [1], whereas in the present article we shall see that below 10^{-10} torr the surface properties of alkali metals generally remain stable for at least 10 hours or so. It seemed, therefore, that the time had come to try and collect definite experimental results that could form an experimental basis for some future theory for the photoemissive effect.

It was obviously not suitable to start these new investigations by investigating photocathodes of the types usually used in modern vacuum devices; in fact these photocathodes consist of layers of complicated structure and composition, formed by successive evaporation of various elements (antimony, various alkali metals) sometimes followed by oxidation or systematic doping and always accompanied by annealing. It was necessary to start with the simplest possible photoemissive surfaces, those of the pure alkali metals. These metals, which crystallize to form a body-centred cubic structure, have such a loose electron-lattice coupling that the material may be said to resemble a gas of quasi-free electrons; as a result of the low bond strength photoemission takes place in the visible and near ultraviolet part of the spectrum, which simplifies measurements. There is also an important technical advantage in using the alkali metals since it is possible to prepare very pure layers of these metals by evaporation *in vacuo*. Some details of this process will now be discussed.

The vapour pressures of the alkali metals are very high compared with those of the elements in the adjacent columns of the periodic table. Nevertheless, at a given temperature, the vapour pressure decreases regularly from caesium to lithium so that if it takes an hour with a particular equipment to obtain a 3 μm thick deposit of caesium by evaporation at a temperature of 110 °C, it is necessary to raise potassium to a temperature of 160 °C, rubidium to 125 °C and finally sodium to over 230 °C to obtain layers of the same thickness in the same time. Now all the alkali metals attack glass, particularly the commonly used glasses which belong to the family of borosilicate glasses; and as with all chemical reactions, the rate of attack increases steeply with temperature. It turns out that the range in which some compromise temperature can be found is quite large for caesium, adequate for rubidium, small for potassium, but would disappear altogether for sodium unless special glasses with certain chemical properties are used for this metal.

Dr. H. Dormont is with the Laboratoires d'Electronique et de Physique Appliquée, Limeil-Brévannes (Val-de-Marne), France. Dr. G.-A. Boutry, formerly Director of these Laboratories, is a Professor at the Conservatoire National des Arts et Métiers in Paris.

[1] Private communication from M. Blanchet of CEA, Saclay.

Experiments carried out up to the present have consequently only made use of caesium, rubidium and potassium. The caesium was prepared at LEP either by reducing the dichromate with silicon, or by reducing the chloride with calcium *in vacuo* (Hackspill's method). The potassium was obtained in ampoules from two different suppliers, as also was the rubidium. In all cases the alkali metal was degassed and purified by carrying out several reflux distillations *in vacuo*. Only after this series of operations was it introduced into the experimental equipment, heated, evaporated and deposited on a plane polished glass surface, held at a temperature of 77 °K by liquid nitrogen. Throughout these operations the pressure in the enclosures never rose above 10^{-10} torr. Layers several microns thick were deposited at an apparent rate of 2 or 3 monomolecular layers ("monolayers") per second. The experimental process was thus the same for all three metals; the only differences were in the evaporation temperatures and in the aperture of the diaphragm limiting the beam of molecules.

As soon as the preparation of a layer was finished, its photoelectric spectral sensitivity at an angle of incidence of 45° was measured, first with natural light, and then with light polarized a) perpendicular and b) parallel to the plane of incidence. The results are summarized in *figs. 1, 2* and *3* (the differently marked points on these figures correspond to different cells). The spectral sensitivity curves in these figures have been normalized to bring all the maxima to unity.

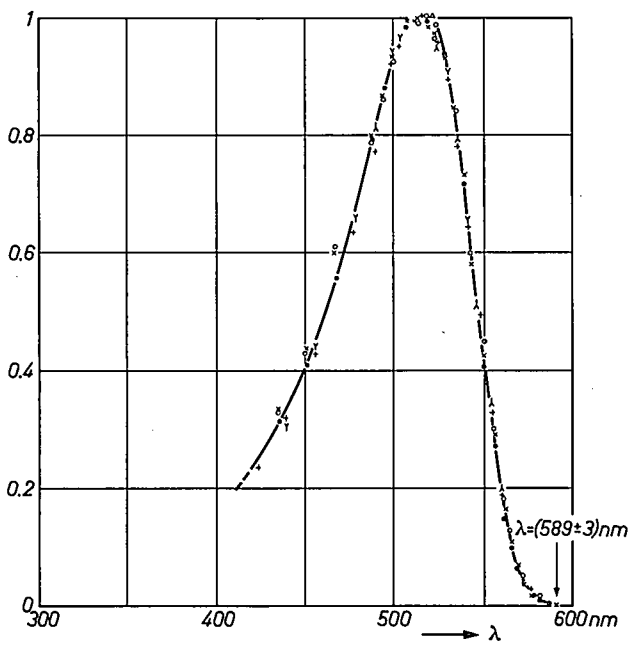


Fig. 1. Spectral sensitivity characteristic of solid caesium at 77 °K. Thickness of layer 3.0 ± 0.2 μm. The photoelectric current (in arbitrary units) per unit incident energy is plotted against wavelength λ.

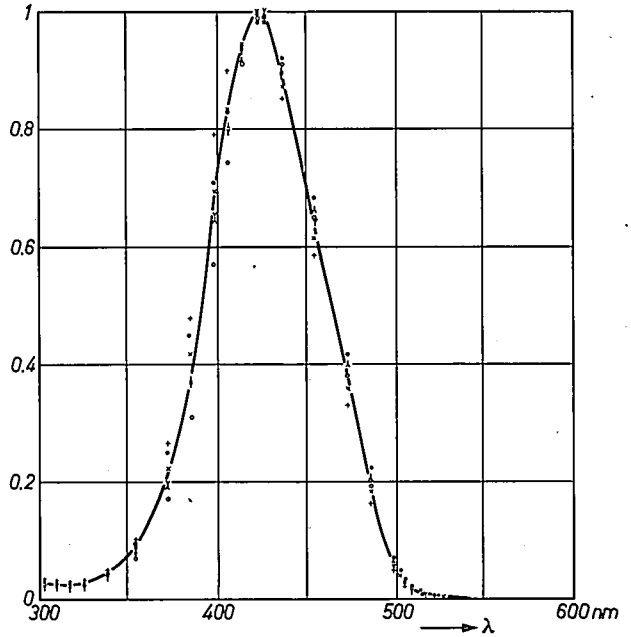


Fig. 2. Spectral sensitivity characteristic of solid potassium, as in fig. 1.

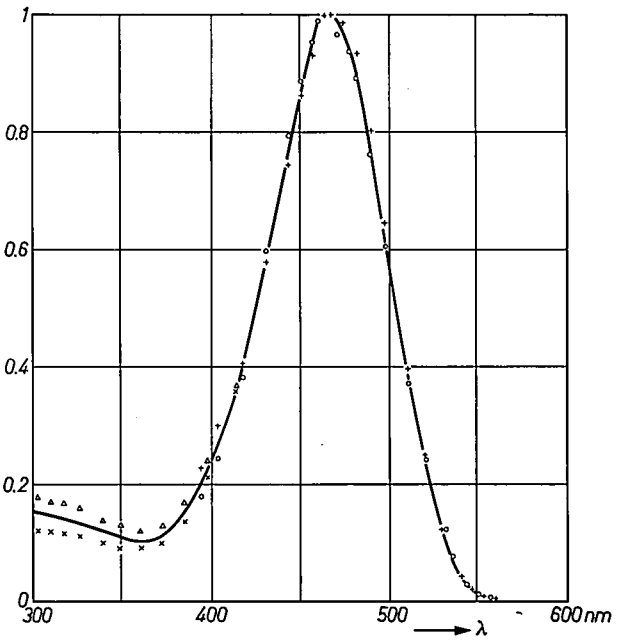


Fig. 3. Spectral sensitivity characteristic of solid rubidium, as in fig. 1.

With the conditions that we used it was always found that the reproducibility of the results was excellent from layer to layer. Extrapolation by Fowler's method of the results obtained allowed the true photoemissive thresholds and the corresponding work functions to be determined (*Table I*). The scatter of the measured results appears to indicate that the uncertainty in the value of the work function was never

Table I.

	threshold nm	work function eV
Cs	589.0	2.14
Rb	560.0	2.21
K	519.5	2.39

greater than 0.05 eV and had a probable value of the order of 0.02 eV. It will be observed that the work functions listed above are considerably higher than those found in the literature; since the last previous determinations for alkali metals were made in 1956, the difference is not particularly surprising. It does not seem unreasonable to assume that our determinations come very close to the real values for the pure metal.

We have already stated that the photoemissive properties of the layers were studied in polarized and unpolarized light. In the publications dealing with this subject it is generally accepted that with alkali metals there is a selective effect which makes itself apparent in two ways:

a) A spectral selectivity, which gives a photoemission maximum very close to the photoemissive threshold when the incident light is polarized so that the electric field vector is in the plane of incidence. This property is only observed for alkali metals (for all the others the spectral sensitivity increases slowly and nearly linearly with the frequency so that any maximum is in the far ultra-violet).

b) A vector selectivity, which makes the effect just described disappear completely when the plane of polarization of the incident light is rotated through 90°.

While our experiments confirm the existence of selective phenomena, they showed that the spectral selectivity and vector selectivity may not be two aspects of a single phenomenon but two distinct phenomena which do not have the same causes. In fact, our polarized-light experiments have always yielded curves showing a very definite spectral selectivity, whatever the orientation of the polarization plane, for all three metals, if the layers were formed at and maintained at a temperature of 77 °K. Although the height of the maximum varied from one layer to another or from one polarization to another by some 30%, the normalized curves could all be superimposed on one another. This is demonstrated by *fig. 4* for potassium. There was thus every indication of a spectral selectivity in these conditions while the vector selectivity was not very marked. The wavelengths at which the maxima were obtained are given in *Table II.*

In view of these results we next carried out experiments based on the following cycle, paying special attention to the particular case of potassium.

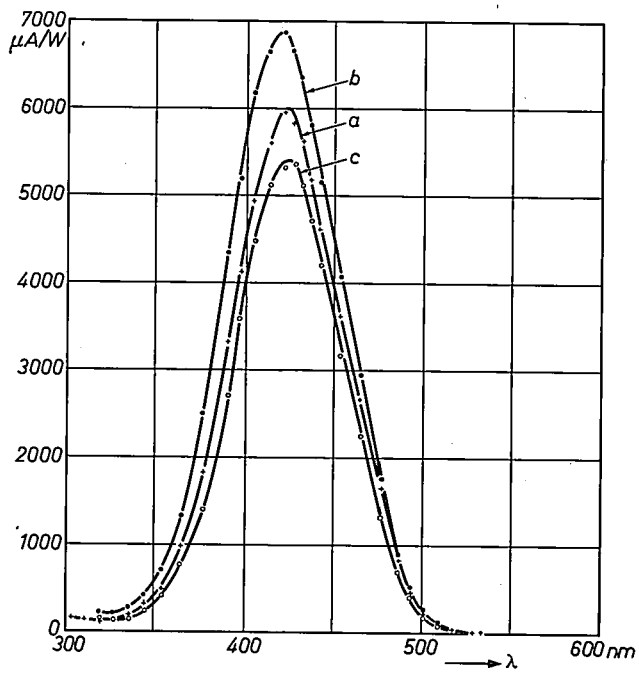


Fig. 4. Spectral sensitivity characteristics (photocurrent in μA/watt as a function of λ) of a layer of potassium at 77 °K in natural light (*a*) and in polarized light with electric vector parallel to (*b*) or perpendicular to the plane of incidence (*c*).

Table II. Wavelength at which the photoemission maximum is obtained for different pure alkali metal layers.

	Measured values nm	Values quoted in literature [*] nm
Cs	505.0	540.0
Rb	467.5	475.0
K	425.0	439.0
Na	—	352.0

[*] These values were measured before 1936.

1) The layer was deposited on a substrate maintained at 77 °K and the photoelectric sensitivity was determined: the results were like those shown in figs. 2 and 4.

2) The target on which the layer was formed was heated from 77 °K to 195 °K. Immediately after heating the photoelectric sensitivity was again measured for the two directions of the electric vector. The results obtained are reproduced in *fig. 5*. It will be seen that the shape of the spectral sensitivity curves, which are very reproducible, differs markedly from that obtained at 77 °K: as with the well-known vector selectivity there is no maximum except when the electric vector is in the plane of incidence. It will also be seen that although the thresholds are the same as those which we obtained at 77 °K, the position of the spectral-sensitivity maxima has changed.

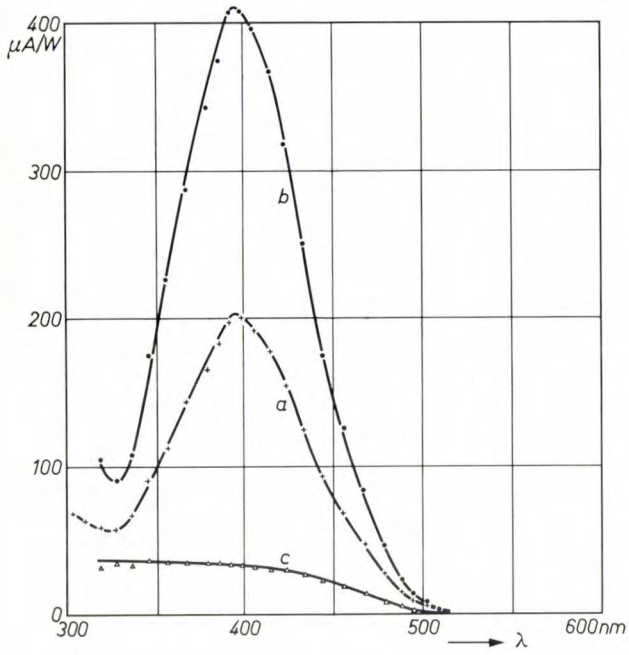


Fig. 5. Same as fig. 4 but at a temperature of 195 °K. (Note the difference in scale.)

3) The targets were then cooled to 77 °K again. This had no effect at all on the results which we have just described.
4) Still keeping the target at 77 °K we then evaporated another layer of potassium; measurement of the spectral sensitivity to polarized light gave a very faithful reproduction of the results shown in fig. 4.

This cycle of experiments was repeated several times and always yielded the same results.

We should add that, at any rate for potassium, the layers deposited at a temperature of 77 °K showed a somewhat diffuse reflection in addition to the specular reflection. The reflection became purely specular when the layers were heated. At the same time, besides the change of the spectral sensitivity characteristics, there was a sharp decrease in the mean photoelectric sensitivity to about 1/20 of its original value.

While it is not possible as yet to draw any formal conclusions from these experiments, they do seem to justify the belief that the existence of a quantum efficiency maximum very close to the photoemissive threshold is a characteristic property of the alkali element, while the vector selectivity as shown in fig. 5 seems to depend on the *structure* of the emitting layer. In view of the possible existence of a relation between the structure of layers and their photoemissive properties, we thought it would be interesting to investigate the properties of monocrystalline alkali layers. The remaining part of this article is devoted to the preparation of these layers. The actual photoemission measurements have only just been started.

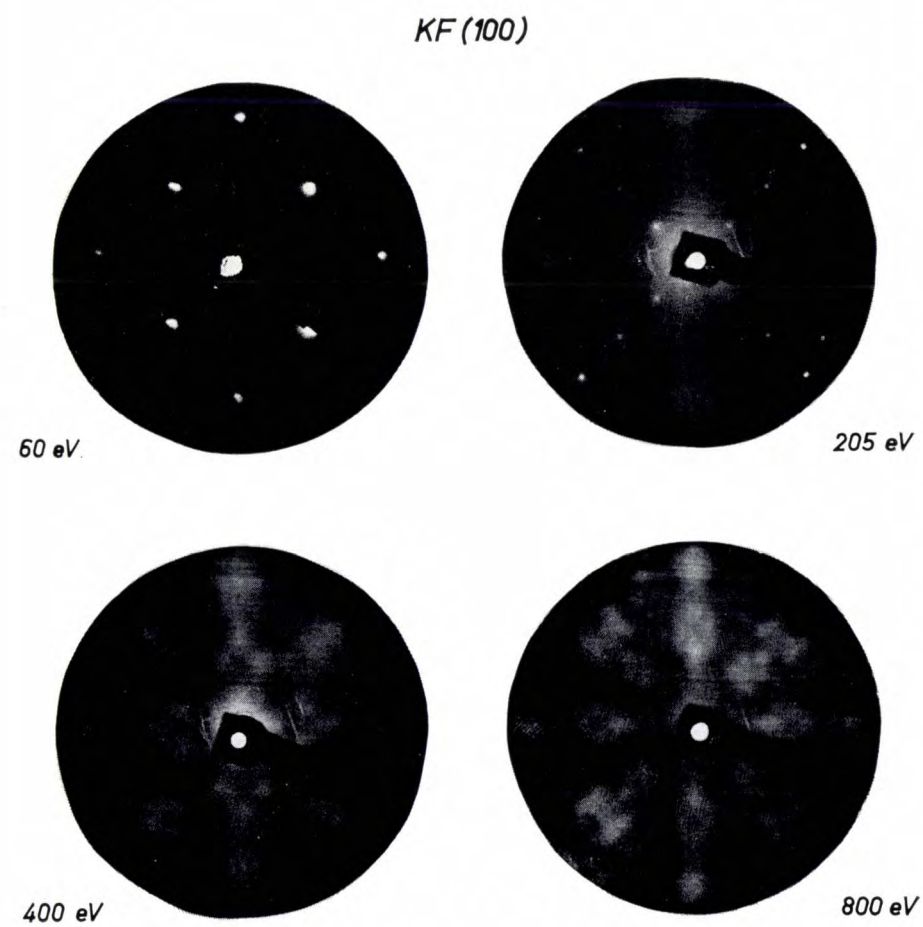


Fig. 6. Electron diffraction diagram obtained from monocrystalline KF showing the (100) face, at different electron energies. Kikuchi lines and bands appear at the higher energies.

The monocrystalline layers were prepared epitaxially, using single crystals of alkali halides as substrates for the photocathodes, and the crystallization process was monitored by the diffraction of slow electrons. The experimental process is different from the one used previously because the epitaxial technique requires the substrate temperature to be only just below the

melting point of the metal to be vaporized. The entire process takes place in an ultra-high-vacuum enclosure: this contains a slow-electron diffraction camera, a system for moving the crystal and varying its temperature from 77 °K (liquid nitrogen) to +100 °C (hot air), the vaporizing device, and furnaces for the degassing processes; these are indispensable if a pressure lower than 10^{-9} torr is to be maintained. The alkali halide crystals are cleaved in air before they are placed in the equipment, and are then subjected to extensive degassing in ultra-high vacuum. They are kept at about 0 °C during the epitaxial process, which is interrupted periodically to allow diffraction tests to be carried out.

Potassium, rubidium and caesium were used. The alkali halides NaCl, KCl and KF were chosen because of their crystal lattice and the ease with which they can be cleaved.

At the start of the process the well-known diffraction pattern for the substrate is obtained; this is shown for KF in *fig. 6*. At low voltages the pattern consists of dots and a weak uniform back-ground; above 100 eV a pattern of light bands with dark surroundings is obtained. If the voltage is then increased, this changes into a pattern known as Kikuchi lines and bands due to scattered electrons subjected to Bragg reflection at the crystal planes. The position of all these lines and bauds corresponds accurately with the properties of a facecentred cubic crystal lattice with lattice constant equal to that of KF and NaCl.

During the epitaxial process it is impossible to obtain any pattern until several molecular layers have been deposited, since the surface scanned by the electrons becomes charged. This is easily explained if it is assumed that the substrate is covered with a discontinuous film which does not have a good electrical connection to earth and does not assume a stable potential since the secondary-emission coefficient is high. As soon as the deposited metal layer is thick enough, the charges flow away and the patterns associated with the face-centred cubic structure of the epitaxial layer are obtained. These, however, can only be seen at a relatively high voltage; at low voltage the pattern is simply a continuous background with no dots at all. Inhomogeneities become visible if the accelerating voltage is increased, and at about 400 eV they give a symmetrical pattern whose clarity increases appreciably if the layer is cooled to about —100 °C. Between 300 and 1000 eV Kikuchi lines can be observed which correspond closely to the lattice planes of the alkali metal being deposited.

A distinction should be drawn between two cases depending on whether the lattices of the substrate and the alkali metal are very well matched or not. The first case is illustrated by potassium on KF; the lattices of the two crystals differ by less than 1 % and the corresponding Kikuchi patterns are shown in *fig. 7*. It can be seen from the lines and bands that the metal shows the (100) face. The second case is that of K and Rb on NaCl; we then obtain the patterns shown in *fig. 8*.

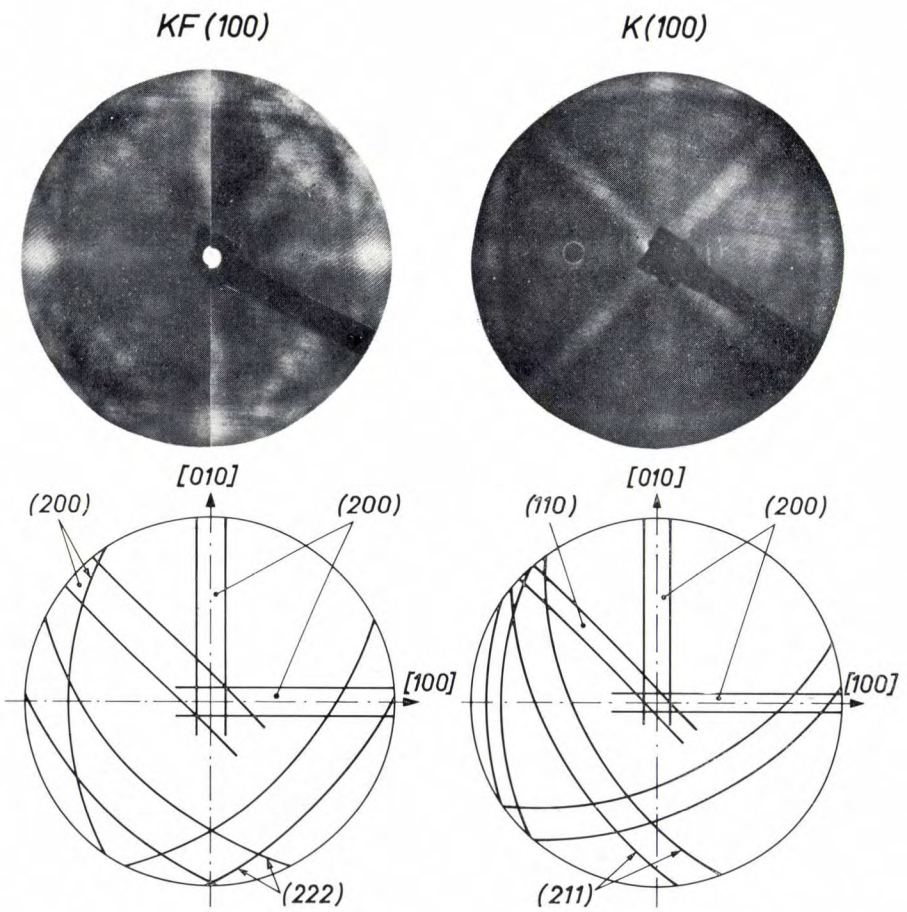


Fig. 7. Electron diffraction diagram at 1000 eV, obtained from an epitaxial layer of K on KF (*right*). Comparison with the diagram for pure KF (*left*) shows that the metal presents the (100) face. The Kikuchi diagrams are drawn underneath.

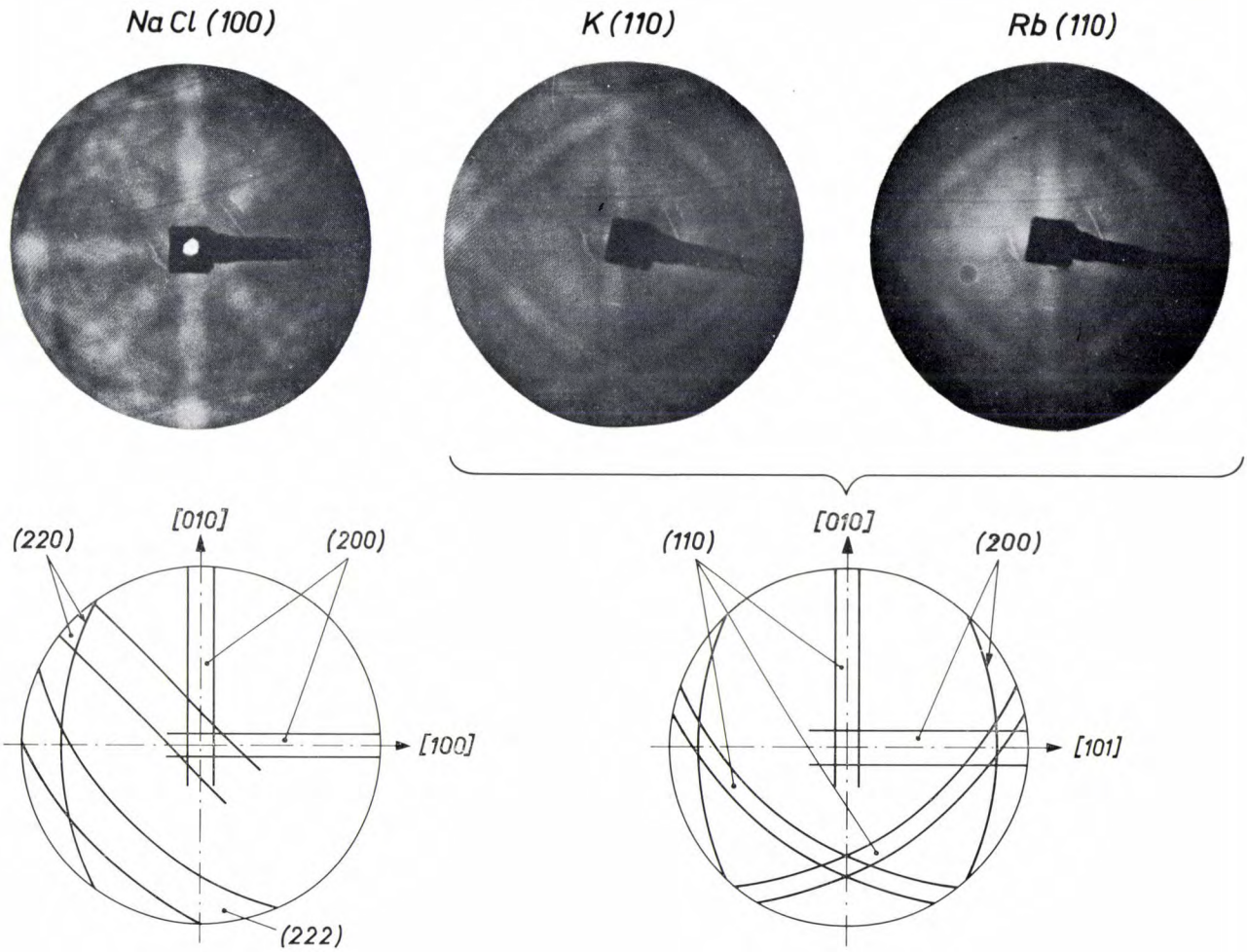


Fig. 8. As fig. 7, for epitaxial layers of K and Rb, grown on NaCl (*right*). Comparison with the diagram for pure NaCl (*left*) shows that in these cases the metal shows the (110) face.

In these patterns the rectangular symmetry is characteristic of the (110) face, which is the face with the highest atom density.

We thus have available a method which can be used to prepare monocrystalline layers of alkali metals, showing either the (100) or the (110) face; this is of particular value in investigations of the photoemissive effect of these metals. The related measurements have only just been started and no results are as yet available.

The work that has been described in this article forms only a part of a much larger programme of studies on surface physics and, to begin with, on the optical and electrical properties of metal surfaces. The volume of work required by these studies is considerable, and this indicated that the best way to proceed would be by cooperation between industrial and university laboratories. The work has in fact been organized in this way, with common leadership of the two groups to coordinate their work as closely as possible. We hope in the near future to be able to give an account of research on the reflection coefficient of the metal surfaces whose photoelectric properties have been discussed here.

Literature (articles dealing with the photoemissive properties of alkali-metal surfaces):

J.-C. Richard, C.N.A.M. Thesis, 1963.
G.-A. Boutry, R. Evrard and J.-C. Richard, C.R. Acad. Sci. Paris **258**, 143, 1964.
G.-A. Boutry, H. Dormont, R. Evrard and R. Perrin, C.R. Acad. Sci. Paris **261**, 383, 1965.
R. Perrin, C. R. Acad. Sci. Paris **267 B**, 58, 1968.
G.-A. Boutry, H. Dormont, J.-C. Richard and P. Saget, C.R. Acad. Sci. Paris **267 B**, 255, 1968.
J.-C. Richard and P. Saget, Le Vide **22**, 99, 1967.

Summary. The study of photoemission of thick films of potassium, rubidium and caesium in an ultra-high vacuum in conditions permitting measurement of uncontaminated surfaces leads to the conclusion that the emission is related to the physical structure of the material. Alkali metals were deposited on alkali halides by the epitaxial process in order to obtain monocrystalline metal layers.

Absolute micromanometers with diamagnetic levitation

R. Evrard

The instruments described here are designed to measure the very weak forces produced by the collision of molecules in a high vacuum. It is known that the pressure can be determined from these measurements, and that the method is both absolute and passive, i.e. it requires no previous calibration and the properties of the residual gas remain unaffected.

The classical systems used for such measurements generally contain a mobile-element assembly suspended from a very thin torsion ribbon, which is of course fragile and extremely sensitive to vibrations. The measurement consists in establishing an equilibrium between the torsional couple of the wire and a couple set up by the molecular collisions. The sensitivity can only be increased by using an even thinner suspension ribbon. The practical limit of this kind of device now seems to have been reached, and such instruments are not really suitable for measuring pressures below about 10^{-7} torr.

At LEP we have now developed instruments that are ten thousand times more sensitive. In these instruments the mobile element is kept in free suspension in the vacuum by magnetic levitation, i.e. there is no mechanical link of any kind.

Unless a complicated control system is used, a stable state of levitation can only be achieved with *diamagnetic* bodies placed in a magnetic field of suitable configuration and strength. In fact, graphite is the only material whose susceptibility is high enough to give levitation in practical fields. In our arrangement we use a very thin disc of polycrystalline graphite (thickness 50 μm) which is held freely suspended in a diverging magnetic field of the order of 1 tesla (10000 gauss). (Pyrolytic graphite, which can be kept in levitation with weaker fields under certain conditions, is not used for reasons explained below.)

The disc floats upright in the gap of an electromagnet whose field has circular symmetry about a vertical axis. The disc can rotate freely, without any mechanical damping, about the diameter coinciding with the axis of symmetry of the field. The disc thus constitutes an ideal mobile element, with a very low moment of inertia, which rotates freely with no mechanical support at all. In the instruments described below, the molecular bombardment causes a couple to act on the disc, and this couple is proportional to the pressure. The work performed by the couple, integrated over a large number of revolutions, is measured by determining the change in the angular velocity of the disc. The sensitivity can be very many times greater than with static measurements which rely on the use of mechanical suspension systems; with our instruments we have been able to demonstrate forces of the order of 10^{-15} newton.

We have applied this principle in the design of extremely sensitive instruments for passive and absolute measurements of pressure. The measuring vessel communicating with the evacuated space is situated in the air gap of an electromagnet. The freely suspended graphite disc is set rotating by means of a rotating magnetic shunt, which is then removed. With the first type of instrument, called a molecular gauge, we simply observe a deceleration of the rotation, which is caused by the molecular bombardment. If the nature of the residual gas is known, the value of the pressure p can be calculated as follows. Let H be the moment of inertia of the disc, M its mass and R its radius. Then if T is the decelerating couple acting on the disc, it can be shown that the angular velocity ω of the disc is given as a function of time t by:

$$\frac{d\omega}{dt} = \frac{T}{H} = -14.24 \frac{R^2\omega}{M\bar{v}} p,$$

where $\bar{v}$ is the mean velocity of the molecules of the gas. We shall go no further into the derivation than to note that the path described by each element of the disc is always perpendicular to the plane of the disc, and that the decelerating couple arises because a surface element moving towards the molecules undergoes more collisions than the element at the opposite surface — which is "running away" from the molecules. It is not very important whether the collisions are completely inelastic or not, and the purely kinetic description holds for all practical purposes. The sensitivity of this type of gauge is limited to 5×10^{-7} torr.

The second type of instrument *(fig. 1)*, which we call a "thermo-molecular" gauge, is much more sensitive. The disc rotates inside a cage of metal blades set at a

R. Evrard is with the Laboratoires d'Electronique et de Physique Appliquée, Limeil-Brévannes (Val-de-Marne), France.

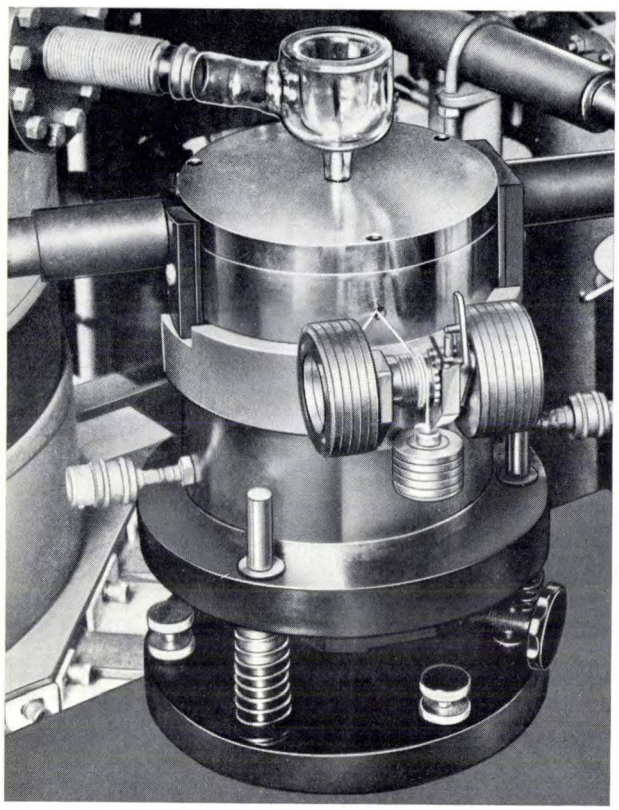


Fig. 1. Experimental version of the thermo-molecular gauge.

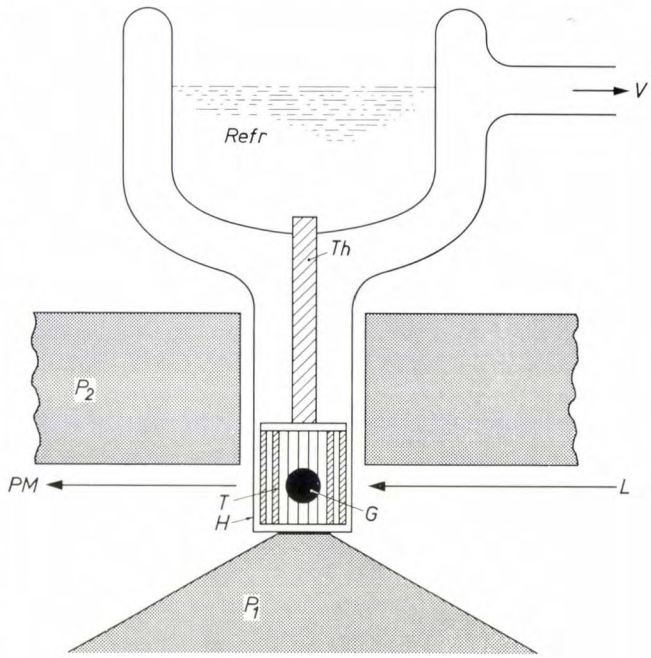


Fig. 2. Cross-section of the thermo-molecular gauge. *G* graphite disc, floating freely in the magnetic field between the pole pieces P_1, P_2. *V* evacuated vessel in which the pressure is to be measured. *T* "turbine" stator with blades. *Th* thermal conductor. *Refr* refrigerant. *L* light source. *PM* photomultiplier.

particular angle (*figs. 2* and *3*). This cage, which is insulated from the walls of the vessel, is cooled to a temperature Θ_2 by means of a heat-conducting rod connected to a cold source. Some of the "hot" molecules from the walls of the manometer, which are kept at the ambient temperature Θ_1, pass between the cooled blades without collision, and collide with the disc, but because of the screening effect of the blades they can strike only one half of each face of the disc (fig. 3). This produces a couple proportional to the total pressure. It can be shown that:

$$T = Cp\left(1 - \sqrt{\frac{\Theta_2}{\Theta_1}}\right),$$

where C is a geometrical factor that can be calculated.

The system works like a kind of molecular turbine. The disc forms an ideal rotor, which has no bearings and a very low moment of inertia; the cooled cage acts as the stator, and the wall is the heat source of the machine.

To determine the work performed by the couple, and hence the gas pressure, it is sufficient to measure the variations in the angular velocity of the disc over a length of time chosen to suit the sensitivity required. The range of pressures that can be measured may be considerably extended by lowering the temperature of the cold source.

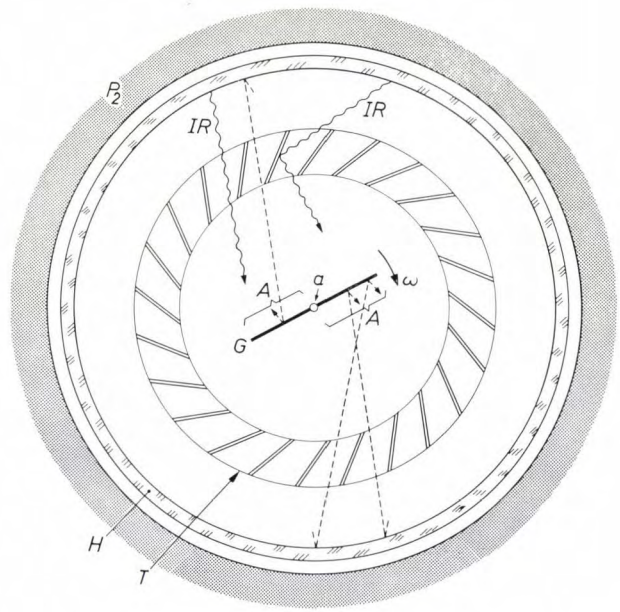


Fig. 3. Section through the actual measuring part. *G* freely floating graphite disc, rotating at an angular velocity ω, in the direction shown by the arrow, about a vertical axis *a*. P_2 pole piece. *H* hot wall of manometer vessel. *T* turbine stator with cooled blades. Since the blades are mounted at a particular angle, the molecules coming from *H* can strike the graphite disc only on the surfaces *A* turning to meet them. They would only be able to reach the opposite side by reflection from the cold blades, but then they have lost most of their energy. The result is a decelerating couple on the disc. The radiation pressure of the infra-red radiation *IR* emitted by *H*, on the other hand, does *not* give a net couple since it is uniformly distributed over the graphite disc by the reflection from the blades.

The measurements are carried out automatically: the rotating disc periodically interrupts a beam of light falling on a photoelectric cell, and the pulses of current are applied to a small computer. The computer gives a direct indication of the value of the deceleration, which is proportional to the pressure. The limit of sensitivity is about 10^{-10} torr, but this could certainly be further improved by at least an order of magnitude. As the last paragraph implies, the sensitivity should in theory increase with the duration of the measurement but in fact it is limited by various parasitic forces not related to the pressure. For instance, the heat source emits infra-red radiation, which acts on the disc in much the same way as the molecules and sets up a couple corresponding to the "radiation pressure". Although this parasitic couple may at first sight appear to be insignificant, if no precautions are taken it may completely mask the thermo-molecular couple at pressures below 10^{-8} torr. We have considerably reduced this radiation-pressure effect by carefully choosing the angle between the stator blades in such a way that the distribution of the infra-red radiation over the disc is made more uniform by the addition of the reflected radiation (see fig. 3).

The other parasitic decelerations, which are caused by imperfections in the symmetry of the field, can be neglected. They are much weaker than the decelerations which arise when pyrolytic graphite is used. This monocrystalline variant can be kept floating with relatively weak fields, but only when particular configurations are used (a ring-shaped rotor in toroidal air gaps). The moment of inertia of the rotor is then higher, and the parasitic couples are rather large. At the present time there are also other difficulties with this arrangement.

The great advantage of the methods of measurement described above is the fact that they are passive. They have virtually no effect on the vacuum to be measured, unlike the ion gauge, for example.

The micromanometers with diamagnetic levitation are thus ideally suited for ultra-high-vacuum studies where any interaction between the residual gas and the measuring instruments should be strictly avoided. Moreover, as "absolute" instruments they can be used for calibrating other types of pressure gauge.

The instruments are now being manufactured by R.T.C.-La Radiotechnique-Compelec.

Summary. The author shows how the use of diamagnetic levitation has made it possible to design extremely sensitive pressure-measuring instruments. A small graphite disc which is kept floating in a magnetic field of circular symmetry can rotate freely about its vertical diameter, which coincides with the axis of symmetry of the field. The disc is subject to a decelerating couple caused by collisions with the molecules of the surrounding gas. In one arrangement the random thermal movement of the molecules is used; in a second arrangement, which is much more sensitive, only molecular movements of suitable directions are effective owing to the use of a fixed stator with cooled blades surrounding the disc. Measurement of the variation of the angular velocity, i.e. the "work" done by the decelerating couples, enables the absolute value of the pressure to be determined with an accuracy of approximately 10^{-10} torr in the second case.

The transfer technique, a new method for activating cathodes of phototubes

R. Legoux

The procedures normally encountered in making vacuum tubes of most types are the actual assembly, pumping, and degassing. If however the tube contains a photocathode, then there is another process which is absolutely vital: the activation of the photocathode, by which an emitting layer of the desired photoelectric properties at the surface of the cathode is produced.

In the existing well-known procedure this is achieved by allowing the vapour of one or more alkali metals to react *in situ* with other materials in a carefully controlled temperature cycle. The materials most commonly used are antimony and silver in that order of importance, and they are deposited beforehand or during the process.

In this process the alkali metal vapour, as it diffuses through the entire volume of the tube, frequently modifies the surface properties of the electrodes within the tube by physical and chemical affinity, and this can adversely affect the characteristics and performance of the tube.

Another disadvantage of activation *in situ* is that sources producing the substances required for the operation have to be placed inside the tube, at places where they do not impede its proper function, and that they have to remain there throughout the life of the tube. Because of their size, these sources prevent the tubes from being made as small as desired, and the production of modern tubes of great complexity and miniature dimensions is made virtually impossible.

Finally, the basic characteristics of the resulting photocathodes (such as overall sensitivity, uniformity of the local sensitivity, and spectral sensitivity) and various other factors (production yield, repeatability) frequently depend on the internal design of the tube, since this design affects the diffusion of the molecules during the activation.

All the difficulties that arise in the established methods of processing can be avoided by using a technique in which a "transfer" is effected under high vacuum in a special enclosure [1] (*fig. 1*). This technique has the following features:

1) The window *F* on which the photoemissive layer is

R. Legoux is with the Laboratoires d'Electronique et de Physique Appliquée, Limeil-Brévannes (Val-de-Marne), France.

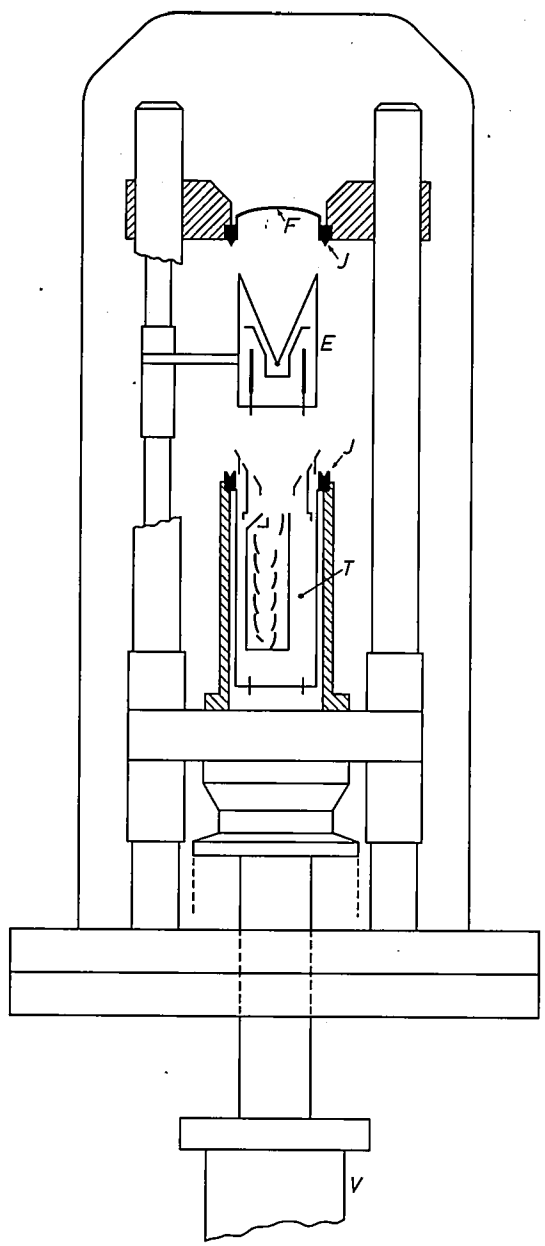


Fig. 1. Diagram of the equipment for activating photocathodes by the transfer technique. The activation enclosure *E* can be moved up and down and sideways, and fits against the window *F* to form an almost closed space. After activation, *E* is swung aside and the window and the main part *T* of the tube are sealed together by an indium pressure-weld (at *J*). The lower part of the tube is moved upwards and forced against the window by the screw-thread arrangement *V*.

to be made is kept separate from the rest of the photoelectric tube T and is the only component exposed to the evaporated materials which form the cathode.
2) The various evaporation sources are all grouped in a small auxiliary enclosure E which is brought right up against the window during activation and is moved away again afterwards.
3) After stabilization of the photoemissive layer the main part of the tube is raised and the window is sealed to it by an indium pressure-weld. This is done with the aid of the screw-thread arrangement V.

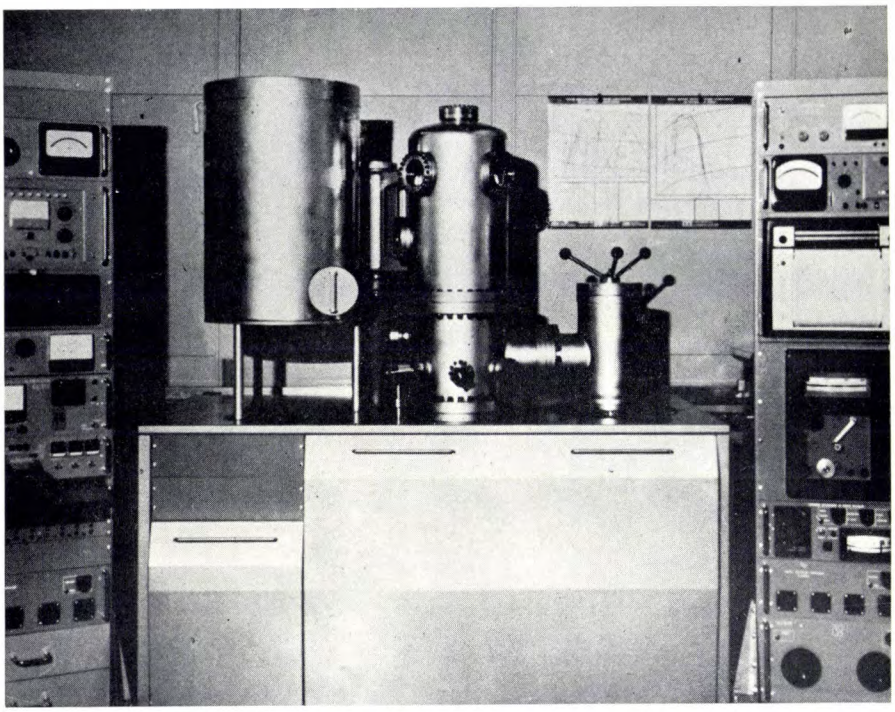


Fig. 2. Equipment for activating photocathodes by the transfer technique.

The technique has many advantages. Noteworthy among these is that the tubes can be pumped and degassed with a large opening to the vacuum system, with no danger of the subsequent increase in pressure that is liable to occur in tubes with a narrow pumping stem, and the fact that the activation is carried out so easily, away from the other parts of the tube which are no longer subject to any harmful effects of the process. There is also an improvement in the characteristics and performance of tubes in which some of the components can be contaminated by chemical or physical action of the diffused vapours. Components particularly susceptible to contamination are: the fluorescent screens of image tubes, high-insulation supports, and electrodes which are normally passive but can give field emission or small photoelectric effects when contaminated by a metal of low work function.

Photoelectric characteristics can be obtained by the transfer technique which are comparable with those obtained by the existing methods. Values better than 200 μA/lm have been obtained with tri-alkaline photocathodes on a glass or metal substrate. Moreover, a smaller spread in the spectral sensitivity curves of photocathodes can be achieved, and since there is plenty of space available in which the evaporation sources can be positioned to give a uniform deposition of material, it is possible to produce emitting layers with much better uniformity of emission.

The photoemission can be continuously monitored, and the final decision to include the photocathode in the tube waiting in the enclosure or to reject it can be deferred to the last moment; thus, if the photocathode should not be satisfactory there is no need to reject a complete tube (an expensive business, particularly if it is a prototype).

Finally, with the new technique we can now make tubes that could not be made at all previously, or could only be made with a limited performance. These are mainly tubes in which activation *in situ* is difficult because of lack of space.

Fig. 2 shows a transfer equipment which can be used in the manufacture of tubes with a useful photocathode diameter of 120 mm and an outside diameter of 180 mm. The limiting vacuum is between 10^{-9} and 5×10^{-9} torr.

Summary. The difficulties encountered in the established methods of processing tubescontaining photocathodes can be avoided by using a new method called the "transfer technique" for activating the cathodes. In the new technique the tube is treated inside an equipment specially designed for the purpose. The window is the only component exposed to the evaporated materials which activate the cathode. The various evaporator sources are grouped opposite to the substrate for the photoemissive layer and are kept there during the activation process. After stabilization the photoemissive layer is connected in its final position to the tube by a displacement inside the equipment.

[1] This method was the subject of a paper presented by P. Dolizy and R. Legoux at the 4th Symposium on Photoelectronic image devices, London, September 1968.

Photomultipliers and the transfer technique

J. Nussli

The improvement in the performance of photomultipliers has long been an important activity in the close cooperation between LEP and the Hyperélec works at Brive-la-Gaillarde.

In its application as a detector of light the photomultiplier is almost unsurpassed in speed of response and sensitivity, and on this account it is widely used for observing the scintillations produced in transparent materials by nuclear radiations. The photomultiplier thus acquired considerable importance in nuclear instrumentation, both for industrial and laboratory applications, and it is of great value in general photometric work. Since the requirements of the physicist are continuously changing, it is necessary to maintain a steady research effort, and in today's situation with new applications in fields such as space astronomy and lasers there is plenty to be done.

The history of these activities and the characteristics of this type of tube have been presented in an article in the previous volume of Philips Technical Review [1]. We shall confine ourselves here to the important contribution made by the transfer technique, a new method for activating photocathodes [2], in this search for ever better performance. The main virtue of the transfer technique is that it reduces the dark current, through an appreciable reduction in the field emission due to alkali metal deposited on all the electrodes of the photomultipliers.

This reduction in field emission also allows a more adequate study of new designs that make use of high voltages or high fields that could not easily be used previously and can give improvement in characteristics such as amplitude fluctuations and rate of response.

Conventional photomultipliers

In a conventional photomultiplier the signal from the photocathode is amplified by an avalanche process at a series of secondary-emission dynodes. To clarify the new developments it is perhaps advisable to recapitulate some facts already described in reference [1].

In most tubes the dark current is far greater than the thermionic emission of the cathode; the thermionic emission is always present, but as a rule forms only a small part of the observed current. For many types of tube the manufacturer is thus forced to adopt a "low-noise" solution. The thermionic emission of ordinary cathodes ($SbCs_3$ or Sb-K-Cs) is of the order of 10 electrons per cm^2 per second. The main source of the additional current is the field emisssion, produced not at the cathode, where the electric fields are weak, but anywhere in the tube, at the irregularities and particles of dust that can never be completely eliminated. The electrons thus produced are rarely accelerated directly towards the useful surface of a dynode, where they would be multiplied by secondary emission. As a rule they are accelerated towards other surfaces (the back of a dynode, for example, or connectors or insulators), where they produce light or soft X-rays. The light and X-rays thus emitted may then reach the cathode or one of the first dynodes, where they cause the emission of photoelectrons or bunches of photoelectrons.

The considerable intensity of field-emission currents emitted at surface irregularities (10^{-9} to 10^{-12} A) is due to the very low work function (typically 1.5 eV) of the metal surfaces which have adsorbed a thin film of alkali metal (especially caesium). This film, which is bound to the metal by an affinity even greater than the affinity that assures the stability of the photocathode, forms because there has to be a high vapour pressure of alkali metal in the tube while the cathode is being activated. In the transfer technique this operation is carried out in a subsidiary enclosure inside the bell-jar of the vacuum equipment, and the vapour pressure of the alkali metals always remains very low in the part of the tube separated from the window.

We should mention a few difficulties that do arise in applying the transfer technique. The film of alkali metal deposited on all the surfaces does not have an entirely negative effect since it also increases the secondary emission of the dynodes. Consequently, when the film is not present, the emission from the dynodes is slightly reduced and somewhat higher voltages have to be used for the same gain. This is not usually a serious handicap, and may even be an advantage if a rapid response is required. Furthermore, the alkali film acts as a getter in the tube, av ery useful role bearing in mind the effect of gases such as oxygen, hydrogen and carbon dioxide on the photocathode: if just a few atomic layers of these gases are adsorbed, the sensitivity is almost completely destroyed. The degassing of the components of the tube, particularly those

J. Nussli, L. ès Sc., formerly with the Laboratoires d'Electronique et de Physique Appliquée, Limeil-Brévannes (Val-de-Marne), France, is now with Hyperélec, Brive-la-Gaillarde (Corrèze), France.

subject to electron bombardment, therefore has to be carried out with particular care. However, the result on the dark current is truly remarkable (*fig. 1*).

A photomultiplier that gave serious problems in this respect in production is the XP 1020. In order to reduce fluctuations in transit time it proved necessary to use rather high electric fields between certain electrodes. Its dark current is on average 10^4 electrons per second. But a few experimental tubes made using the transfer technique have a dark current which on average is a hundred times smaller.

One final point should be made: masks can be placed in the activation enclosure during the evaporation, in order to obtain a sharp delineation of photocathode surfaces. When such masks are used either a cathode of small surface area, and hence giving little thermionic emission, can be deposited on the window, or several electrically separated cathodes (PM 450 FH, PM 450 FQ [3]). The emission of each cathode element can then be switched on or off at will by switching pulses. A photomultiplier of this type can be used as a detector of position.

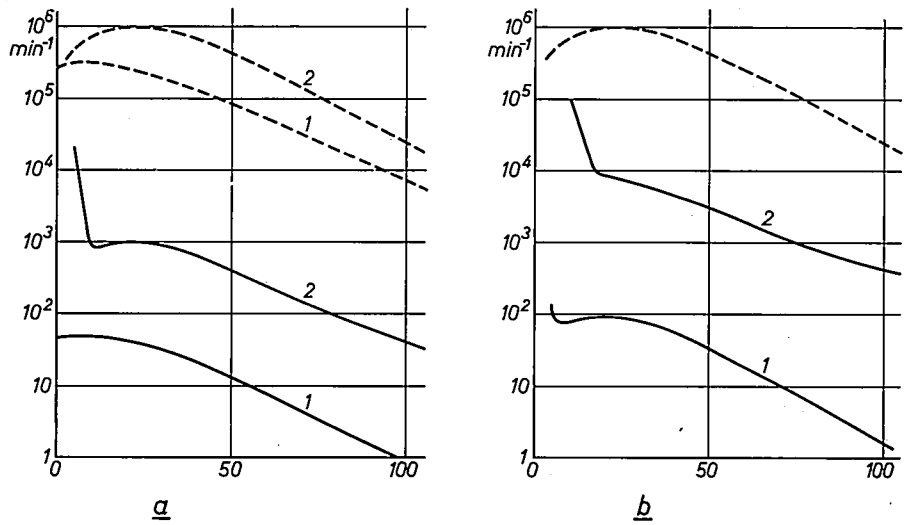


Fig. 1. Amplitude distribution of signal and noise pulses for various photomultipliers.
a) Noise signal (solid curves) and signal obtained for continuous illumination (dashed curves), from the tube 403 F made by the transfer technique (curves *1*) and the tube 56 AVP 03 made by the conventional pumping method (curves *2*).
b) Noise of XP 1020, made by the transfer technique (solid curve *1*) and by the conventional pumping method (solid curve *2*). The signal obtained for continuous illumination has the same amplitude distribution in both tubes (dashed curve).

The transfer technique is now being used at the Brive works in making short runs of ruggedized tubes derived from the PM 400 [1]. These tubes, represented typically by the PM 403 F [3], are suitable for use in space vehicles and satellites; their dark current lies between 20 and 50 electrons per second. A fact typical of the purely thermionic character of this current is that the amplitude distribution [1] of the dark-current pulses is identical with that of the pulses produced by photoelectrons (curves *1* and *5* of fig. 1).

The example of the XP 1020 illustrates clearly the advantage to be gained by using the transfer technique in making fast-response tubes. In fact the speed of response increases with the voltage, and this is limited in practice to a value at which it does not give field emission. The transfer technique thus makes it possible in principle to break through this barrier and use tubes at markedly higher voltages. We hope to be able to exploit this possibility in the near future.

Let us now turn to two rather newer types of photomultiplier for which the transfer technique has made an even more noteworthy contribution.

Hybrid photomultiplier

In a hybrid photomultiplier (*fig. 2*) the amplification of the photoelectric signal is no longer produced by secondary emission; the photoelectrons are accelerated to an energy E (10 to 40 keV) and then absorbed at a semiconductor junction, where each photoelectron will cause a large number of electron-hole pairs, given by:

$$\frac{E - E_w}{E_0},$$

[1] G. Piétri and J. Nussli, Design and characteristics of present-day photomultipliers, Philips tech. Rev. **29**, 267-287, 1968 (No. 8/9).
[2] R. Legoux, The transfer technique, a new method for activating cathodes of phototubes, page 234 of this issue.
[3] Developed under contract from Centre National d'Etudes Spatiales.

where E_w is the energy lost in the window of the detector, and E_0 is the mean ionization energy of the semiconductor material (in general $E_0 = 3.6$ eV for silicon). A reverse bias is applied to the junction and the depth of the space-charge region, where the charge carriers

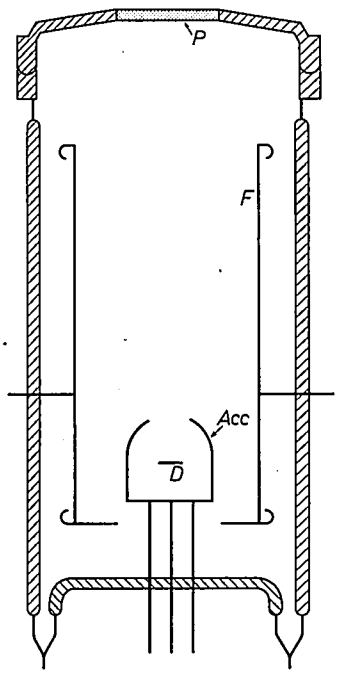


Fig. 2. Hybrid photomultiplier, schematic. *P* photocathode. *F* focusing electrode. *Acc* accelerating electrode. *D* semiconductor detector.

are liberated, is chosen to suit the depth of penetration of the photoelectrons, some tens of microns. The surface area of the junction, between 10 and 100 mm², has to be chosen as a function of the geometric spread of the electron beam (i.e. the focusing), of the permissible capacity, and of other parameters. The window, which is one of the electrodes of the junction, has a thickness which should be sufficiently small to absorb only a small part of the energy of the photoelectrons. These junctions are thus closely related to those currently used for the detection of charged particles. Without going further into the theory and practice of these detectors, we can just mention that their dark current and breakdown voltage are related mainly to leaks at the surface of the junction; these parameters are therefore extremely sensitive to any desorption or adsorption of impurity atoms.

Thus, two particular difficulties made themselves apparent during the early attempts to incorporate such elements in a phototube: The characteristics of the junction have to be maintained not only in the vacuum but also after baking out at 350 °C for several hours; on the other hand the alkali metals deposited on the detector surface must not increase its leakage current to an undesirable degree.

These semiconductor detectors are designed and produced at R.T.C. La Radiotechnique-Compelec (Caen laboratories). Various solutions for the first difficulty in making a hybrid photomultiplier were therefore investigated at these laboratories. The second difficulty, however, could be overcome by using the transfer technique. This is all the more important since the problems of field emission, which had already proved difficult with the voltages used in conventional photomultipliers (less than 3 kV), threatened to become insuperable at voltages between 10 and 40 kV.

The first hybrid photomultiplier was produced by R. Kalibjian [4], who was attempting to achieve a linear region for pulses up to and above 5 A. Work of this type is being carried out at Hyperélec (Brive) and also at R.T.C. Caen, where a detector based on the principle of "inverse epitaxy" [5] has been developed.

We shall now turn to another type of hybrid photomultiplier, developed at LEP, which has been devised for a different purpose, namely to give a reduced statistical fluctuation in the gain.

In a conventional photomultiplier [1] these fluctuations, which are proportional to $1/\sqrt{\delta}$, are large because the gain δ is small (about 3) and also because the number of secondary electrons collected varies from one dynode to the other. For scintillation detection (γ-spectroscopy) this increases the fluctuations due to the photomultiplier by 20-30% compared with the ideal case in which only the number of photoelectrons (n) fluctuates. If n is very small, any information about its value is blurred to the point at which it is no longer possible to distinguish pulses corresponding to 1 or 2 photoelectrons by their amplitudes. The improvement obtained with the hybrid photomultiplier can be illustrated as follows: The absolute fluctuation Δq in the collected charge is given approximately by $\Delta q \approx G \times q \times 1.5\sqrt{n}$, where G is the gain of the photomultiplier and q is the charge of the electron. On the other hand, in a hybrid photomultiplier with a supply voltage of 35 kV the fluctuation about the mean value of 10^4 electron-hole pairs produced, although not as small as the corresponding theoretical value of $1/\sqrt{10000} = 0.01$, is nevertheless considerably smaller than in a conventional photomultiplier. This fluctuation, which is usually quoted as a resolution expressed in keV, is in the hybrid photomultiplier related to the effect of various processes that take place in the semiconductor, to the fluctuation of its dark current, and to the noise of the amplifier. At the present time values of 10 keV are obtainable, which corresponds to an absolute fluctuation of the charge in the region of $0.3\ Gq$ for a hybrid photomultiplier with a supply of 35 kV ($G = 10^4$). Retrodiffusion (inelastic reflection of a small part of the photoelectrons)

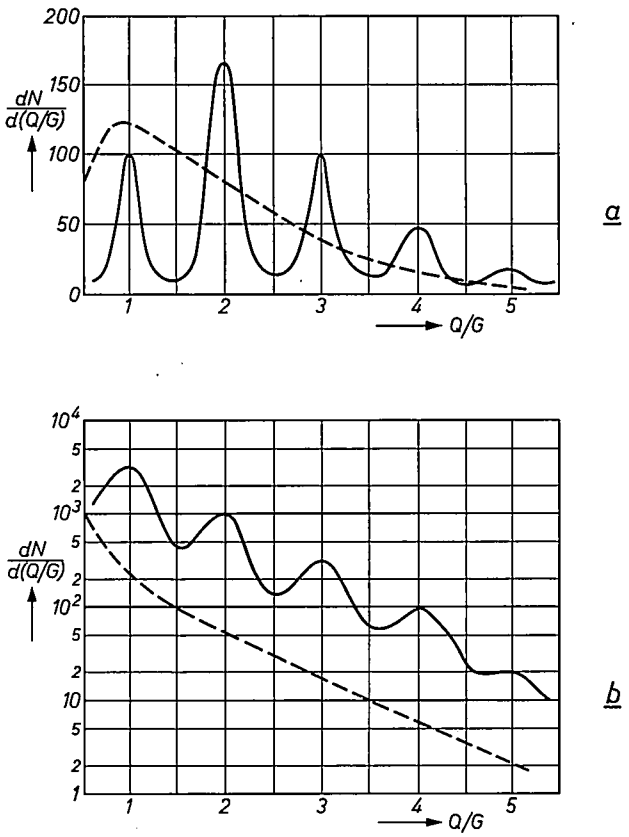


Fig. 3. Amplitude distribution of pulses obtained from a hybrid photomultiplier developed at LEP, *a*) with luminous pulses produced by a brief spark discharge, *b*) with the scintillations produced by the β-rays of tritium in a liquid scintillator.

The density of pulses counted is plotted against the number of electrons liberated at the photocathode (total charge Q — measured in the electronic charge q — divided by gain G). The dashed lines indicate the spectra that would be obtained with a conventional photomultiplier.

introduces an inaccuracy that becomes greater as n, the number of electrons, increases. *Fig. 3* shows various spectra obtained with a hybrid photomultiplier. The semiconductor used is a *P-N* junction (of planar configuration) with a guard ring; its edges have been passivated by SiO_2 to limit leakage currents. Its surface area is 12 mm².

It is true that a photomultiplier of this type requires a power supply that gives very high voltage and low-noise amplifiers of high gain, but such units are now readily available. On the other hand, since the gain now varies linearly with the voltage (and not as V^{10} or V^{12}), the regulation does not have to be so accurate. Finally, such a photomultiplier may well have a much better stability for variations in parameters such as the counting rate or the temperature.

[4] R. Kalibjian, A phototube using a semiconductor diode as the multiplier element, IEEE Trans. **NS-13**, No. 3, 54-62, 1966.
[5] J. Fertin, B. Lach, J. Meuleman, J. Dupuis, L'Hermite and R. Petit, Reverse epitaxial silicon diode for hybrid photomultiplier tube, IEEE Trans. **NS-15**, No. 3, 179-189, 1968.
[6] J. Adams and B.W. Manley, The channel electron multiplier, a new radiation detector, Philips tech. Rev. **28**, 156-161, 1967.
[7] G. Eschard and R. Polaert, The production of electron-multiplier channel plates, page 252 of this issue.

Microchannel photomultipliers

The general properties of channel-plate multipliers have been described earlier in this journal [6]. Such a plate, which has a surface area of several square centimetres and a thickness between 1 and 10 mm, can be traversed by 10^6 multiplier channels. An electron flux incident on one of the faces leads to the emission, at the other face, of a flux 10^4 to 10^5 times greater, with a voltage between the faces which is less than 1.5 kV. Methods for producing the channel plates are described in another article in this issue [7].

It seemed attractive to investigate the possibilities of using such multipliers in photomultipliers. One drawback immediately revealed itself: a large part of the photoelectrons are intercepted by the cross-section of the walls, so that the detection efficiency is greatly reduced. This means that such photomultipliers are of no use for nuclear spectroscopy. However, these multipliers could offer certain advantages because the thickness in the direction of propagation of the electrons is small (and hence the transit time is very short and consequently the fluctuations in transit time are small), while the surface area in the plane at right angles to the direction of propagation is large, allowing wide beams to be amplified. Moreover, the use of the transfer technique allows tubes to be made with the photocathode and multiplier set very close to each other, and also allows high electric fields to be employed between these plane electrodes with very little field emission. The obvious configuration to adopt therefore is the one shown in *fig. 4*. This shows a cross-section of a photomultiplier (see photograph *fig. 5*) which has a calculated transit

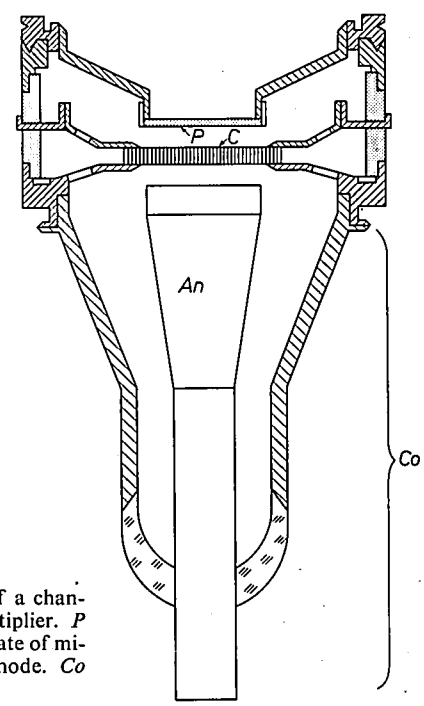


Fig. 4. Diagram of a channel-plate photomultiplier. *P* photocathode. *C* plate of microchannels. *An* anode. *Co* coaxial output.

time of 1 ns, a rise time of 0.2 ns, and a linear region extending beyond 1 A. Its gain is not very high, about 10^4, and this limits its use to relatively intense luminous pulses, like the ones observed during rapid phenomena in plasma. The gain is limited by the generation of positive ions, which travel towards the cathode, liberating electrons and thus causing after-pulses or even a self-maintaining discharge. This tube therefore lies somewhere between the photoemissive diodes of the biplanar type, which are very fast, and the conventional photomultipliers with high gain but slower response.

One of the characteristics of this type of photomultiplier, however, prevents it from being used in certain applications: its electrical resistance should never be much less than 10^8 Ω because it is difficult to conduct away the heat generated. The d.c. component of the signal cannot therefore exceed 10^{-5} A, which limits the counting rate to 10^7 electrons per second at a gain of 10^7 (and in practice, these values are 10 or 100 times smaller). The maximum charge transported by a short pulse (e.g. 1 μs) is limited to about 10^{-9} C/cm^2. This restricts the duration of 1-ampere pulses to about 1 ns.

Fig. 5. Photomultiplier whose cross-section is shown in fig. 4. Total length 80 mm, diameter 40 mm.

Conclusion

It is as yet difficult to predict how widely the transfer technique will be used in making photomultipliers in the future. Possibly its rather high cost will limit its use to high-performance tubes, but this is a technique that may undergo considerable development. The encouraging results obtained at the Brive works with ruggedized photomultipliers with bi-alkali or tri-alkali cathodes indicate that photomultipliers of even better performance will become available to the physicist.

Summary. When the transfer technique for activating photocathodes is applied to photomultipliers of conventional design, there is a considerable improvement in their noise characteristics and, in certain cases, in their sensitivity. This new technique also makes it possible to develop new types of tube such as the "hybrid" photomultiplier, in which the multiplier is in fact a semiconductor electron detector (the electrons being accelerated to an energy ranging from 10 to 40 keV), or the microchannel photomultiplier, which includes an arrangement of channel electron multipliers. The hybrid photomultiplier is less subject to statistical fluctuations in amplitude and the microchannel multiplier gives a faster response.

Ultra-high-speed photography

H. Bacchi and G. Eschard

This article describes several pieces of research devoted to high-speed photography, some of which have been undertaken at the request of the French Atomic Energy Commission and carried out in close collaboration with E. Laviron and C. Delmare, of the Limeil Research Centre, and M. Marilleau, of the Vaujours Research Centre.

High-speed full-image cameras can be grouped into three categories, namely: mechanical, electro-optical and electronic cameras.

Since 1954 mechanical cameras have provided exposure rates of 10^7 pictures per second, with the possibility of taking a very large number of pictures of the same phenomenon (20 to 100, depending on the version). With these cameras there is a lower limit to the exposure time of about 50 ns.

Electro-optical cameras are composed from an input lens, a shutter consisting of an electro-optic cell — usually a Kerr cell situated between crossed polarizers — and a repeater lens. With these cameras, exposure is obtained by applying a voltage of 15-20 kV across the cell. The exposure time can be reduced to a value of about 1 ns. The chief drawback is that the shutter transmits only 10-20 % of the incident light when it is open.

Electronic cameras, which first appeared on the market in 1962, give performances which are superior to those of the cameras already mentioned. They are also suitable for a much wider range of applications. In this article we shall deal only with this type of camera. The heart of the camera in this case is an *image-converter tube* which acts as a shutter and is used in conjunction with an input objective and a repeater objective.

The image converters which are used in these cameras all consist essentially of a photocathode and a luminescent screen. An image of the scene is produced on the photocathode by the objective lens; the luminescent screen receives the emitted photoelectrons and reproduces the image for a time which is relatively long (0.1 to 10 milliseconds) compared to the time the tube is "open" (e.g. 1 nanosecond).

There are three main types of converter tube:

1) Diode-type converters, consisting of a photocathode and a screen placed in two close parallel planes; we shall describe one version in greater detail below.
2) Triode-type converters, in which operation of the tube is controlled by an electrode in the form of a grid or ring [1].
3) Converters with deflection plates, in which the shutter function is achieved by shifting the electron beam past a diaphragm of very small dimensions [2].

Other types of converter exist which offer possibilities for information storage [3] or charge storage, but they are not yet widely used.

In addition to its function as a shutter an image-converter tube gives a gain in photons and can if necessary be coupled to an image intensifier. This makes it possible for an electronic camera to photograph with very short exposure times phenomena which are not bright enough to be photographed by an electro-optical camera.

As its name indicates, the image converter can also effect conversion of the light frequencies, and it can be sensitive to radiation of different wavelengths, depending on the photocathode used. For example, when the type S 1 photocathode which is sensitive up to 1.1 μm is employed, frequency conversion takes place since the screen which reproduces the image delivers radiation at a wavelength situated in the visible spectrum, centred on 0.55 μm. This conversion makes it possible to use films of very high sensitivity which are practically non-existent for the higher wavelengths.

The lower limit of exposure time provided by these tubes is governed either by the transit times of the photoelectrons between the photocathode and the screen or by propagation phenomena at the surface of the photocathode. These phenomena, which are peculiar to the tube, lead in practice to a lower exposure-time limit of the order of 100 picoseconds.

A converter-tube shutter is opened by applying a high-voltage pulse, usually between 15 and 20 kV, to the electrodes. The width of this pulse approximately

H. Bacchi, Ingénieur A. M. and E. S. E., and G. Eschard, Ingénieur E. S. E., Dr. Ing., are with Laboratoires d'Electronique et de Physique Appliquée, Limeil-Brévannes (Val-de-Marne), France.

[1] See for example J. A. Jenkins and R. A. Chippendale, High-speed photography by means of the image converter, Philips tech. Rev. **14**, 213-225, 1952/53.

[2] A. E. Huston, Proc. 3rd Symp. on Photoelectronic image devices, London 1965, pp. 963-966.

[3] R. W. Smith, in: High-Speed Photography, Proc. 8th Int. Congress, Stockholm 1968, pp. 18-20.

determines the exposure time. If it is desired to obtain exposure times approaching the limiting value allowed by the tube, the control circuits have to deliver pulses whose rise and fall times are of the order of one hundred picoseconds.

The essential characteristics of the tubes and the control circuits are a direct consequence of these requirements. In fact, a voltage variation occurring in such a short time can only be effectively applied to a tube if it is transmitted by a transmission line whose highest operating frequency is of the order of 3 GHz. The tube, of course, has to form an integral part of this transmission line, a circumstance which determines the geometry of the structure on which the sensitive layers have to be formed.

Allowing for the relatively short distance between the photocathode and the screen (cf. formula 1) and for the need to have a useful surface area of several square centimetres, we are obliged to use a transmission line with a characteristic impedance of approximately 25 Ω, and this means that the current will reach a value of about 600 A when a pulse of 15 kV is applied. Therefore, the trigger circuits have to include a high-speed spark gap, and our best results were obtained with a gap of coaxial structure [4][5].

We shall see later that the rise time of the leading edge of the wave produced by a circuit comprising a spark gap is proportional to the distance between the electrodes of the gap [6]. A lower limit is imposed on this distance by two factors: a) the useful life, which will be shortened when the gap is reduced, b) the pressure of the gas in the spark-gap cavity. In fact, the operating voltage of the spark gap is determined by the voltage required for the shutter tube, and, if for a given voltage we decide to decrease the distance between the electrodes, we are obliged to increase the pressure (Paschen's law). Technological problems limit the pressure to twenty or thirty bars, and consequently the distance between the electrodes is made a few tenths of a millimetre.

The final parameters which determine the minimum length of the rise time are therefore the distance between the electrodes of the spark gap and the quality of the various matching elements along the transmission line linking the spark gap to the shutter tube.

Additional requirements concerning the drive circuit come into play if it is desired to obtain a perfectly uniform transmission factor during the exposure and for an accurately defined time. If before and after the actual exposure time the camera has to present the maximum opacity permitted by the converter tube, the waveform of the applied pulse must remain fairly accurately rectangular despite the extremely short duration.

The necessary circuits can now be realized thanks to the progress made at LEP in the field of spark gaps with ultra-high-speed triggering, inserted in coaxial lines. The performance of these spark gaps will be improved still further as a result of research at present in progress.

We now propose to give a brief description of the converter tubes developed at LEP and of several types of camera in which these tubes are used. A description of the tubes has been given by Eschard and Polaert [7] and various equipments have already been described in references [4] and [5]. These equipments were also the subject of a paper read before the Eighth International Congress on High-Speed Photography in 1968 [8]. Several examples of spark gaps with two or three electrodes, developed at the Laboratoire Central de l'Armement and LEP have also been described in these articles.

Image-converter tubes

For a number of years now we have been developing a whole family of biplanar diode shutters [9]. These shutters are very similar in their underlying principle and are based on the technique of activation of photocathodes by the transfer method [10]. On page 235 in this issue a laboratory equipment is shown with which it is possible to make tubes whose useful diameter may be as much as 120 mm. This is important since one of the interesting points about this type of shutter is that it enables the definition of the image to be increased by using increasing diameters.

The main characteristics of these shutters were discussed in a recent paper [11]. We shall first briefly recapitulate the main features to bring out the points which have been the subject of new developments, particularly the possibility of using cathodes which are sensitive up to the near infra-red and constructions suitable for very short exposure times.

Quality of the image

If we place the screen a distance d from the photocathode, the defocusing due to initial velocities results in the following formula for the resolution R (line pairs per mm):

$$R = \frac{0.6}{d}\sqrt{\frac{V}{\frac{hc}{e}\left(\frac{1}{\lambda}-\frac{1}{\lambda_0}\right)}}, \qquad (1)$$

in which h is Planck's constant, c the velocity of light, e the electron charge, λ the wavelength of the incident light, λ_0 the wavelength of the transmission threshold of the photocathode, and V the voltage applied at the terminals of the tube (all in SI units).

A resolution of the order of 15 line pairs per millimetre can be obtained when the distance d is sufficiently small. When the useful diameter of the image is as

much as 60 mm, the resolution of the image can be as great as 1000 line pairs per diameter. The absence of distortion is very important, especially when it is required to measure physical characteristics, particle velocities or deformation of materials, or to examine spectra.

Opacity factor

In this range of shutters attenuation of the incident light is due chiefly to a metallic backing on the luminescent screen. The opacity factor depends on the thickness of the metal layer; it is limited by the number and dimensions of micro-fissures in the layer. If the image of a uniformly illuminated surface is examined when the tube is not supplied with power, a series of dots corresponding to faults in the backing will be observed. We have developed a method which allows an opacity factor of over 10^5 to be attained with a P 11 screen.

Shutter speed

The main factors limiting the shutter speed are the transit time of the photoelectrons, the conductivity of the electrodes, space-charge effects and the manner in which the tube is integrated into the transmission line.

If allowance is made for the shape of the voltage pulse, it can be shown that a *transit time* of less than 80 picoseconds allows the tube to be used with a minimum exposure time of 100 picoseconds. This is not, therefore, the limiting factor in present tubes but we shall see that it will affect future developments, and the same can be said of the *space charge*.

The way in which the problem of the *conductivity of the electrodes* is tackled can be explained as follows. Both the photocathode and the luminescent screen should have sufficient conductivity for the voltage pulse to travel to the centre of the tube without undue deformation. The screens commonly used in this type of tube meet this requirement. The normal photocathodes however are very poor conductors. We have measured surface-resistivity values of 10^8 ohms per square. A conducting sub-layer must therefore be applied, but this will absorb part of the incident light. For an exposure time of 1 ns the surface resistance of the sub-layer should amount to some dozens of ohms per square. This is obtained at a loss of the order of 50 % in gain. For the development of shutter tubes of even higher speed the use of highly conductive and therefore highly absorbing sub-layers will be unavoidable. In this case the light losses may be reduced by applying the conducting layer in an array of very narrow metallic bands, and the types of layer now in use will serve the purpose.

To obtain operating times of less than a nanosecond, we also have had to reconsider the fourth factor, the *matching of the tube in the transmission line.* In the first generation of shutter tubes, it is inserted in a strip-line terminated with its characteristic impedance. In the model which we are developing at present it is the tube itself which forms the strip-line, and the transition between this strip-line and the coaxial supply cables is inside the tube. Indeed, the presence of a dielectric (glass or ceramic) needed to seal the vacuum tube constitutes a considerable discontinuity which it is difficult to compensate for when one is interested in pulses having leading edges with a width of 0.1 ns. We have now developed vacuum-tight coaxial transitions which can transmit 50 ps pulses without any perceptible distortion. *Fig. 1* shows the pulse response measured with a reflectometer.

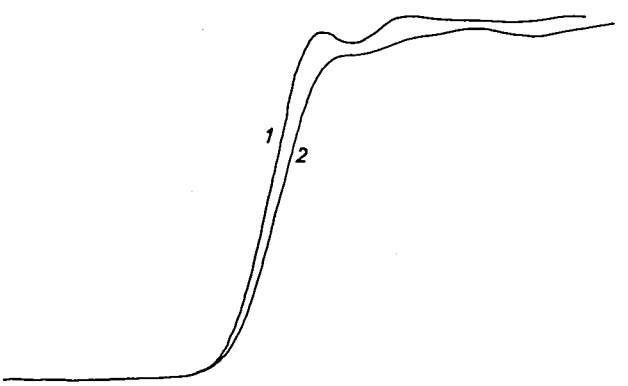


Fig. 1. Pulse response of ultra-high-speed shutter type 42-25, measured by reflectometer. Curve *1* input signal, rise time 100 ps. Curve *2* output signal, rise time 120 ps.

To obtain a sensitive zone with sufficient useful width on this tube, we had to use a characteristic impedance of 25 Ω. Transmission is then carried out with two matched coaxial cables with a characteristic impedance of 50 Ω which are joined inside the tube.

[4] E. Laviron, C. Delmare and H. Bacchi, Caméras électroniques de 1 ns de durée d'ouverture, Onde électr. **48**, 421-425, 1968 (No. 494).
[5] H. Bacchi and M. Blanchet, Générateurs d'impulsions haute tension de 1 nanoseconde à mi-hauteur, Onde électr. **48**, 430-437, 1968 (No. 494).
[6] G. A. Mesjac and V. V. Kremnev, Raideur maximale du front de montée dans un circuit *RLC* avec éclateur à étincelles, Bull. Acad. Sci. U.R.S.S., Sci. techn. En. Transp., 1963, No. 1, pp. 53-57 (in Russian).
[7] G. Eschard and R. Polaert, Tube obturateur pour photographie ultra-rapide, Onde électr. **48**, 426-429, 1968 (No. 494).
[8] H. Bacchi and J. Marillleau, in: High-Speed Photography, Proc. 8th Int. Congress, Stockholm 1968, pp. 57-60; E. Laviron and H. Bacchi, ibid. pp. 61-63.
[9] G. Eschard and R. Polaert, in: High-Speed Photography, Proc. 8th Int. Congress, Stockholm 1968, pp. 54-56.
[10] P. Dolizy and R. Legoux, Proc. 4th Symp. on Photoelectronic image devices, London 1968; see also R. Legoux, page 234 of this issue.
[11] G. Eschard and R. Polaert, Proc. 4th Symp. on Photoelectronic image devices, London 1968.

Photocathode

With the transfer technique mentioned above [10] it is possible to produce various types of photocathode, particularly the S 20 and S 24. We recently adapted this technique for the production of tubes with silver-caesium layers (S 1), which are sensitive in the near infra-red. *Fig. 2* shows the image of a test pattern obtained with radiation having a wavelength between 1 and 1.2 μm. A resolution of 18 line pairs per mm is obtained.

In view of the increasing interest in experiments using laser beams with wavelengths in this spectral region, shutters equipped with this photocathode may have many applications.

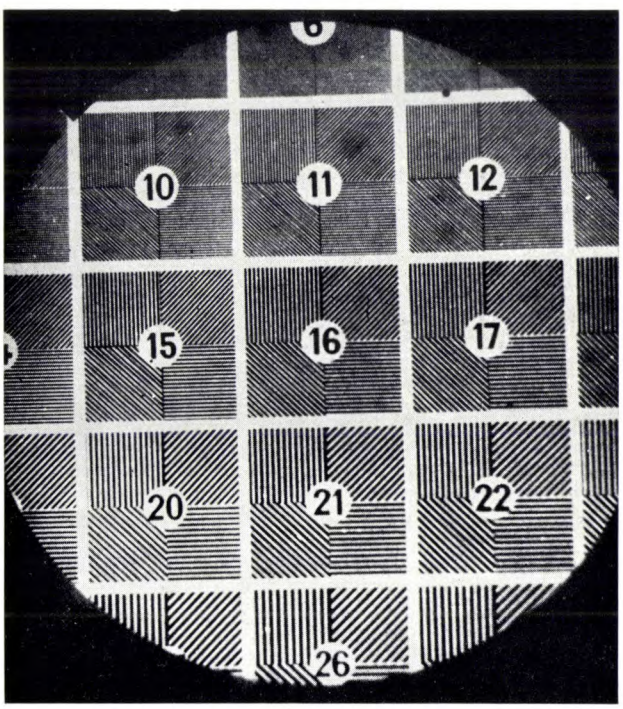


Fig. 2. Image of test pattern obtained from an S 1 photocathode with radiation having a wavelength between 1 and 1.2 μm.

Gain

Tubes developed so far allow us to attain photon gains of 30 at 430 nm.

In the cameras which we have developed, however, the repeater optics use only a small part of the luminous flux emitted by the screen. With tubes of large dimensions it is practically impossible to achieve high relative apertures without introducing troublesome distortion. The use of fibre-optic windows enabled us to resolve this problem. With the type 1325 shutter shown in *fig. 3* an overall gain of the order of 30 can be obtained between the light received by the photocathode and that which falls on the photographic plate.

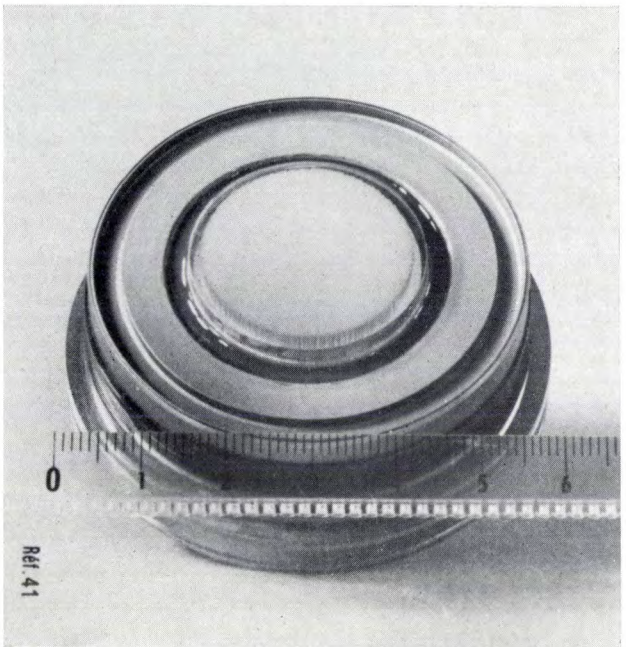


Fig. 3. The type 1325 shutter.

In a later phase image intensifiers incorporating channel multipliers will be used. First experiments show that this line of development is very promising.

Cameras

We have produced two families of cameras using different types of image-converter tube. Some of these tubes only have one stage, i.e. they consist of a single photocathode and a single screen used in accordance with the principle outlined above. Others comprise two single-stage tubes coupled by a fibre-optic plate inserted between the screen of the first tube and the photocathode of the second; these are called "two-stage tubes".

The first family consists of cameras whose exposure time is relatively long (5-500 ns). The chief applications for these cameras occur in ballistics and the study of explosives. The second group comprises cameras with extremely short exposure times which are at present in the nanosecond range, and which eventually may be reduced to several hundredths of a picosecond. These cameras are chiefly employed in the field of plasma physics.

5-500 ns camera with single-stage tube

A camera from the first family (5-500ns) is shown in *fig. 4*. It is equipped with a single-stage tube with a diameter of 38 mm or 60 mm. Its main components are the optical head, which contains the shutter tube and has a photographic attachment for polaroid film, and the pulse generator, of very small dimensions, con-

nected to the optical head by two flexible coaxial cables.

The pulse generator delivers a 15 kV pulse of nearly rectangular shape, with a width ranging from 5 to 500 ns and with a rise time of the order of a nanosecond.

The instant of occurrence of the high-speed phenomena to be photographed is usually not very well

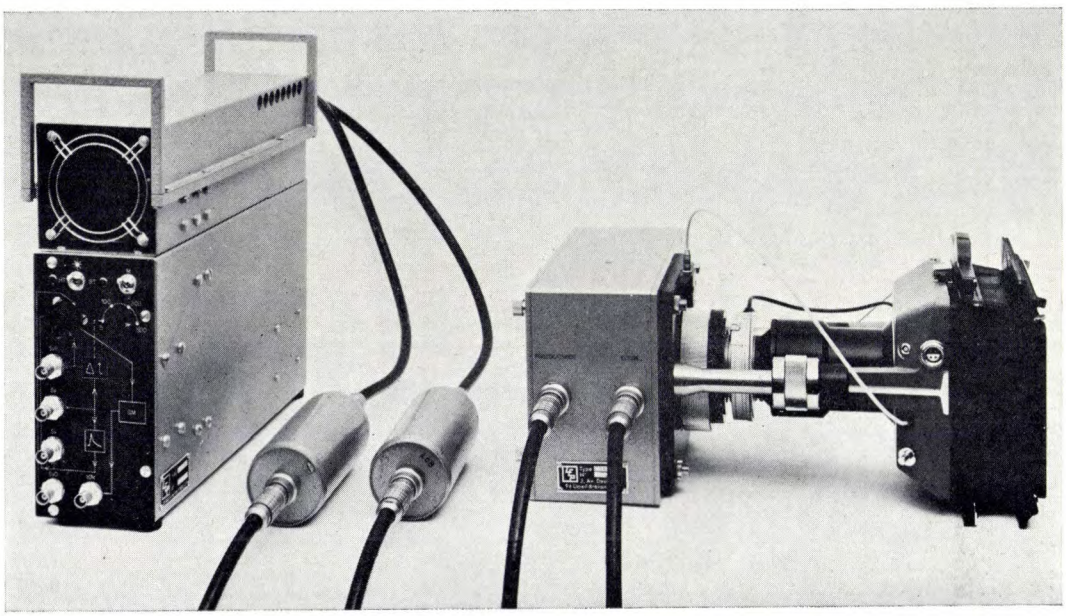


Fig. 4. Camera with 5-500 ns exposure time. The pulse generator with spark gap is shown on the left, the optical head with shutter tube on the right.

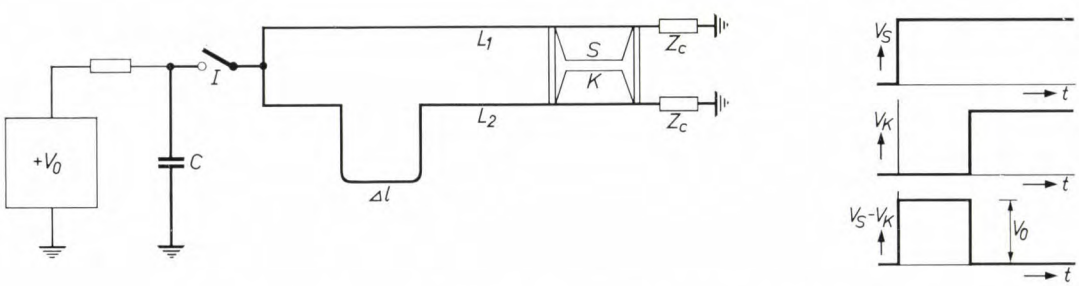


Fig. 5. Principle of pulse formation for the camera with 5-500 ns exposure time.

wave-front with an amplitude V_0 is transmitted through each line in the direction of the terminal load. The tube thus receives between its two electrodes a pulse $V_S - V_K$ of amplitude V_0 whose length is determined by the difference in length Δl between the two lines (see right-hand side of fig. 5).

defined. The exposure has therefore to be triggered by the phenomenon itself. Thus the low values of the trigger delay time (35 ns) and jitter (3 ns) are essential characteristics of the camera.

Fig. 5 shows the circuit diagram of the high-voltage pulse-forming circuit. It comprises a storage capacitor C, a switch I and two transmission lines L_1 and L_2 of unequal length terminated with their characteristic impedance Z_c. The shutter tube with photocathode K and screen S is placed between these two lines near their terminal loads. When the switch is closed, a

Fig. 6 shows a more detailed circuit diagram. The circuit is closed by triggering a 3-electrode spark gap. The trigger circuits comprise essentially a Marx generator built from transistors operating in the avalanche mode and a vacuum-tetrode power amplifier. These introduce a delay of the order of 15 ns.

The basic elements of the pulse-forming circuit are the storage capacitor and the spark gap. The essential feature of the storage capacitor is that it presents extremely low stray inductance of the leads. Its capacitance is 0.1 μF, its stored energy is about 10 joules

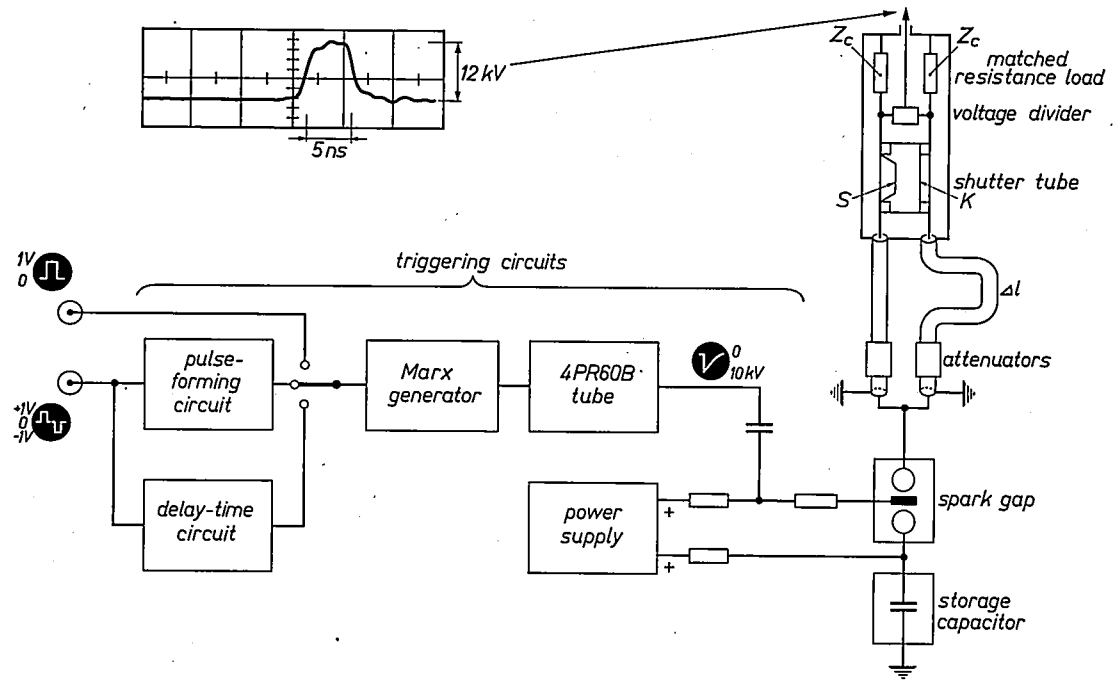


Fig. 6. Block diagram of the camera with 5-500 ns exposure time. The pulse forms at several points are indicated schematically. A more detailed diagram of the output pulse of the voltage divider is shown in the upper left-hand corner.

and its lead inductance does not exceed 2 nH. It is a product of the British firm Hivotronic.

The 3-electrode spark gap is, in fact, a combination of two 2-electrode spark gaps connected in series. Before sparking, the common electrode shows a very high input impedance, equivalent to that of a 2-3 pF capacitor. This arrangement makes it possible to use a vacuum tube for application of the high-voltage pulse. The delay which ionization phenomena would cause in a gas tube is thus eliminated. The time taken to set up the current is more or less proportional to the distance between the electrodes [6]. To ensure that this time is short enough, the distance between the electrodes has to be of the order of several tenths of a millimetre, which means that the gas pressure inside the spark gap has to be raised to approximately 15 bars.

The operation of a spark gap can be divided into two successive stages: ionization and discharge. The ionization time and its jitter depend mainly on what is called the overvoltage (the ratio of the voltage applied between the electrodes to the break-down voltage). *Fig. 7* shows how the ionization time and its jitter change as a function of the overvoltage. During discharge the equivalent circuit of the spark gap consists of an ideal switch in series with an inductance whose value is proportional to the distance between the electrodes. The proportionality constant depends mainly on the intensity of the current through the spark gap; it also

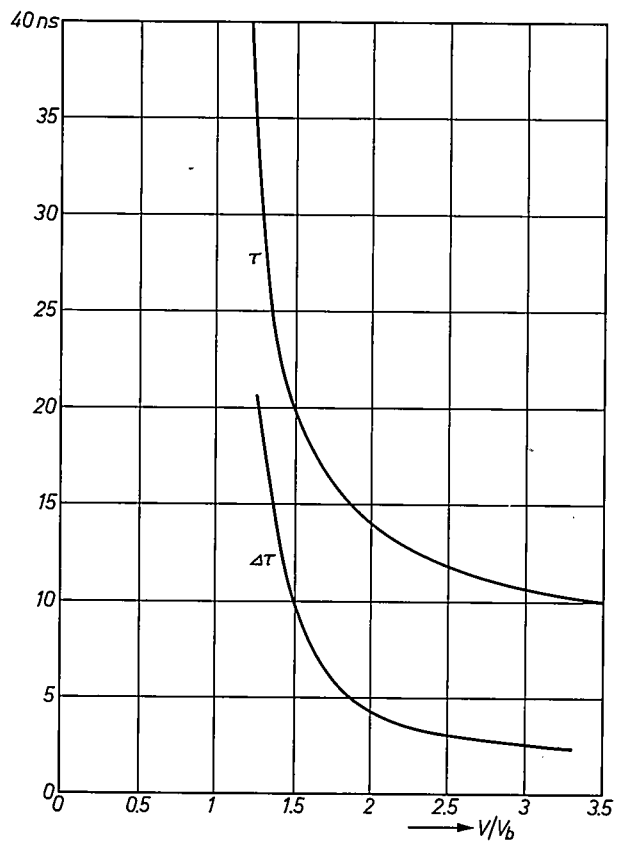


Fig. 7. Variation of the ionization time τ and jitter $\Delta\tau$ as a function of the "overvoltage" (voltage V applied to the electrodes divided by breakdown voltage V_b).

depends, but to a lesser degree, on the nature of the gas and the shape and nature of the electrodes, and it is practically independent of the pressure. For a current of approximately 500 A, the proportionality constant is of the order of 45 nH/mm.

We have studied the effect of a certain number of parameters, particularly the distance between the electrodes, on the service life of high-speed spark gaps. It was found that the further the electrodes are apart the longer the life of the spark gap is. The choice of the electrode spacing is therefore a compromise between operating life and a low rise time. Thus, in the case of the camera with which we are concerned here, an electrode spacing of 0.3-0.4 mm gives a rise time shorter than 1.5 ns and operation is still satisfactory after 30 000 discharges.

The main characteristics of this camera are the following. The high-voltage pulse applied to the shutter tube has an amplitude of 10-15 kV, a width of 5-500 ns, a rise time of 1.5 ns and a fall time of 10% of the pulse width. The trigger delay time is 35 ns, with a jitter of 3 ns. The maximum repetition rate is one pulse per minute.

A camera with a performance similar to that described here is being developed by SODERN (see page 214). It will have a sealed spark gap, which makes it easier to operate.

Test facilities for exposure times between 5 and 500 ns

We have built two test facilities (which in fact constitute a kind of experimental cameras) to aid the development of electronic shutters: one for use with single-stage 120 mm diameter tubes and the other for use with two-stage tubes. The circuits are of the same type as those described above.

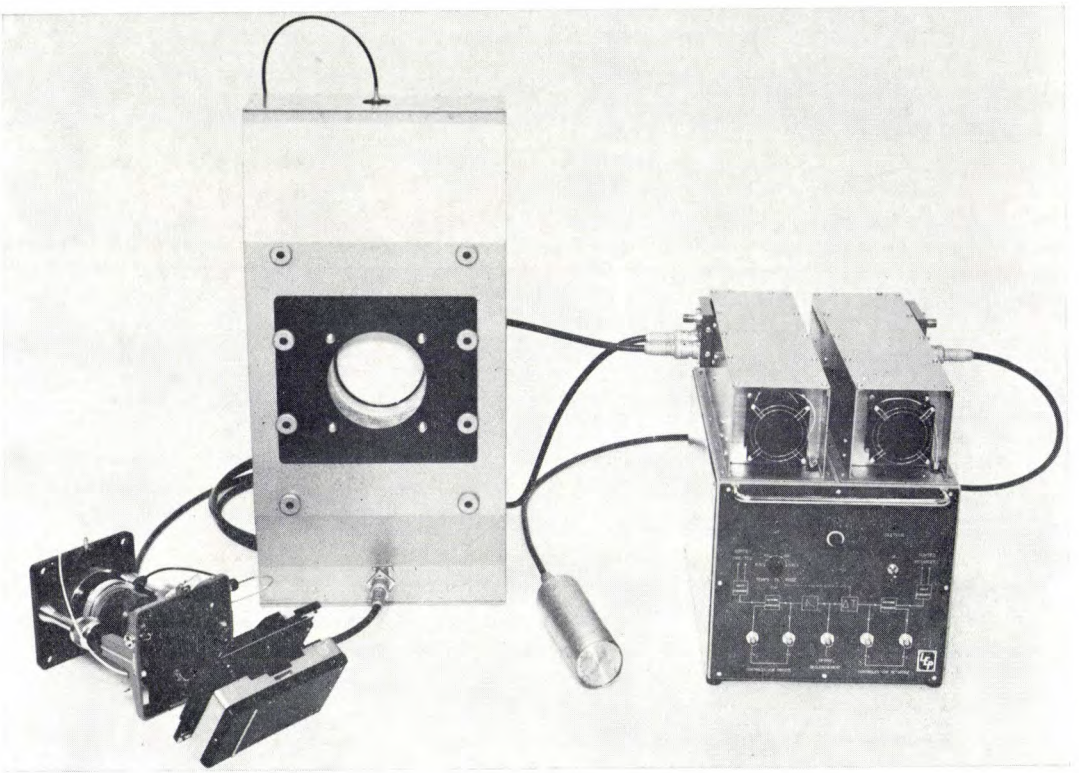


Fig. 8. Test facility for single-stage tubes with diameter of 120 mm.

Fig. 8 is a photograph of the test camera for tubes with a diameter of 120 mm. The amplitude of the high-voltage pulse is adjustable between 8 and 30 kV. The difficulties which had to be overcome to obtain a short exposure time are: 1) the capacitance presented by the tube, which is roughly 100 picofarads; and 2) the impedance mismatch caused by the presence of such a large tube in a strip transmission line which otherwise is connected to coaxial cables. We had decided to use strip-line about 230 mm wide. Because of the weight of the tube it is difficult to incorporate its supports in this line. On the other hand, this line should form the extension of an RG 214 U 50-Ω cable whose dielectric has an external diameter of 7.2 mm. The transition from this cable to the strip-line is dif-

ficult to achieve with good matching of the characteristic impedance and a sufficient dielectric strength to withstand 30 kV. Despite these difficulties the exposure time is still less than 10 nanoseconds. The performance in other respects is identical to that of the camera described above.

Fig. 9 is a photograph of the test camera for two-stage tubes. This system comprises two generators

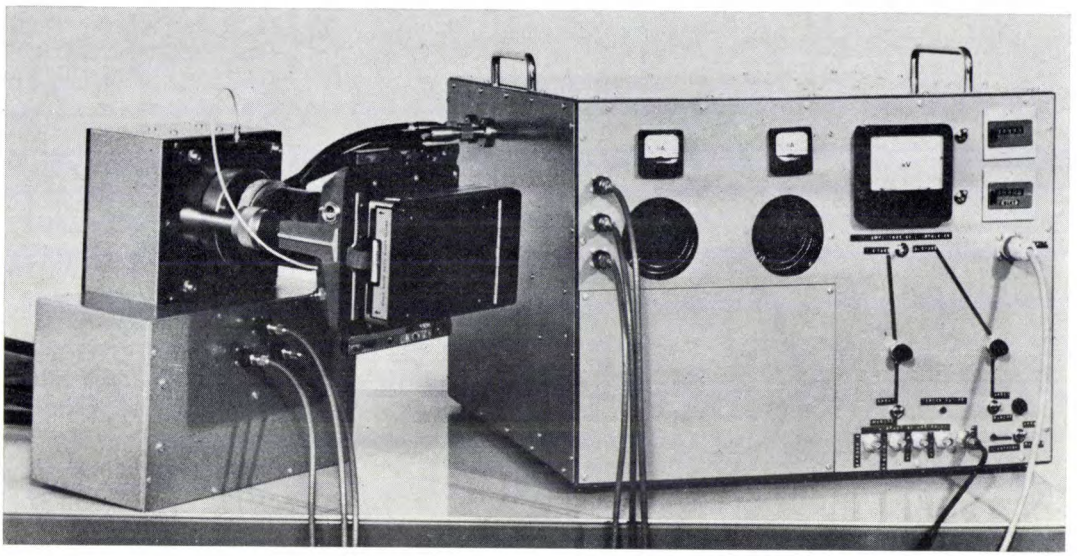


Fig. 9. Test facility for two-stage tubes.

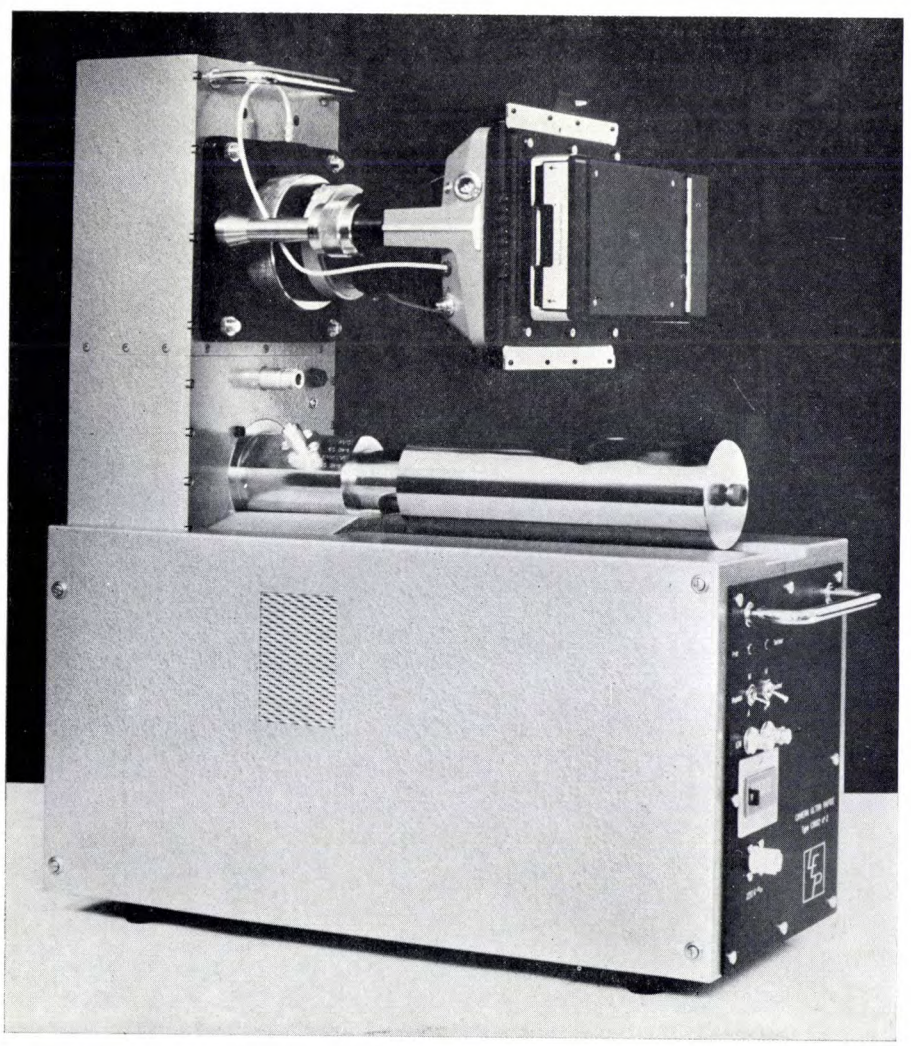


Fig. 10. Camera with 1-2 ns exposure time. The horizontal cylinder in the centre contains the pulse-forming line.

housed in a single box but independently controlled, one for the first stage and the other for the second stage of the tube under test. The generator for the first stage is identical to that used in the camera already described, but incorporates facilities for adjusting the voltage applied to the tube between 8 and 15 kV. The high-voltage pulse applied to the second stage should be rather wide because of the persistance of the luminescent screen of the first stage. The width of this pulse can be varied between 1 and 100 μs, and its amplitude between 8 and 15 kV.

Camera with exposure time of 1-2 ns

We shall now describe an apparatus belonging to the second family of cameras, allowing extremely short exposure times. The minimum exposure time is at present 1 ns. This should be reduced in the near future to 0.3 ns.

The pulse-forming line is contained in the removable horizontal cylinder.

To photograph a single random luminous phenomenon, such as a break-down in the air caused by the focusing of a laser beam operating in the triggered mode for example, the camera can be operated by a low-level pulse of about 6-10 V taken from a photocell receiving part of the laser beam. The image of the luminous phenomenon has then to be stored in an optical delay line for a time which is at least equal to the trigger delay time of the camera. In our case the optical delay lines used have delay times of approximately 35 ns. It is therefore absolutely essential that the 12 kV pulse which causes the tube to open should be applied to the tube with a delay of less than 35 ns.

Another essential characteristic of the camera is the jitter in this delay. To analyse a luminous phenomenon it is necessary to be able to shift, in steps of, say, 1 nano-

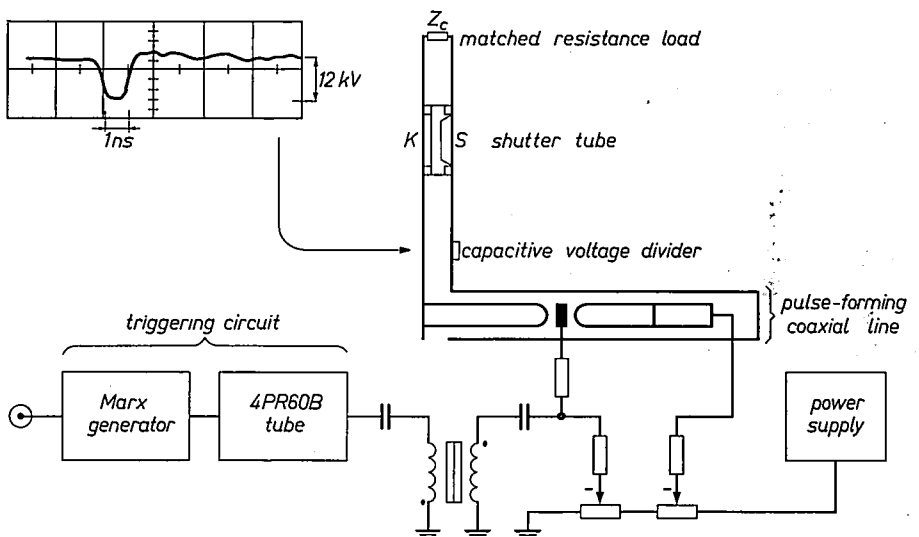


Fig. 11. Block diagram of camera with 1-2 ns exposure time. A diagram of the output pulse of the voltage divider is shown in the upper left-hand corner.

A camera of this type is again made up of two essential parts: the optical head, with the image-converter tube, and the high-voltage pulse generator (*fig. 10*). The converter tube, with which the input optics, the repeating optics and the photographic attachment are associated, can be seen at the top. A reflex viewfinder is used for focusing the image on the photocathode. The lower part contains the electronic control circuits. These consist mainly of a pulse amplifier and a high-voltage pulse-forming circuit which delivers a single pulse with an amplitude of 12 kV and a half-height width of 1 or 2 ns each time it is triggered.

second, the instant of exposure in relation to the triggering pulse applied to the camera (i.e. relative to the phenomenon studied). The jitter constitutes a limit to the precision with which this shift can be made, and is therefore a characteristic which is as important as the trigger delay time. In the camera described here we have reduced it to approximately 3 ns.

The circuit diagram (*fig. 11*) shows that the design of the camera has been influenced by a desire to reduce the delay and jitter at each amplification stage. The system comprises the following circuits:

1) A Marx generator employing transistors operating

in the avalanche mode. The main features of this circuit are its high sensitivity, its very short delay (less than 5 ns) and negligible jitter (0.2 ns). The amplitude of the pulse delivered readily attains 1 kV with a rise time of 2 ns. The difficulty is in coupling the generator to the vacuum tube which follows it, because the input impedance of this tube is that of a capacitor of about 50 pF, which, for a rise time of 2 ns, means currents of approximately 20 A. This is a value which is difficult to achieve with semiconductors, and a balance has to be struck between the rise time of the output pulse and the reliability of operation. Our circuit has a rise time of approximately 5 ns due to a series resistor inserted at the output of the Marx generator.

2) A power amplifier using a vacuum tube type 4 PR 60 B. This delivers the required output pulse of 15 kV at the required power (15 A for a rise time of 5 ns across a 5 pF capacitor). The input drive required for this output is 1 kV at the grid. A gas tube (e.g. a hot or cold-cathode thyratron) would also give this performance but the vacuum tube has the advantage that it introduces a delay of less than 10 ns, practically without jitter, whereas a gas tube would have a delay of 100-200 ns and cause a jitter of 20-50 ns.

3) A 3-electrode spark gap inserted in a coaxial arrangement consisting of an open-circuit pulse-forming line with an electrical length of 1 or 2 ns at one end, and a transmission line at the other, the latter being linked by a very carefully matched transition to a strip-line providing connection to the shutter tube. The characteristic impedance of the device is 25 ohms. The d.c. voltage to which the pulse-forming line is charged may be as high as 30 kV (for a 15 kV pulse applied to the tube). The spark gap is pressurized with high-purity nitrogen at a pressure of 40-50 bars.

As already explained, a distance of the order of one tenth of a millimetre has to be maintained between the electrodes to obtain the 300-400 ps rise time required for an exposure time as low as 1 ns. This meant that considerable difficulties had to be overcome to ensure a suitable mechanical strength at 50 bars pressure without preventing a rapid change of the exposure time; a dielectric strength to withstand 300 kV d.c.; a negligible mismatch; sufficient reproducibility of spark-gap adjustments after each change in exposure time; sufficiently long service life, despite the spark erosion and the spacing of the order of 1/10 mm between the electrodes; and finally the exclusion of spontaneous or untimely triggering.

By submitting the effect of each of the parameters of the pulse-forming circuit to extremely detailed analysis we were able to give the system optimum characteristics for each of the above requirements, and at the cost of extremely stringent mechanical tolerances, strict inspection of surface conditions, and far-reaching technological precautions during assembly we have succeeded in ensuring satisfactory reproducibility and reliability. The performance obtained can be summarized as follows:

a) Overall trigger delay time of less than 30 ns.
b) Jitter of less than 3 ns.
c) Spontaneous triggering occurrence of less than 2%. Incidentally, complete suppression of unwanted triggering has been achieved with a system of automatic cut-out of the extra-high voltage; with this system the camera has to be "set" in advance, one second before it is used.
d) Service life exceeding 50 000 discharges.
e) Rise and fall time of approximately 350 ps.
f) Spurious signal amplitudes measured on the tube of less than 10% of the pulse amplitude.

The electrical pulse applied to the tube is measured by a capacitive divider inserted in the transmission strip-line near the tube. This divider introduces an initial attenuation of 54 dB, with an inherent rise time of 0.2 ns. The difficulty was to obtain this performance while giving the divider a "time constant" (the period during which it introduces dynamic errors of less than 5%) of approximately 5 ns.

Fig. 12 shows the transition between the pulse-forming coaxial line and the strip-line and also how the tube is inserted in the latter.

To sum up, we may say that the principal feature of the cameras which we have just described is that they permit full-image exposures with the following characteristics: very short exposure times, e.g. 1 ns; an extremely wide range of exposure time, e.g. 1-500 ns; a very low trigger delay time and jitter (30 and 3 ns, respectively); a very substantial light gain; and an almost total absence of distortion.

This last feature in particular opens up prospects for many interesting applications such as three-colour printing, photo-feedback [12] and relief photography. All these processes involve the superimposition of several images and thus require an almost complete absence of distortion.

Improvement of the cameras is possible in a number of directions, such as the development of luminance intensifiers using channel multipliers giving a considerably higher light gain; the development of new super-high-speed tubes characterized chiefly by the fact that the sensitive layers (photocathode and screen) are integrated into a strip transmission line, which will

---

[12] M. L. Perl and L. W. Jones, The regeneration image intensifier and its application to the luminescent chamber, Proc. 4th Symp. on Photoelectronic image devices, London 1968.

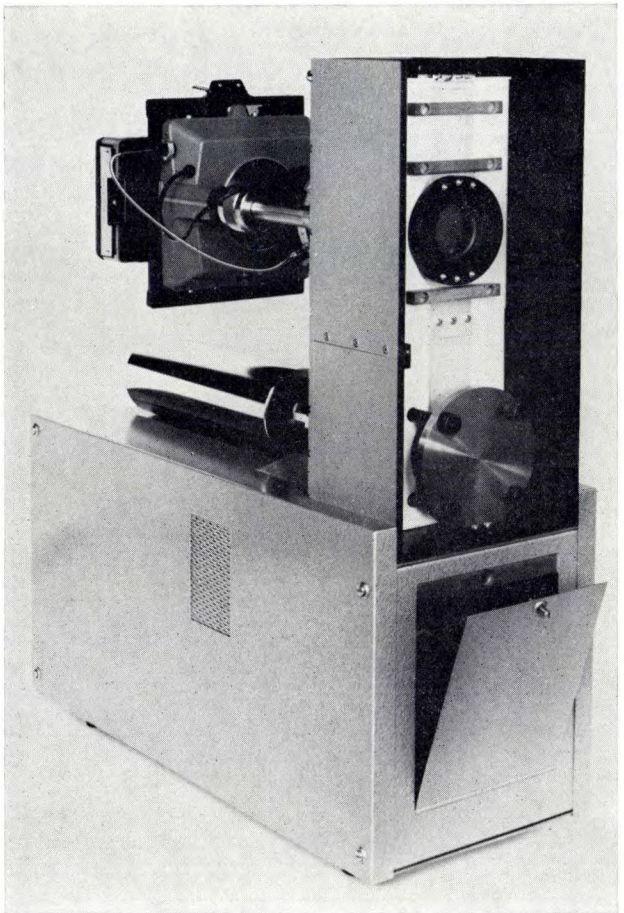


Fig. 12. The same camera as in fig. 10, seen from another angle. The transition between coaxial line and strip-line is seen, as well as the insertion of the shutter tube in the strip-line.

allow the exposure time to be reduced to approximately 300 ps; the development of new types of triggered spark gaps whose essential characteristic will be that they have a jitter as low as 300 ps; the miniaturization of the optical head, enabling several shutter tubes to be grouped together within a small space, so that full-image exposures of several successive stages of the same phenomenon can be made.

Summary. After briefly reviewing the various types of camera which can be used for ultra-high-speed photography, the authors describe several image-converter shutter tubes and various cameras developed and produced at LEP. The shutter tubes are of the planar-diode type. By making use of a fibre-optic plate they can be connected in cascade to form a two-stage tube or combined with channel multipliers for a much higher light gain. They may have a diameter of 38 mm, 60 mm or 120 mm. The cameras developed can be grouped into two main "families": a) Cameras with exposure times of 5-500 ns; a brief description is given of a single-stage tube version with a diameter of 38 mm, a two-stage version with a diameter of 35 mm and a single-stage version with a diameter of 120 mm. b) Cameras with exposure times of the order of 1 ns. Two versions have been made, with exposure times of 1 ns and 2 ns, using a single-stage tube and a diameter of 38 mm. Development still in progress should result in a camera with an exposure time of 300 ps. All these cameras are actuated by an electrical pulse with an amplitude of several volts. The trigger delay is 35 ns and the jitter, at present 3 ns, will be reduced to about 300 ps as a result of progress made in the field of ultra-high-speed spark gaps. The principal features of these cameras are their light gain and the absence of distortion on one hand and the extensive exposure range and the low delay and jitter values on the other.

The production of electron-multiplier channel plates

G. Eschard and R. Polaert

The concept of an electron multiplier with a single continuous dynode has led to the rapid development of channel multipliers. These devices have important applications; in particular they can be used in sensitive photoelectric tubes for image intensifiers and as particle detectors in space research[1][2]. The amplification produced by these channel electron multipliers results from the secondary emission which occurs when a semiconductor surface is bombarded by electrons. Unlike the multiplying surface of the electron multipliers with separate dynodes, that of the channel multipliers is continuous, and is formed by the partially conducting inner wall of a glass tube. A potential difference is applied between the ends of the tube, producing a uniform electric field inside. When an electron or ionizing particle strikes the inner surface of the tube, several secondary electrons may be emitted. These electrons are accelerated into a parabolic trajectory by the electric field, strike another point on the inner wall, and in turn give rise to the secondary emission of a shower of electrons. The number of times this collision process is repeated depends on the length and diameter of the tube and the magnitude of the applied voltage. The ratio between the number of electrons at the output to those at the input of the tube may be as great as 10^7.

In practice, the single multiplier tube is replaced by a bundle or stack of very narrow tubes. By taking a section of the bundle a device known as microchannel plate is obtained.

These novel components can be used, for example
1) in photomultipliers, where they improve the speed of response by cutting down the overall transit time of the electrons[3];
2) in image-intensifier tubes, where they give very high gains in a small volume and with low supply voltages [4][5];
3) in cathode-ray tubes, where they can be used to increase the writing speed.

This article first considers the number and size of the channels in a plate and the choice of the material. Some methods of producing these plates are then outlined.

Dimensions of the microchannel plates

For the plates used in image intensifiers the number and diameter of the microchannels are determined by the required definition and the dimensions of the image. For example, an image with 400×400 lines and a side 40 mm in length requires the use of a plate with at least 64×10^4 channels, and a channel diameter of less than 40 microns. For accurate calculation it is necessary to take into account the thickness of the walls, the quality of the stacking of the channels, the electron-optical system in front of and behind the plate, and the grain of the screen.

The length of the channels (i.e. the thickness of the plate) is determined by the desired electron gain. Theoretical calculations show that the gain of a multiplier is mainly a function of the ratio of the length to the diameter of each channel. This ratio lies between 40 and 60.

Choice of material

The basic material is glass. It is used in the form of a tube and is chosen because it can easily be drawn out and has stable properties under high-vacuum conditions. However, glass is usually an excellent insulator, and for this application the conductivity must be increased[6]. There are two methods of achieving this. A conducting layer can be deposited inside the tube; this method is difficult to put into practice because of the small diameter of the channels. Alternatively a glass of special composition is used, which has either a bulk or a surface conductivity.

Production technology

The formation of a bundle of several million glass channels of microscopic dimensions can be carried out in various ways.
1) The channels can be made by piercing a glass disc with thousands of holes. This can be done by photo-etching methods or by electron bombardment.
2) A glass tube can be drawn out to obtain a tube with a microscopic internal diameter. This hollow fibre is cut into regular lengths which are finally assembled into a bundle.
3) A glass tube with a solid core is drawn and compressed. Only the last two methods have been developed into useful manufacturing techniques.

Drawing a hollow glass tube

A glass tube which has been brought to its softening temperature can be drawn out into a hollow fibre of

G. Eschard, Ingénieur E.S.E., Dr. Ing., and R. Polaert, Ingénieur H.E.I., Dr. Ing., are with Laboratoires d'Electronique et de Physique Appliquée, Limeil-Brévannes (Val-de-Marne), France.

microscopic diameter. This fibre is wound on a drum and then cut into regular lengths, which are assembled into a bundle using enamel as a sealing material.

For this operation two production processes are used. The first, known as the "direct drawing" process, is generally used to obtain hollow fibres with a diameter of more than 200 μm. The second process consists of drawing in two or three stages, bundling being carried out after each stage to facilitate the regular arrangement of the fibres.

This latter method is preferred for the fabrication of hollow fibres with a diameter of less than 100 μm. Such fibres can also be obtained by direct drawing, but in that case the wall thickness will be too large with respect to the channel diameter. This can be explained as follows. When the tube passes through the hot zone of the drawing furnace, the viscosity of the glass decreases. The glass can then be pulled, but at the same time the relative thickness of the tube wall increases because of surface tensions, which increase as the radius of curvature becomes smaller. This is undesirable since it is important that as many as possible of the electrons accelerated at the entry of the channel plate should be able to strike the interior wall of the channels and take part in the multiplication process. The thickening of the walls diminishes the useful area of the channel plate and reduces the efficiency of detection. We have developed a method of maintaining a small excess gas pressure inside the glass tube during the drawing process, and this results in a tube with a useful cross-section of 80%.

To ease assembly of a very large number of fibres of small diameter it is preferable to proceed in several stages as stated above. The glass tube is first drawn out until the inner diameter of the hollow fibre is less than 1 mm. This fibre is rolled on a drum, and then cut into pieces of constant length which are assembled, enamelled and arranged in a hexagonal mould. This is placed in a furnace to seal the fibres; the arrangement of the fibres is perfectly regular if they all have the same diameter. The number of fibres contained in the hexagonal mould is determined by the relation:

$$N = 1 + 3m(m + 1),$$

where m is the number of adjacent fibres arranged along one side of the hexagon.

The second drawing operation is then performed to reduce the fibre diameter even more, but this time it is no longer possible to wind the bundle on a drum. It is simply pulled along by the movement of two belts pressed against each other.

To ensure that the final dimensions are uniform it is vital that the input velocity of the tube, the output velocity of the fibre, and the furnace temperature should all be held constant during both drawing processes.

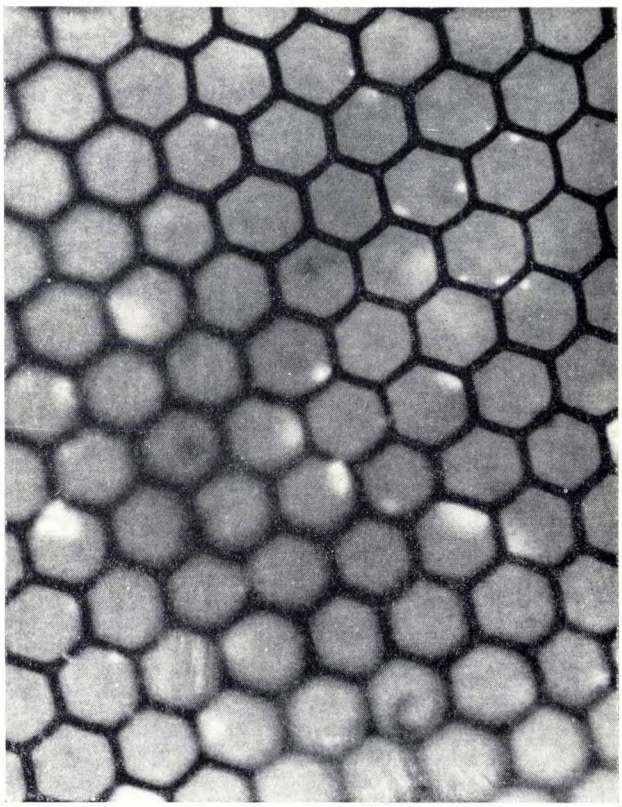


Fig. 1. The cross-section of a microchannel plate produced by the technique of drawing a glass tube in two successive steps. Channel diameter 40 microns. A useful cross-section of 80% is achieved by maintaining a small excess gas pressure inside the glass tube during the drawing process.

Fig. 1 shows the cross-section of a channel plate obtained by this process.

Use of a tube with solid core

If a tube with a solid core is used the core is dissolved away after the plates have been cut. This technique offers several advantages: the fibres are less fragile, manipulations are easier, the fibres can be subjected to strong pressure in order to compress them without fear of crushing or deforming them, all the channels

[1] W. C. Wiley and C. F. Hendee, Electron multipliers utilizing continuous strip surfaces, IRE Trans. **NS-9**, No. 3, 103-106, 1962.

[2] Cf. the article by J. Adams and B. W. Manley, The channel electron multiplier, a new radiation detector, Philips tech. Rev. **28**, 156-161, 1967.

[3] J. Nussli, Photomultipliers and the transfer technique, page 236 of this issue.

[4] A. Guest, R. T. Holmshaw and B. W. Manley, Channel multiplier plates for imaging applications, Proc. 4th Symp. on Photoelectronic image devices, London 1968.

[5] G. Eschard and J. Graf, Quelques problèmes concernant les multiplicateurs canalisés pour intensificateur d'image, Proc. 4th Symp. on Photoelectronic image devices, London 1968.

[6] H. J. L. Trap, Les effets de quelques traitements conventionnels sur la conductibilité superficielle du verre, Verres et Réfr. **23**, 28-37, 1969 (No. 1).

maintain the same diameter, and they therefore have the same electron gain.

In the application of this technique a *metal* core or a *soluble glass* core can be used. In the first case a metal wire of the exact diameter of the desired channel is placed inside a glass tube. The tube is then passed through a furnace and drawn out so as to cover the metal wire and assume the required wall thickness. The resulting fibre is then wound directly on to an octagonal drum with the turns touching. The coil is cut into eight parts which are put together to form a rectangular block. The stack is subjected, while hot, to strong pressure in order to seal the fibres to each other. It is then cut transversely into thin sections forming the microchannel plates. Finally the metal contained in each channel is dissolved by chemical or electrolytic means.

The channels obtained in this way are perfectly regular, as shown in *fig. 2*.

The alternative process is to draw a tube with a core of soluble glass. This is an adaptation of the practice used for making optical fibres [7]. In contrast to the latter process the glasses are not chosen for their difference in refractive index but for the solubility of the glass forming the core of the fibres and for the electrical properties of the multiplier glass. As in the production of optical fibres the drawing of the multiplier channels is generally done in several stages. The regularity in diameter is not quite as good as in the metal-core method, but is still quite satisfactory.

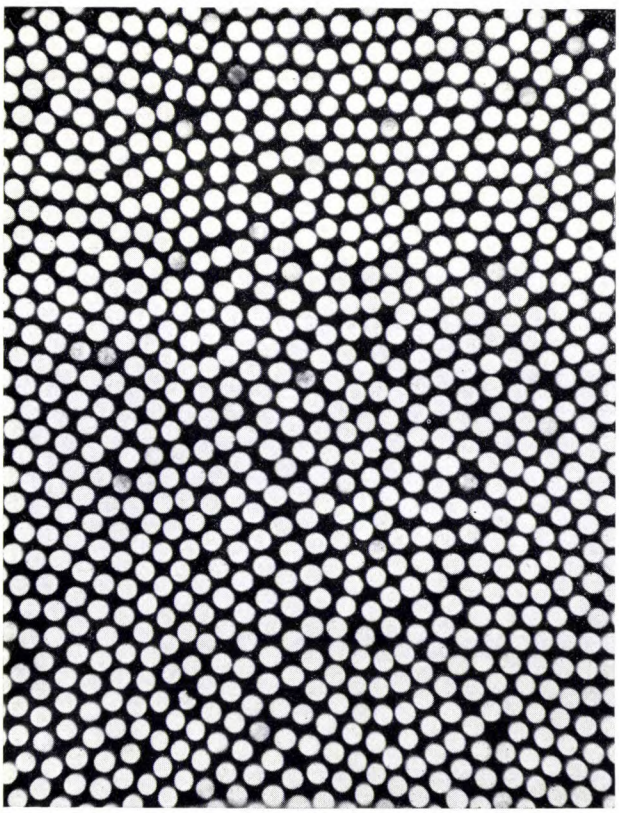


Fig. 2. The cross-section of a microchannel plate produced from a glass tube with a metal core. Channel diameter 80 microns.

Measurement and control of the diameter of the glass fibres

The above description has shown the need for precise control of the diameter of the fibres during the drawing process. This is necessary to obtain a regular arrangement of the fibres, and to obtain an identical electron gain in all channels.

Measuring the diameter is difficult because as the glass leaves the furnace it is fragile, hot and moving rapidly. For the hollow fibres in any case it is impossible to make the measurement by mechanical contact. The technique developed at LEP is purely optical; see *fig. 3*. The fibre passes through one of two uniform parallel beams of light intercepted by two photoelectric cells.

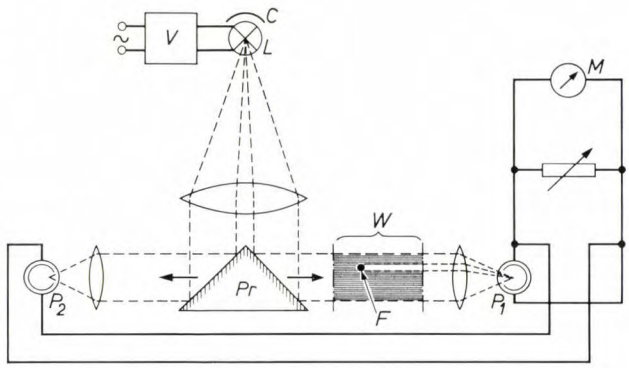


Fig. 3. Method for continuous measurement of the diameter of the fibre leaving the drawing furnace. *F* cross-section of fibre; the movement of the fibre is perpendicular to the plane of drawing. *L* lamp with condenser *C* and stabilized power supply *V*. *Pr* prism dividing the light between photocells P_1 and P_2, whose outputs are compared in meter *M*. Part of the beam directed on to P_1 is intercepted by the fibre *F*. Lateral displacement of *F* within the "window" *W* of 4×25 mm is permitted.

The difference in luminous flux picked up by the two cells is proportional to the outer diameter of the fibre. The fibre can be displaced to either side of its mean position, within a "window" of 4×25 mm, without affecting the accuracy of measurement.

The measuring accuracy for fibres of 0.5 mm diameter is $\pm$ 1%; in the case of very small diameters (of the order of 20 μm) an uncertainty of $\pm$ 1 μm must be added to this.

Besides the fact that there is no mechanical contact, this method of measurement has the advantages that it is accurate, quick and unaffected by vibration and ambient light.

The same optical arrangement can be used for both

opaque and transparent fibres, but the diffraction and refraction effects are different. The system must therefore be calibrated under the conditions of use.

Control of the diameter of the fibres is obtained by amplifying the error signal generated by the diameter reader. This signal controls a servomotor which drives a speed control. In this way the speed of drawing and any consequent variation in diameter can be corrected. The time constant of the control system is about 0.7 s. Part of this interval represents the time elapsing as the fibre is pulled from the drawing zone to the place where its diameter is measured.

A new method of measurement, which is based on dielectric losses caused in a resonant microwave cavity, is now being developed. The fibre is made to pass through a cavity forming part of a transmission circuit, and the resulting mismatch gives a very accurate indication of the amount of material present in the cavity at any moment. By combining this method with the optical measurement it is possible to obtain a continuous check on both the diameter and the thickness of the fibre.

Conclusion

The technology for producing microchannel plates, in whose development LEP have cooperated closely for several years with the Mullard Research Laboratories at Salfords, England, and the Philips Research Laboratories at Eindhoven, has certain parallels with that for optical fibres. The distinctive feature is that the channel diameter has to be very uniform: this diameter is important because it determines the electron gain. To achieve this uniformity the equipment must incorporate extremely accurate control of the diameter, the furnace temperature and the speed of drawing.

A good demonstration of the capabilities of these microchannel plates is provided by a particular photomultiplier of original design. This photomultiplier, which is illustrated and more fully discussed in another article in this issue [3], consists of a microchannel plate placed between the photocathode and the anode, which is the open end of a coaxial line with an impedance of 50 Ω. This arrangement gives a transit time of less than 0.1 ns for the electrons. In a conventional photomultiplier the electron transit time is usually 15 ns or longer. The performance of the new photomultiplier opens up new prospects in the measurement of the duration of light pulses.

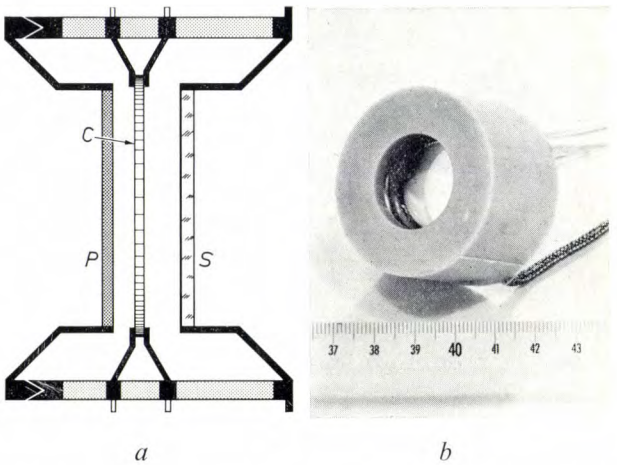


Fig. 4. *a*) Schematic cross-section of a novel image intensifier tube, consisting of a photocathode *P*, a microchannel plate *C* and a luminescent screen *S*.
b) View of the tube in its housing.

Another example of the application of a microchannel plate is given in *fig. 4*. This shows an image intensifier, in which the photocathode is mounted immediately in front of the channel plate. A very high electron gain is thus achieved in a simple arrangement. The photocathode of this tube is activated by means of the transfer technique described in another article in this issue [8].

Summary. The principle of operation of channel electron multipliers is outlined. Some applications of arrays of these multipliers in plates consisting of several million channels are considered and their physical and electrical properties summarized. The chief methods of producing these channel plates are described. The basic material is generally glass in the form of tubes several millimetres in diameter, which can be reduced to the final diameter by direct drawing. The tubes are then stacked and compressed to form the channel plates. Intermediate steps may be introduced with two or three successive drawings. Other methods involve filling the tubes with a substance which is dissolved at the end of the drawing and stacking operation. To ensure the uniformity of channel diameter (which is necessary for a uniform electronic gain) all of these methods require very precise control. An optical device is described which enables the diameter of a tube or a glass fibre to be measured continuously and to be kept constant to within about ± 1%.

[7] See for example N. S. Kapany, Fiber optics; principles and applications, Academic Press, New York 1967.
[8] R. Legoux, The transfer technique, a new method for activating cathodes of phototubes, page 234 of this issue.

A high-speed oscilloscope for real-time use

G. Andrieux and C. Loty

Introduction

The oscilloscope is probably the one instrument that the electronic engineer regards as an absolute essential for his various tests and measurements. Since the oscilloscope is so vital a measuring instrument, its development should not just match the general progress of electronics; it should always be one step ahead. Maintaining such a lead requires a great deal of effort, since the oscilloscope must at the same time remain a general-purpose instrument and yet give a performance equal to that of the most sophisticated electronic equipment.

Progress in electronics has usually been associated with faster action or response. Thus, thermionic valves gave pulses which had a duration of tens of nanoseconds, present-day solid-state devices give pulses measured in fractions of a nanosecond, and the laser offers the promise of pulses in the picosecond region. Such high-speed signals can only be displayed on an oscilloscope if it has a very large bandwidth. Now, the best amplifiers that can be produced today have a rise time of the order of a nanosecond. Beyond this, two techniques are available: real-time operation without amplification, and sampling oscillography. The sampling technique, in which normal amplification can be used, has permitted remarkable progress to be made, but it is not applicable to all signals, particularly those that only occur once, for which real-time observation is the only possible method.

High-speed oscilloscopes are thus divided into two types. With the first, the sampling oscilloscope, recurring signals can be displayed in the usual way in which oscilloscopes have been used since their inception. The other type, the real-time high-speed oscilloscope, was developed for a different procedure: it enables fast and unique signals to be studied by means of a photographic record, whereas recurrent signals can be displayed as a subsidiary function. This procedure, together with the absence of amplification, imposes fresh limitations on the real-time oscilloscope.

In the classical oscilloscope, the cathode-ray tube is no more than a display element whose own performance does not have much effect on the performance of the equipment. Similarly, the ancillary circuits for the time-base, for synchronization, beam switching, etc., are only of secondary importance. The actual measuring element is really the signal-channel amplifier. (This may seem a strange statement, but its truth becomes apparent to anyone who has to write out an order for one of the modern oscilloscopes with plug-in units.) On the other hand, in the real-time high-speed oscilloscope the cathode-ray tube is the vital element whose intrinsic qualities determine the overall performance of the instrument. Moreover, the ancillary circuits, which have to satisfy very difficult conditions because of the high speed of the signal to be examined, acquire a much greater importance. This is particularly so for the delay line — an element which is not used in an ordinary oscilloscope, but whose characteristics can have a marked effect on the design of the time-base circuits and the cathode-ray tube itself.

Development of the cathode-ray tube [1]

We have already stated that the essential characteristics of the oscilloscope will be those of the cathode-ray tube. It is therefore useful to list these characteristics before going on to consider how to improve them together rather than separately. In fact, we shall show that they are not independent, and that as in other situations a figure of merit can be applied. There are four essential characteristics: the bandwidth, the sensitivity, the useful scan on the screen and the maximum writing speed. The first three are obviously related to the signal channel of the tube. The fourth, which refers to the ability to record a non-recurrent fast signal, is directly dependent on the energy density in the electron beam.

The bandwidth is of prime importance since we are interested first and foremost in *high-speed* oscilloscopes. In fact, the rate of response to a signal is limited fundamentally by the transit time of the electrons between the deflection plates. This transit time can be reduced by making the plates shorter but if this is done the sensitivity is also reduced. The same thing happens if the velocity of the electrons is increased by increasing the accelerating voltage. Finally, the spacing of the plates and their length can be reduced together; the sensitivity would then remain the same but there is then less room for the beam to pass between the plates. This

G. Andrieux, Ingénieur E.R.B. and E.S.E., L. ès Sc., and C. Loty, Ingénieur C.N.A.M., are with Laboratoires d'Electronique et de Physique Appliquée, Limeil-Brévannes (Val-de-Marne), France.

means that the current has to be reduced, and hence also the maximum writing speed. We have thus the first indication that there is a figure of merit relating bandwidth, sensitivity and writing speed.

One obvious solution to the problem of the bandwidth is to arrange a series of pairs of small deflection plates, for which the transit time is very short, and to incorporate these in a transmission line in such a way that the velocity of propagation is the same as that of the electrons. A series of additive deflections is then obtained and the total sensitivity becomes normal again. This is a well-known principle which is applied in every travelling-wave tube, and it has already been applied in the most advanced oscilloscopes of the classical type. This has been done by making use of a cathode-ray tube with divided plates which are linked by external inductances to form a lumped-element transmission line.

To carry this development one step further, it is necessary to put the inductances inside the tube and to reduce them to simple loops (*fig. 1a*). The delay line formed in this way is in fact a constant-K filter, and has a cut-off frequency. It also has a phase dispersion, which limits the rise time well before the value corresponding to the cut-off frequency has been reached. In fact, it can be shown that the rise time as limited by the phase dispersion always exceeds the transit time per element. In order to obtain an appreciable increase in bandwidth, it is therefore necessary to divide the deflection line into a large number of elements each of which is very small. The capacitance and inductance becomes more or less distributed and in the limit the line becomes a zigzag line (fig. 1*b*). The practical limit of such a line is in the region of 1 GHz.

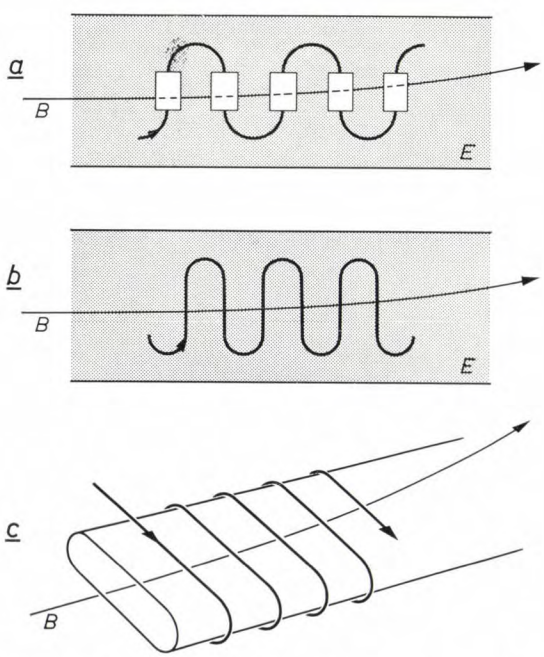


Fig. 1. *a*) Deflection line for a large bandwidth. *E* earthed deflection electrode. The inductances linking the small deflection plates have been reduced to simple loops. *B* electron beam. *b*) Zigzag line. *c*) Helical line.

The zigzag line is not a line whose constants are perfectly distributed: because of the alternating change in the direction of propagation there is an appreciable coupling between adjacent turns, and this coupling may be taken as being responsible for the phase dispersion just mentioned. The dispersion is much smaller for a deflection line wound in the form of a helix (fig. 1*c*). It can in fact be made negligibly small and the bandwidth is then only limited by the transit time for a single turn. However, although the bandwidth that can be obtained in this way is very much greater than before, a new limitation appears: in order to make the phase dispersion sufficiently small the transverse dimensions of the line have to be reduced and this reduces the space available for the passage of the beam, and hence the maximum writing speed. Moreover, to increase the sensitivity, the deflection line has to be made longer by giving it more turns, but its losses at the higher frequencies cannot then be neglected. We thus find another, more sophisticated, relationship between the sensitivity, the bandwidth, and the maximum writing speed. We should also add that although making the deflection line longer increases the sensitivity, it reduces the useful area of the screen by intercepting the beam.

As a result, the four principal characteristics are linked by a figure of merit whose best value is usually obtained when a helical deflection line is used. A mathematical analysis of the problem shows that there are only two independent means of increasing this figure of merit: reducing the diameter of the spot or increasing the supply voltages on the tube electrodes. For the first case, it would for example be possible to reduce the distance between the screen and the deflection line; other things being equal, this would reduce all the dimensions of the image, including the spot diameter. If the sensitivity and the useful scan are expressed in terms of the spot diameter, these characteristics remain unchanged, but since the same luminous flux is distributed over a smaller area, the brightness is increased and hence the maximum writing speed.

In the second case, the length of the tube could for example be increased, with the main objective of increasing the useful area of the screen. Even though the tube will be longer, the spot diameter and the beam current can remain unaltered since the higher voltage reduces the effect of the space charge; nevertheless, the

[1] See the articles in Acta Electronica **10**, No. 4, 1966, which all relate to the real-time oscilloscope. — This study was supported by the Commissariat à l'Energie Atomique, Direction des Applications Militaires.

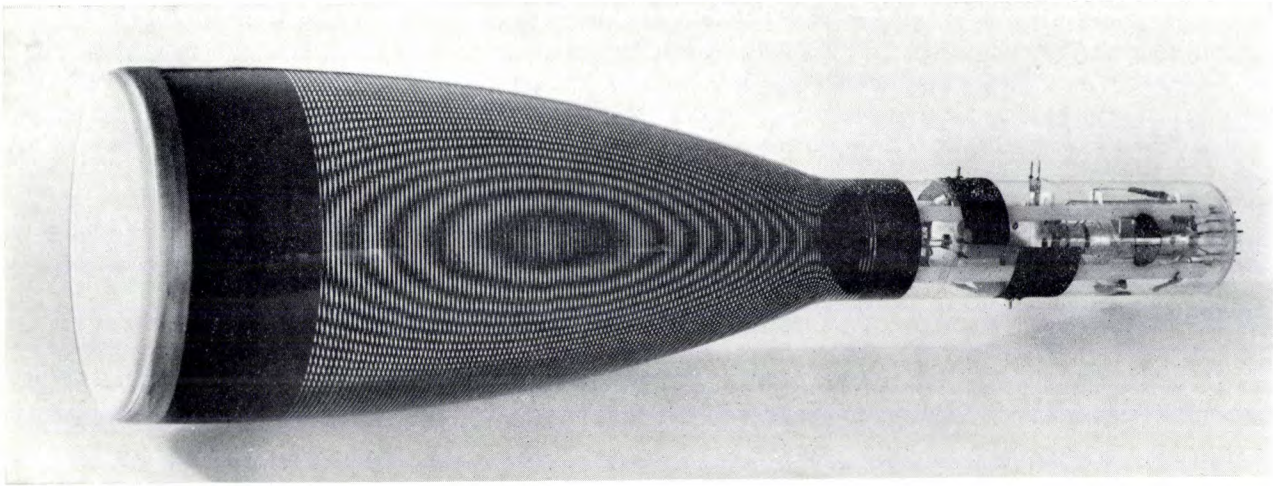


Fig. 2. 5 GHz cathode-ray tube 50 D 13 BE, developed for the Direction des Applications Militaires du Commissariat à l'Energie Atomique of the French government. This tube is used in the prototype of an oscilloscope referred to in this article (fig. 6) and also in a series of special-purpose oscilloscopes made by Société F.E.R.I.S.O.L. for the Commissariat à l'Energie Atomique. Total length of tube 48 cm.

second approach never permits an increase in sensitivity.

A helical deflection line like the one discussed is used in the cathode-ray tubes 42 D 13 BE and 50 D 13 BE. The latter tube is shown in *fig. 2*, and the deflection line of the 42 D 13 BE is pictured in *fig. 3*.

We have mentioned earlier the influence of the ancillary circuits on the concept of the oscilloscope as a whole. One ancillary device whose influence is quite marked is the delay line which has to be inserted in the signal channel between the trigger circuit and the tube input to compensate for the delay introduced by the time-base circuits. This line should have a delay time of several tens of nanoseconds, i.e. a length of several metres. Even if the very best coaxial cable is used, its losses cannot be neglected when the bandwidth is of the order of 5 GHz. Taking these losses into account a well-defined transit time per turn is necessary for the deflection line of the cathode-ray tube in order to optimize the pulse response of the equipment. This shows how impossible it is to separate the study of the cathode-ray tube from the study of the circuits associated with it, if the best performance is to be achieved.

Problems connected with the associated circuits

The principal circuits that have to be associated with a cathode-ray tube to form a high-speed oscilloscope for real-time use are:

1) In the signal channel:

A triggering coupler which serves to supply the synchronizing circuits with timing information related to the incident signal.

A delay line whose function has been explained earlier.

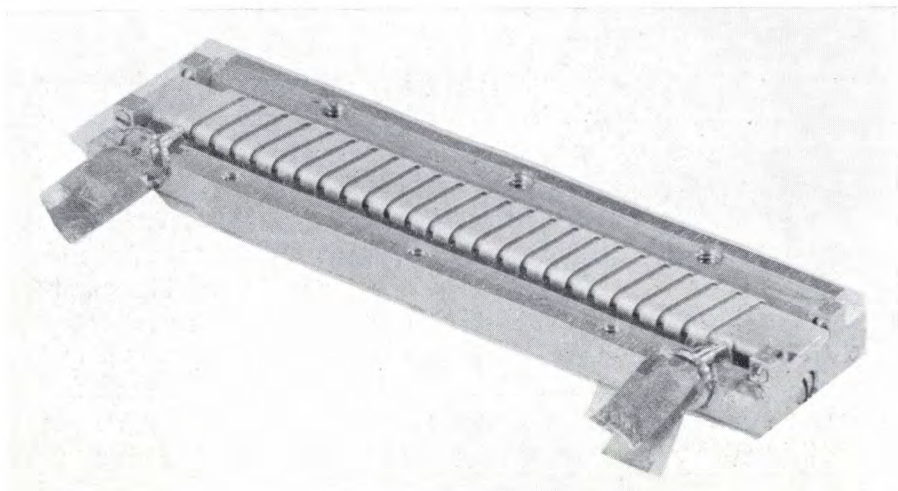


Fig. 3. Helical deflection line for the 42 D 13 BE tube.

2) Outside the signal channel:

A time-base circuit producing a signal for high-speed horizontal sweep.

An unblanking circuit which provides a pulse that switches the beam on for a time equal to that of the useful part of the sweep.

Synchronizing and trigger circuits for obtaining exact synchronization between signal, unblanking pulse and horizontal sweep.

The triggering coupler has to tap off a fraction of the energy of the incident signal, but it is important that it should not distort the signal, however fast it is. This condition can only be satisfied by a resistive type of coupling whose reponse is practically independent of frequency. Moreover, the amount of energy tapped off must be small enough not to introduce an appreciable reflection of the signal. A wide-band coupler designed at LEP is shown in *fig. 4*.

We mentioned earlier that the losses of the delay line have a marked effect on the response of the oscilloscope. These losses cannot be made negligible by increasing the cross-sectional dimensions of the cable, because it would then become too bulky. For cable of the highest quality the attenuation is mainly due to skin-effect losses and therefore varies as the square root of the frequency. Since the attenuation is of course proportional to the length of the cable, it follows that for a given attenuation the intrinsic bandwidth of the cable is inversely proportional to the square of its length. In the same way, the effect of the cable on the rise time of a pulse signal will be proportional to the square of its length. It can thus be seen that it is vitally important to keep the length of the delay line as small as possible, and consequently to reduce the triggering delay of all the circuits linked with the tube.

The time-base speed should be directly related to the bandwidth of the oscilloscope: for example, the accurate determination of a 100-picosecond leading edge requires a time-base speed of the order of 2 cm per nanosecond. This means that the time-base signal should have a slope of up to 150 V per nanosecond and a linear range of about 200 V. Classical time-base circuits are not capable of giving this kind of performance and new techniques have to be brought into play. A method used at LEP consists in forming a high-voltage step with as steep a leading edge as possible, and then shaping this edge by means of a series of filters which permit different time-base speeds to be obtained. The voltage step can be generated by ultra-high-speed switching circuits using avalanche transistors. This method allows extremely high slopes to be obtained at sufficient amplitude. A time-base generator based on this principle is shown in *fig. 5*.

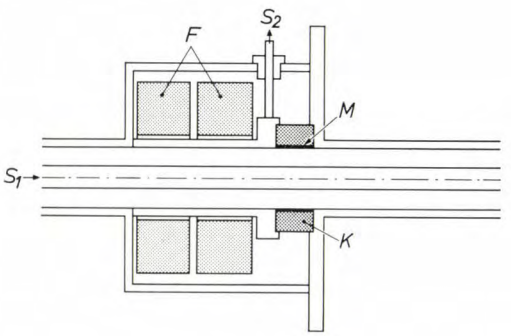


Fig. 4. Schematic diagram of a wide-band coupler designed at LEP. The metal film M deposited inside the ceramic bushing K forms a 2.5 ohm resistance R_s in the outer conductor of the coaxial line carrying the signal S_1 to be displayed. A fraction of S_1 appears across this resistance and is used as a synchronizing signal S_2. The ferrite toroids F prevent the resistance R_s from being short-circuited by the walls of the coupling cavity.

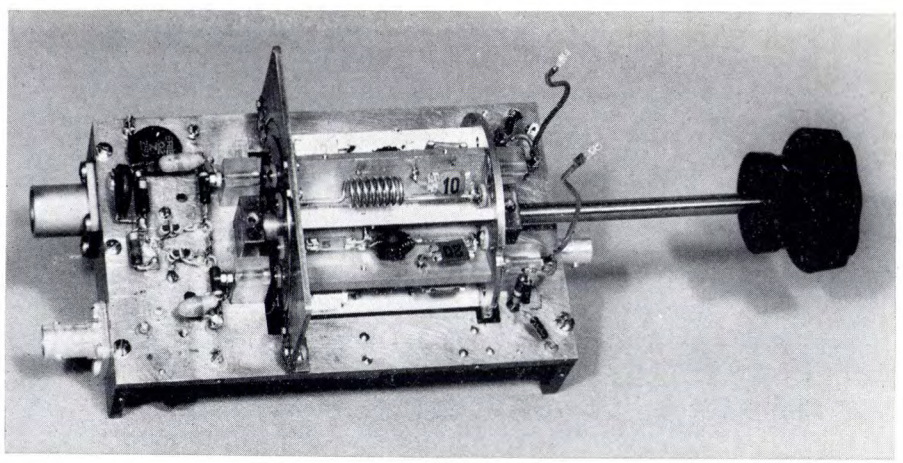


Fig. 5. Time-base generator with choice of four speeds.

The introduction of the filter increases the delay of the time-base signal, and this delay becomes more pronounced as the desired time-base speed is reduced. Now we saw earlier that this delay should be kept within reasonable limits to avoid limiting the bandwidth of the oscilloscope. This explains why only high time-base speeds can be obtained in this way: unless oscilloscope should have, at the highest speed, an amplitude of about 150 volts, and a flat top of about 1.5 nanoseconds duration with leading and trailing edges of the order of 0.5 nanosecond. Special precautions have to be taken in transmitting such a pulse to the control grid of the cathode-ray tube: in fact, in the region of the maximum writing speed, any accidental

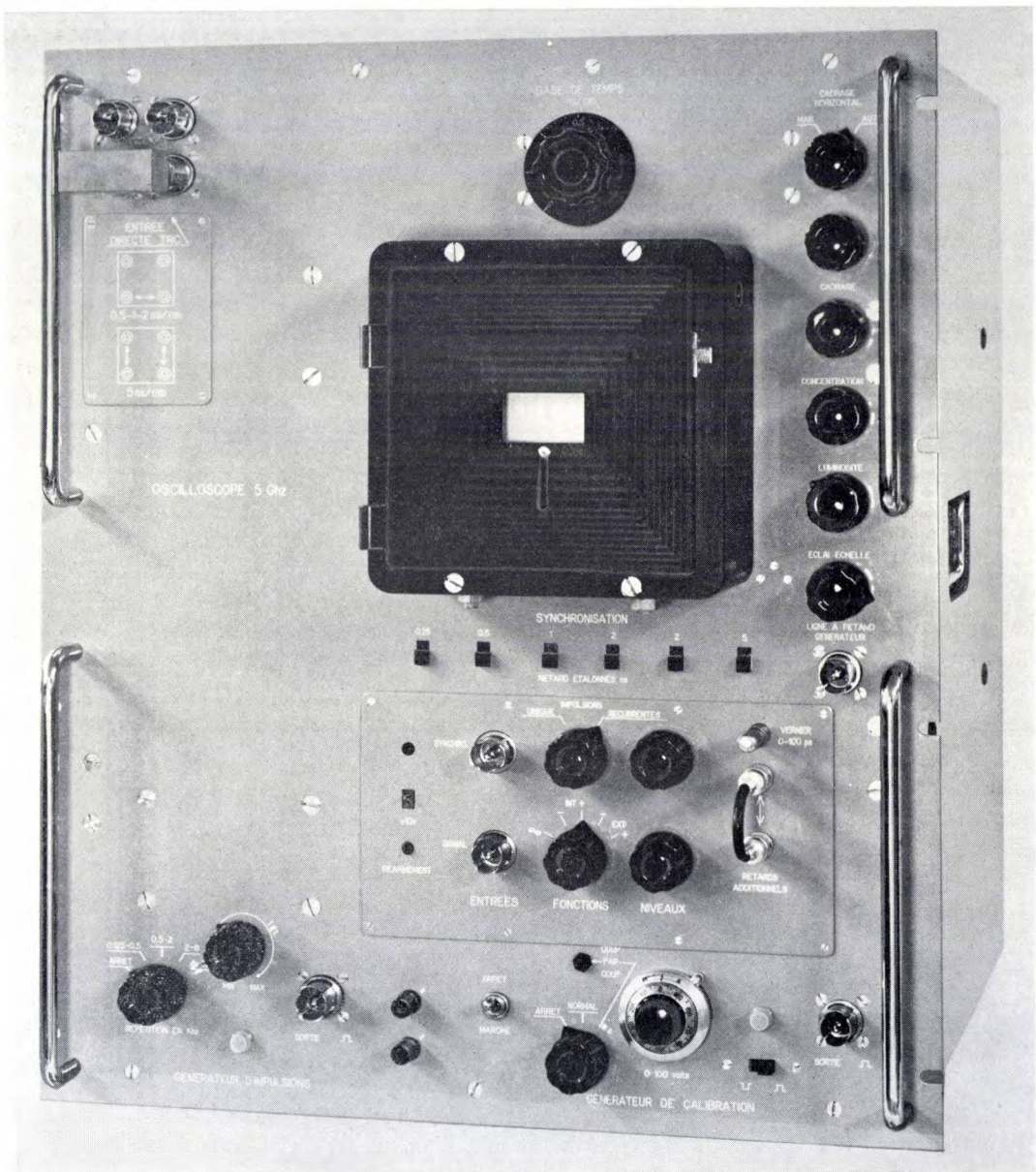


Fig. 6. Experimental oscilloscope with a bandwidth of 5 GHz.

extra circuits of classical design are included to give low time-base speeds, a high-speed real-time oscilloscope only has a limited range of speeds.

Associated with fast time-base speeds is the necessity of producing very short beam-switching pulses with a very steep leading edge. By way of example, the beam-switching pulse for the cathode-ray tube of a 5 GHz overshoot of the unblanking pulse results in an unacceptable modulation of the intensity of the trace. It is also important to avoid any coupling between the beam-switching circuit and the deflection circuits of the tube.

It is obvious that the synchronizing circuits should be capable of exceptionally accurate triggering. We

have spoken earlier of time-base speeds reaching 2 centimetres per nanosecond. On the other hand, the thickness of the trace on the screen is always very small, a typical value being 50 μm. If the maximum accuracy corresponding to this trace thickness is to be obtained, it can be seen that the maximum error in the triggering should be of the order of 2 picoseconds. Since there is a limit to the repetition rate of the time-base circuits, it is also necessary to introduce scaling circuits if it is desired to examine signals with very high repetition frequencies (several GHz). These various requirements, including the short delay time specified, result in synchronizing or triggering circuits that are much more complex than in classical oscilloscopes.

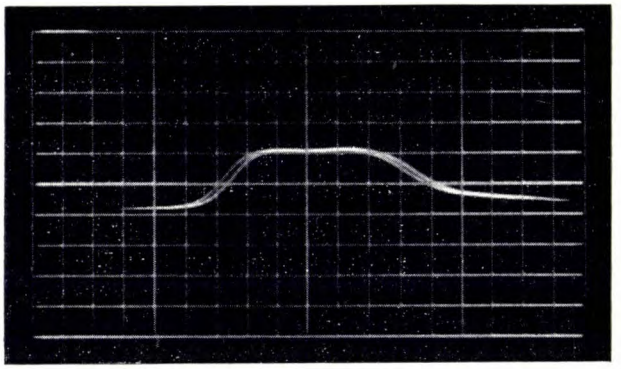


Fig. 7. Photograph of two closely spaced pulses (time difference 25 ps) as displayed on a 5 GHz oscilloscope at a time-base speed of 2 cm/ns. The rise time of each pulse was 100 ps, the amplitude 12 V. The photograph is enlarged about 2×.

Conclusions

We have briefly discussed the varied problems encountered in the design of a high-speed oscilloscope for real-time operation, and described the close interplay between them. The solution of these problems requires the existence and close cooperation of a team whose expertise ranged over fields as widely separated as vacuum physics, electron optics, and thermionic-valve techniques on one hand, and ultra-high-speed pulse techniques and microwave circuits on the other. The fact that LEP was able to build up a close-knit team with these capabilities has led to the production today of a prototype 5 GHz oscilloscope that has no equivalent anywhere else on the world market, and has already been found extremely useful at these Laboratories. This oscilloscope is shown in *fig. 6*. An oscillogram of two closely-spaced pulses recorded by the equipment is reproduced in *fig. 7*.

With the progress made in the last few years the real-time oscilloscope has become a measuring instrument which is indispensible in the study of the many transitory phenomena which are found in the various branches of modern science.

Summary. Since real-time amplification is not yet practical for signals of less than 1 ns duration, these signals can only be displayed by applying them directly to the cathode-ray tube of the oscilloscope. The cathode-ray tube must at the same time have large bandwidth, high sensitivity, high definition and a maximum writing speed high enough to permit the display of a very fast non-recurrent signal. The high-speed time-base circuit and the synchronization and unblanking circuits raise problems which are not encountered in conventional high-speed oscilloscopes and require new solutions.

Method for growing single crystals of cuprous chloride

J.-J. Brissot and A. Lemogne

Cuprous chloride CuCl in its cubic modification shows a strong Pockels effect, and for this reason is an interesting material for making electro-optical modulators in a spectral region extending from 0.4 to 20.5 microns.

It solidifies at 422 °C, but at this temperature a hexagonal wurtzite structure is formed, which changes into the cubic blende structure at 407 °C [1]. Since the phase transition takes place in the solid state, it is extremely difficult to obtain a cubic single crystal of useful dimensions from a pure melt [2]. It is therefore necessary to

J.-J. Brissot, Ingénieur E.N.S.C.P., L. ès Sc., is with Laboratoires d'Electronique et de Physique Appliquée, Limeil-Brévannes (Val-de-Marne), France; A. Lemogne, D. ès Sc., was formerly with these Laboratories.

[1] M. R. Lorenz and J. S. Prener, Acta cryst. **9**, 538, 1956.
[2] F. Sterzer, D. Blattner and S. Miniter, J. Opt. Soc. Amer. **54**, 62, 1964.

prepare this crystalline form at a temperature below 407 °C.

Since the material should have high optical transparency, it is essential to obtain it in the purest possible form. Commercial cuprous chloride, usually supplied in powder form, contains cupric ions. These are removed by dissolving the material in concentrated pure hydrochloric acid and selectively precipitating CuCl by dilution with de-ionized water, in which it is relatively insoluble. The white precipitate obtained is washed in acetone and then dried in a vacuum at room temperature. The product is then subjected to fractionated sublimation in a vacuum of $\leq 10^{-5}$ torr. After melting the condensate in an atmosphere of pure argon, a vitreous colourless mass is obtained which can straight away be used for preparing the single crystals.

Among the various methods of preparing single crystals at a temperature lower than their melting point, the flux method is of particular interest [3].

In the Bridgman method a crystal can be grown from a cuprous-chloride solution using a minimum flux concentration in order to keep the melting point lower than 407 °C.

We have adopted a somewhat different method which makes it possible to work at a considerably lower temperature, thus reducing the thermal stresses generated in the crystal.

The eutectic mixture CuCl(0.65 mol)-KCl(0.35 mol), whose melting point is 130 °C, is prepared and a small quantity is placed at the bottom of a tapered quartz tube. A cylindrical piece of polycrystalline CuCl is placed on top of the eutectic mixture and the tube is sealed at a pressure of 400-500 torr of pure argon. By means of a heating coil made from a few turns of resistance wire the eutectic-CuCl interface is heated to a temperature close to 200 °C; the eutectic mixture melts completely and dissolves a small quantity of CuCl. The quartz tube is then lowered into the coil at the rate of 0.2 mm/h, enabling the flux to diffuse freely at the level of the crystallization interface, while the molten solvent zone moves up the rod. A single crystal of CuCl of high optical transparency gradually forms from the tapered end of the tube (*fig. 1*). In the same way a single crystal seed with a particular crystallographic orientation can be placed at the bottom of the tube, with a eutectic pellet above it, and above that a polycrystalline rod of CuCl.

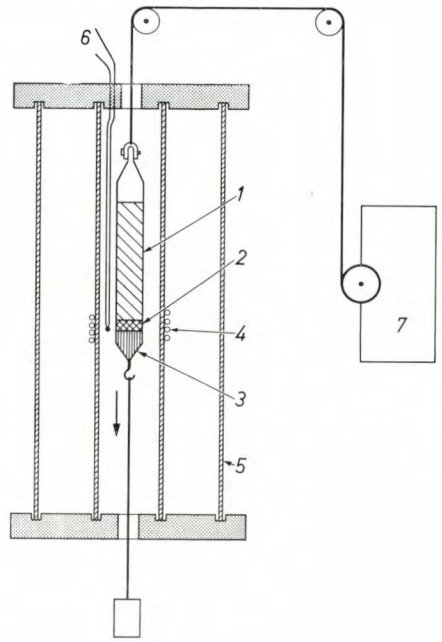


Fig. 1. Apparatus for growing single crystals of CuCl. *1* bar of polycrystalline cubic CuCl. *2* molten solvent zone. *3* growing single crystal. *4* heating coil. *5* heat-insulating wall. *6* thermocouple. *7* motor.

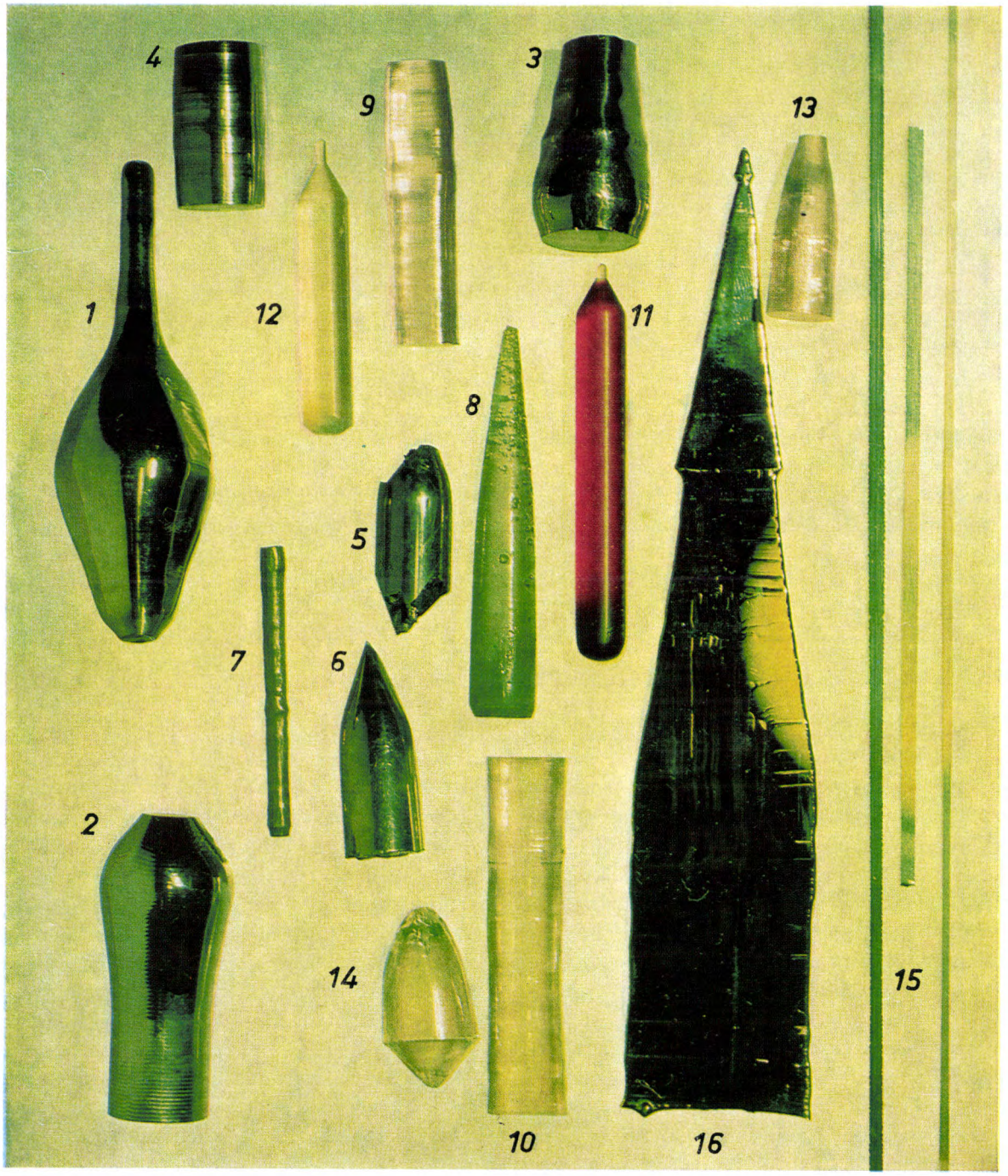
The application of this method, in which the molten solvent zone is displaced, enables single crystals of cubic CuCl to be prepared at low temperature and with low thermal gradients. Crystals of diameter greater than 10 mm can be prepared.

Since CuCl is very sensitive to moist air, the single crystals must be protected from it during use. Gallium arsenide crystals, which are chemically very stable, are therefore generally preferred at present for the modulation of infra-red radiation mentioned in the introduction.

Summary. CuCl has a transformation point below its melting point (422 °C). Single crystals of the cubic form, which is stable at low temperatures, can therefore be grown only at temperatures well below the melting point. This has been achieved by means of a flux method in which a molten solvent zone is displaced through a bar of polycrystalline CuCl.

[3] J. Rivera, L. A. Murray and P. A. Hoss, J. Crystal Growth **1**, 171, 1967; M. Soga, R. Imaizumi, Y. Kondo and T. Okabe, J. Electrochem. Soc. **114**, 388, 1967.

Single crystals for semiconductor devices and lasers



J.-J. Brissot and co-workers at the Laboratoires d'Electronique et de Physique Appliquée have investigated the preparation of single crystals by applying a variety of production processes: Czochralski's method (C), the floating-zone method (Z), the Bridgman method (B) and the travelling solvent zone method (S).

The photograph shows a number of specimens that have been grown; the largest crystal shown (No. 16) is 18 cm long.

1 Germanium (C)
2 Silicon (C)
3 Indium antimonide (C)
4 Indium arsenide (C)
5 Cadmium telluride (B)
6 Mercury telluride (B)
7 Gadolinium (Z)
8 Cuprous chloride, CuCl (S)
9 Neodymium-doped calcium tungstate (C)
10 Pure calcium tungstate (C)
11 Uranium-doped calcium fluoride (B)
12 Pure calcium fluoride (B)
13 Neodymium-doped yttrium aluminium garnet (C)
14 Pure yttrium aluminium garnet (C)
15 Three germanium dendrites (C modified)
16 Single-crystal strip of germanium (C)

Photoelectric solid-state devices and the perception of images in the infra-red

F. Desvignes, J. Revuz and R. Zeida

Introduction

It is generally realized nowadays that the observation of images in that part of the infra-red region of the spectrum in which objects at normal temperatures radiate strongly, namely from 3 to 30 microns, is of considerably practical interest. This region can be used for military applications, some of which are fairly obvious, as well as for numerous scientific, medical and technical purposes.

For many years now, this problem has been investigated at LEP in a number of different ways. The problem is very different from that of image perception in the ultra-violet, visible and even near-infra-red regions (up to 2 μm) so that it will be useful to briefly recall here its essential points. At ordinary temperatures, bodies emit approximately 150 watts per square metre and per steradian (or 10^{22} photons per m^2 sr s). Less than 2/10000 of this power — or fewer than 3 photons per million — is found in the wavelength region lower than 3.2 microns; if the threshold is raised to 5.5 microns, then one obtains 2% of the power and 0.2% of the photons, and with a threshold at 12 microns these figures become approximately 25% and 15%. These wavelengths have not been chosen arbitrarily: they correspond to the thresholds of the photoelectric materials indium arsenide (3.2 μm), indium antimonide (5.5 μm), mercury-doped germanium and cadmium-mercury telluride (both 12 μm). It should be recalled that the atmosphere has pronounced absorption bands between 2.6 and 2.8 μm, between 5.5 and 7 μm and beyond 14 μm.

Another important point is that in an almost isothermal enclosure, such as a closed room, the contrasts in the infra-red scene are very low. The relative emission difference between two bodies at different temperatures is at a maximum if they are "black bodies" and depends only on the difference between their respective temperatures. If the total emitted energy is considered, this relative difference is 1.3% for a temperature difference of 1 °C. When it is recalled that the minimum luminance difference perceptible to the eye (in the visible spectrum) is 1-2%, and realized that for the majority of applications it is desirable, for example, to perceive temperature differences of 0.1 °C, it will be appreciated that the instruments required must have a differential sensitivity which is 10 to 20 times better than that of the eye. There is a further difficulty caused by the fact that the image to be observed must be produced by an infra-red optical system whose relative aperture and transparency are rather limited and that, unless special precautions are taken, the low-contrast image provided by the objective will be swamped by the ambient radiation. The precautions in question consist in protecting the photoelectric body by non-radiating (i.e. absorbing and cooled) screens.

The most natural course — which is also the most ambitious — is to try to use in the infra-red the solutions which have proved best in the visible region of the spectrum, i.e. to produce vidicons; successful attempts along these lines are described in another article in this issue [1]. So far these attempts have led only to the production of tubes whose photoelectric threshold does not exceed 2.5 μm and which are not capable of detecting bodies whose temperature is lower than 400 °K.

Theory and practice show in fact that with a normal television standard the target must supply 10^{10} photoelectrons per second or 1 nA to give a detectable signal, and 10^{12} photoelectrons per second or 0.1 μA to give an acceptable picture with 40 or 50 levels of contrast. With a quantum efficiency equal to 1 and an optical system whose aperture is equal to *f*/2 (cf. [2]), calculation shows that this last condition can only be satisfied for a scene at normal temperatures if the threshold is beyond 3 microns.

On the other hand, there is a drawback in choosing a threshold which is too high, for the signal is then so large that there is a risk of it exceeding the capacity of the target to store information (maximum current: several tenths of a microampere). If the threshold is at 5.5 μm, the radiation at ordinary temperatures contains 500 times more useful photons than at 3.0 μm. In order to prevent this excess it is necessary to use a cold diaphragm with an aperture equal to *f*/40, or to use a target whose quantum efficiency is only 2/1000.

These things are not impossible to achieve but technical difficulties arise when it is necessary to detect very

F. Desvignes, Ingénieur E.S.O., Ingénieur C.N.A.M., and J. Revuz, L. ès Sc., are with Laboratoires d'Electronique et de Physique Appliquée, Limeil-Brévannes (Val-de-Marne), France. F. Desvignes is also a Professor at the Ecole Supérieure d'Optique. R. Zeida, D. ès Sc. Appl., formerly with these Laboratories, is now with the Faculdad de Ciencias, Universidad de Buenos Aires.

weak contrasts corresponding to temperature differences of several tenths of a degree in the vicinity of 20 °C. It can be shown that these difficulties are reduced if thresholds lower than 5.5 μm are used, because the contrast is then slightly better.

With these considerations in mind we undertook a detailed study of the technology of *P-N* junctions, first of InSb (threshold about 5.5 μm) and then of InAs (threshold about 3.2 μm); our aim was not to produce photodiodes but camera tubes of the type described in reference[1], with a target consisting of a mosaic of *P-N* junctions. As will be seen, the results obtained are perhaps too good for the production of photodiodes but not quite good enough for the targets of camera tubes. An alternative and well-known method of observing infra-red images uses a single detector with a small sensitive surface and a high response speed. By sweeping the two-dimensional infra-red image past this detector in two directions in accordance with a certain scanning code by opto-mechanical means, the image is analysed dot by dot. Systems of this type are described in this issue [3] and possess the great advantage that they completely eliminate those faults in the final image which are due to local variations in the target sensitivity. However, they are restricted in their space- and time-resolving power by the high speed which they impose on one of the two mechanical movements required for scanning.

We now believe that a solution halfway between the two preceding methods may prevail in the next few years. This solution consists of using a linear array of single-element detectors (a "cell-strip") in a system with only one mechanical scanning movement. Its advantages over the two methods described above are threefold, namely, the sole remaining mechanical movement can be relatively slow, the problem of storing the photoelectric information in the target no longer occurs in certain cases, and local sensitivity variations can be compensated in a relatively simple way. The "cell-strip" of very high quality indium-antimonide photodiodes, developed and produced by LEP, is, like the InAs and InSb photodiodes themselves, a by-product of the work done during research on television camera tubes.

While it is of advantage in the case of television camera tubes to use a photoelectric target whose threshold is not too far into the infra-red (situated at about 5 μm), is it preferable with single-cell detector systems to have a detector whose threshold is farther in the infra-red since it is better to receive as large a signal as possible. The designers have still not decided whether to choose detectors which are sensitive up to 12 μm or up to 5.5 μm, the relatively modest benefit given by the first (minimum detectable temperature difference 3 to 4 times smaller) being offset by more complicated cryogenic problems and therefore increased cost. In any case, the lower the temperatures of the scenes to be observed, the greater the importance of the larger wavelengths.

Conditions to be satisfied for the target of the camera tube

The charge-storage effect used in television camera tubes whose target consists of a mosaic of *P-N* junctions requires these junctions to possess very definite properties (see article [1]). These properties are the following:

a) very low dark current,

b) break-down voltage as high as possible,

c) high capacitance per unit area.

The theory of the *P-N* junctions, borne out by experiments made here, underlines the importance of certain electronic and geometric parameters, affecting the production of junctions with the required properties. Thus:

1) The dark current can be considered as consisting of two terms: one depending on bulk effects in the semiconductor, the other on the surface properties. The current due to bulk effects will be small if the lifetime of the minority carriers is high, if the thickness of the depletion layer is small and if the concentrations of the dopants are sufficiently high. The contribution due to surface effects will be relatively small if the number of impurity ions and atoms per unit area that act as recombination centres is small. Moreover, the crystal quality of the semiconductor can be responsible for a high dark current; in fact, defects such as dislocations, grain boundaries, vacancies, etc. are the origin of local break-downs in the neighbourhood of a junction.

2) A high break-down voltage and a thick depletion layer are related to a low concentration of impurities in the most lightly doped zone. Indeed, the electric field existing there must not exceed a value which is characteristic of each semiconductor material and for a given break-down voltage the field will be less for greater thickness.

3) High capacitances permit the storage of considerable charges for a given voltage change. The capacitance of the junction is linked to the thickness d of the depletion layer by the equation:

$$C \propto d^{-1/n},$$

in which n equals 2 or 3 depending on the type of junction (abrupt or gradual).

---

[1] M. Berth and J.-J. Brissot, Targets for infra-red television camera tubes, page 270 of this issue.

[2] C. Hily, Objective lenses for infra-red image formation, page 290 of this issue.

[3] M. Jatteau, Infra-red thermography equipment for medical applications, page 278 of this issue.

4) The photoelectric sensitivity of the junction is of less importance in the problem which concerns us, since it is possible to make do with very low values if the photoelectric threshold is situated appreciably beyond 3 μm. Thus the detection qualities of the junction will not be a vital point in a pilot study.

When all these points are taken into account it is seen that the above important parameters must satisfy conditions which are sometimes contradictory. In particular, the concentration of impurities in the semiconductor will have to be sufficiently high to increase the capacitance value and decrease the intensity of the dark current without, however, reducing the breakdown voltage of the junction to excessively low values or influencing too strongly the life-time of the minority carriers. The choice of the parameters mentioned with III-V compounds is further restricted at the present time because it is not possible to obtain sufficiently pure crystals (indium antimonide: minimum concentration of majority carriers 10^{14} cm^{-3}; indium arsenide: 2×10^{16} cm^{-3}).

Technological problems raised by the making of InSb and InAs junctions

Numerous technological problems are encountered at various stages when making *P-N* junctions of III-V compounds. They start with the sawing, cutting and polishing operations. InSb and InAs are very fragile materials, so it was necessary to develop special mechanical methods to avoid disturbing the crystal structure of the semiconductor excessively. The various cleaning and etching processes which follow the mechanical operations demand particular care; chemical agents were chosen after numerous tests designed to find the best results. In spite of all the precautions taken, additional cleaning is necessary as a final operation for InAs in order to obtain the surface condition which is indispensable for the success of subsequent operations; a mechanical polishing operation with chemically active substances has been found suitable.

The diffusion process used to obtain the two regions of different conductivity on either side of the junction must prevent both the loss by diffusion of the most volatile component and the introduction of impurities into interstitial sites in the crystal structure of the atom.

These problems are not completely understood but lengthy experimentation has made it possible to apply the method of diffusion from the vapour phase using the elements zinc and cadmium and to define operational conditions such that junctions made from the diffused wafers have optimum characteristics.

The ohmic contacts with the *N* and *P* regions are made using different alloys one of whose constituents should preferably be indium. Special precautions are taken

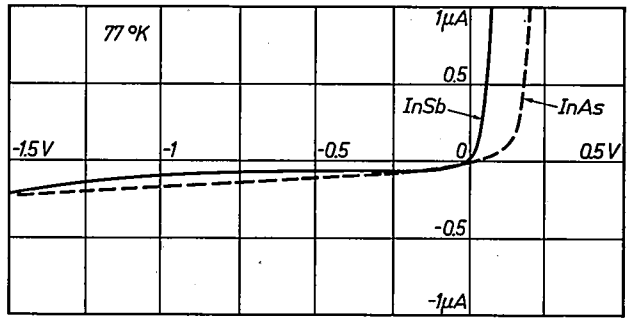


Fig. 1. Current-voltage characteristic of a *P-N* junction in InSb at a temperature of 77 °K. The dashed curve applies to InAs at this temperature.

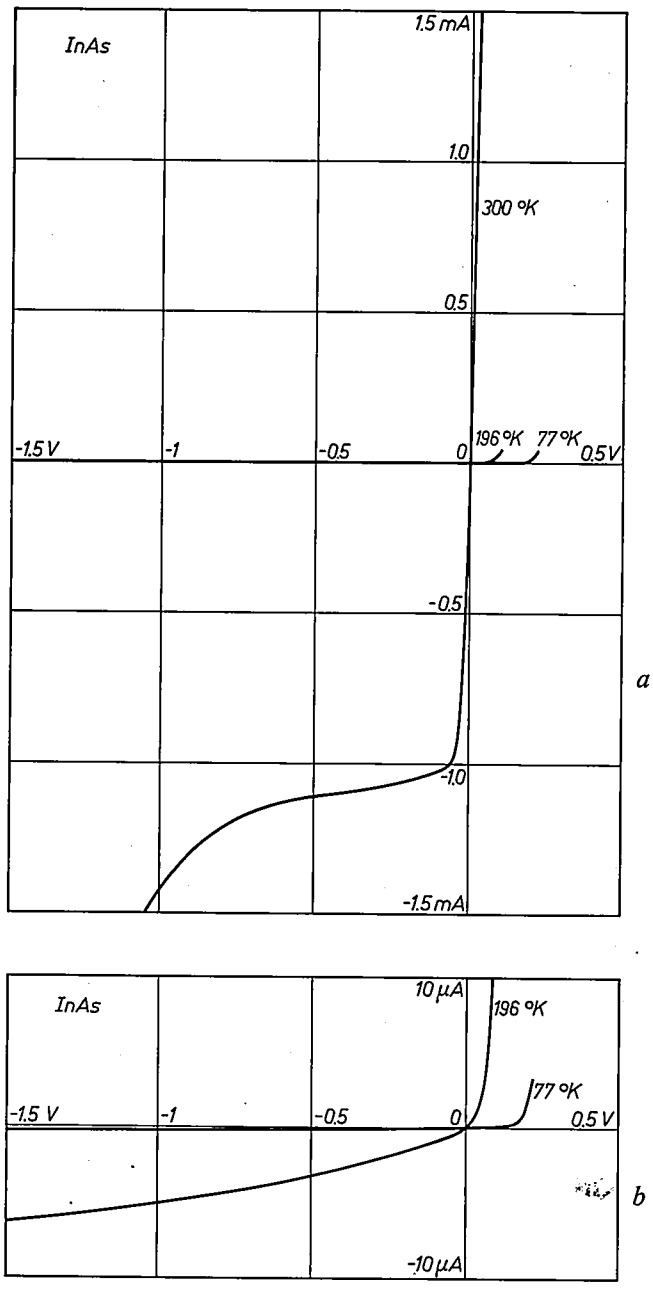


Fig. 2. *a*) Current-voltage characteristics of a *P-N* junction in InAs at three different temperatures. *b*) The curves for 77 °K and 196 °K on a scale more similar to the one of fig. 1.

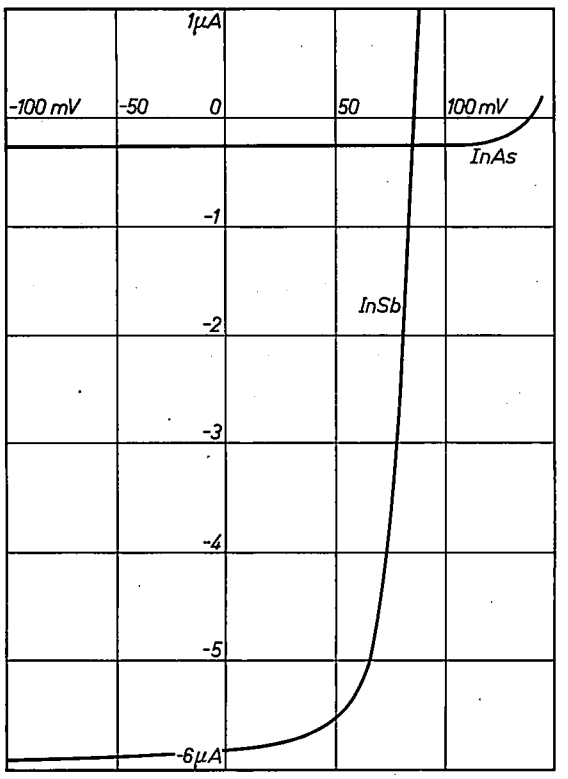


Fig. 3. Photoelectric characteristics of a *P-N* junction in InSb and a *P-N* junction in InAs, both at a temperature of 77 °K, for irradiation by a black body at 300 °K (half-aperture angle of the cone of radiation 45°).

to ensure that contact with the diffused layer is sharply localized. This means that only the area of the joint must be submitted to the necessary heating.

In order to limit the junction area to the required value and to remove (by chemical means) most of the surface defects of the crystal in the region where the junction reaches the free surface, an etching operation is needed. The process has been worked out in detail and leads to almost complete elimination of the contaminated zone.

Since the junction in this state is particularly sensitive to the environment, it must be immediately inserted in an air-tight enclosure serving also as a cryostat. This set-up is necessary for the use of detectors which have to operate at temperatures very much lower than ambient. To facilitate the diode assembly operations and also to improve the diode's reliability, surface-passivation research, chiefly concerned with InAs junctions, was done at LEP and enabled a suitable method to be adopted.

Characteristics obtained

Indium antimonide and indium arsenide behave differently in one important respect: while indium-arsenide *P-N* junctions can be used at ambient temperatures, indium-antimonide *P-N* junctions can only operate at temperatures below —150 °C. Consequently, the results reported below refer to InSb junctions cooled to 77 °K and to InAs junctions kept at the following three temperatures: 300 °K, 196 °K and 77 °K.

The change of the current-voltage characteristics of the InAs *P-N* junctions with temperature has clearly shown the importance of etching and of passivation of the junction, and also has given an indication of the kind of residual vapours present in the cryostat enclosure.

The current-voltage curve of an InSb junction is shown in *fig. 1*, while *fig. 2* shows three curves for an InAs junction. The maximum reverse voltage (breakdown voltage) is of the order of 4 to 5 volts in all four cases. For a given range of voltages corresponding to reverse-polarization operation of the diodes, the capacitance is inversely proportional to the cube root of the voltage applied. Typical capacitance values per unit area at zero voltage are 7 nF/cm² for InSb diodes and 12 nF/cm² for InAs junctions.

The photoelectric current-voltage curves obtained when the photodiode was held at 77 °K and exposed to the radiation of a black body at 300 °K are given in *fig. 3* for InSb and for InAs. The spectral sensitivity curves for InSb and InAs junctions (the latter at different temperatures) are reproduced in *fig. 4*.

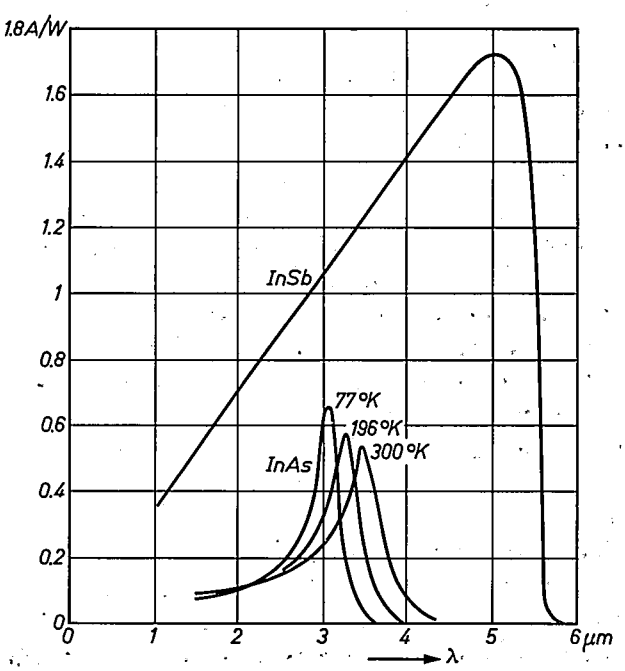


Fig. 4. Spectral response curves of a *P-N* junction in InSb at 77 °K and in InAs at 77 °K, 196 °K and 300 °K. The photocurrent per unit of power received is plotted in amperes per watt.

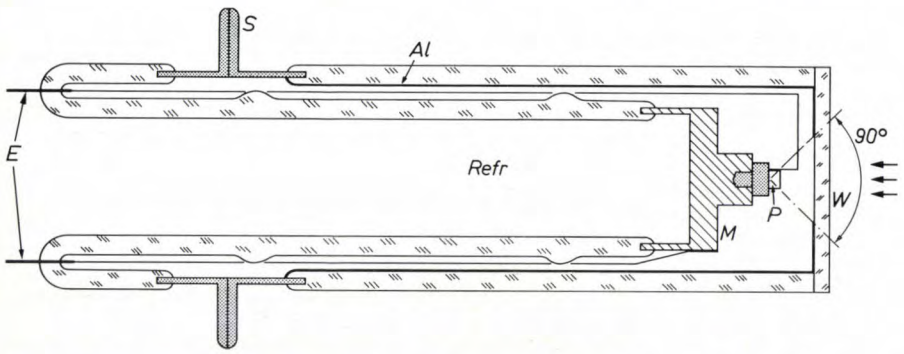


Fig. 5. A photodiode (*P*) of InAs or InSb is mounted in a vacuum-tight enclosure that also functions as a cryostat. The photodiode is mounted in a molybdenum holder *M* whose rear surface is in direct contact with the cooling fluid *Refr*. The radiation to be detected reaches the photodiode through a sapphire window *W*. *Al* aluminium coating, functioning as a radiation shield. *S* soldering flanges of Dilver P. *E* external connections for the photodiode.

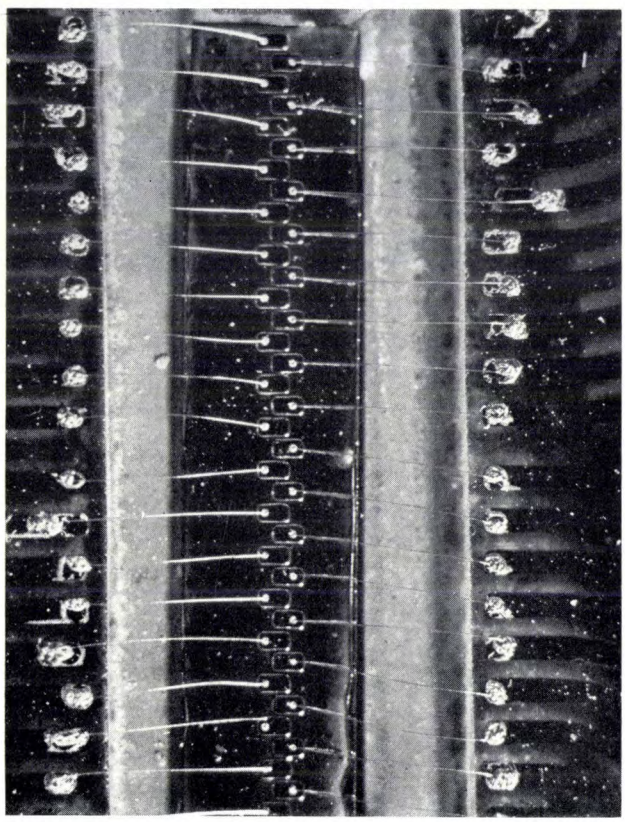


Fig. 6. Linear array of 40 *P-N* diffused photodiodes made with a mesa technology on a single crystal of indium antimonide (a single wafer 20 mm long). Each mesa is 400×650 μm, the distance between centres is 500 μm. The responsivity of individual photodiodes differs less than 15% from the mean value.

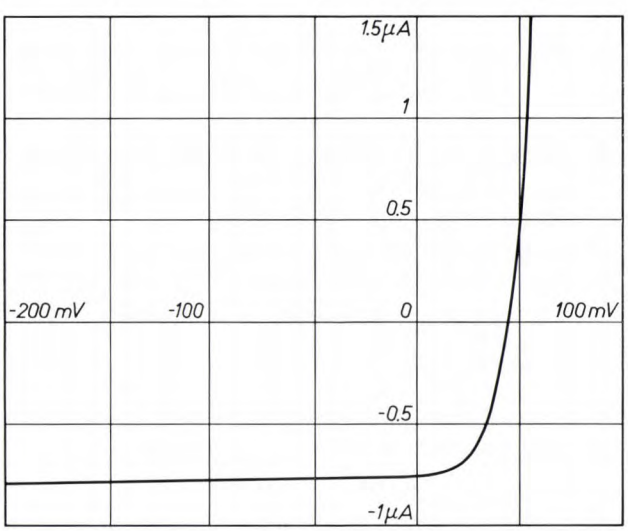


Fig. 7. Photoelectric characteristic of one of the photodiodes of the strip, at a temperature of 77 °K, when irradiated by the surroundings at 300 °K.

The specific detectivity was calculated from measurements performed using radiation from a black body at 500 °K modulated at 500 Hz. The following results were obtained:

45×10^8 W^{-1} cm Hz$^{1/2}$ for InSb (77 °K),
8×10^8 W^{-1} cm Hz$^{1/2}$ for InAs (300 °K),
50×10^8 W^{-1} cm Hz$^{1/2}$ for InAs (77 °K).

Fig. 5 shows a section through the photodetector assembly.

Linear arrays of photodiodes (cell-strips)

A few remarks should be made about the design of the linear arrays of photodiodes mentioned in the foregoing. In one case, which is presented here as an example, a row of 40 *P-N* junctions is spread out over a distance of 20 mm (*fig. 6*) and each of the junctions is connected individually to an external circuit. The individual photodiodes have a sensitive area of 0.4×0.4 mm and the mean current-voltage characteristic with ambient illumination is shown in *fig. 7*. The envelope, of glass and metal, includes an internal cooling device and a crown of 40 insulating passages as shown in *fig. 8*. *Fig. 9* is a photograph of the device shown diagrammatically in fig. 8. Under the window can be seen the cold diaphragm with an oval aperture and behind this is mounted the cell-strip with 40 indium antimonide photodiodes; the electrodes are visible around the circumference; on the left, the tube in which a Joule-Thomson expansion valve is inserted to produce liquid nitrogen from compressed nitrogen at ambient temperature.

If these *P-N* junctions are illuminated by a body at room temperature and the short-circuit current is measured, differences not exceeding 15% of the mean current value will be observed between the diodes.

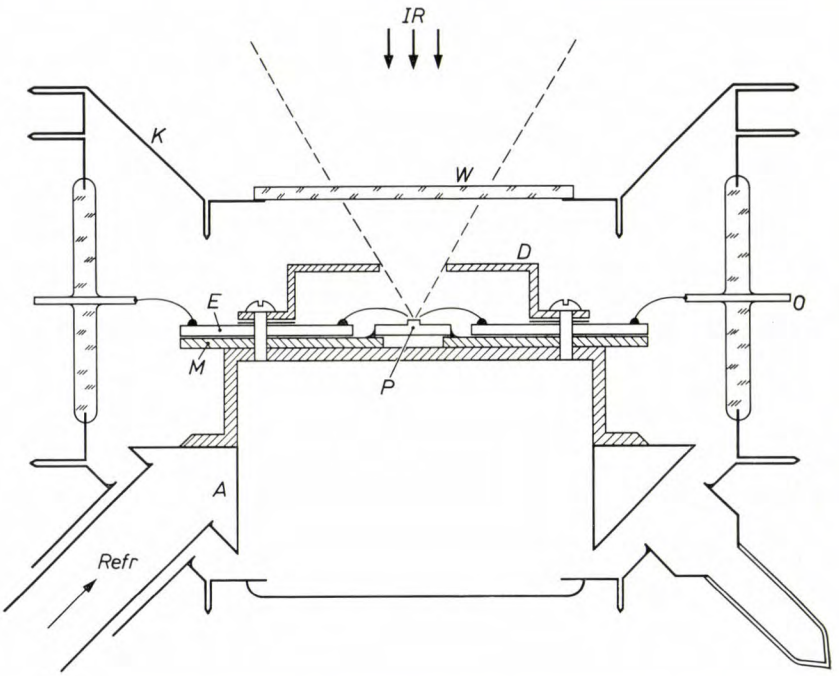


Fig. 8. Cross-section of a pick-up tube with a linear array of InSb photodiodes. The infra-red radiation *IR* arrives at the strip *P* via a sapphire window *W*. *M* molybdenum sheet, which carries the photodiode strip and is in thermal contact with the chamber *A*, cooled by liquid nitrogen (*Refr*). The window is mounted in a cup *K* of Dilver P. *D* cooled diaphragm. *E* disc carrying the 40 connections for the 40 photodiodes in the strip. *O* one of the output electrodes.

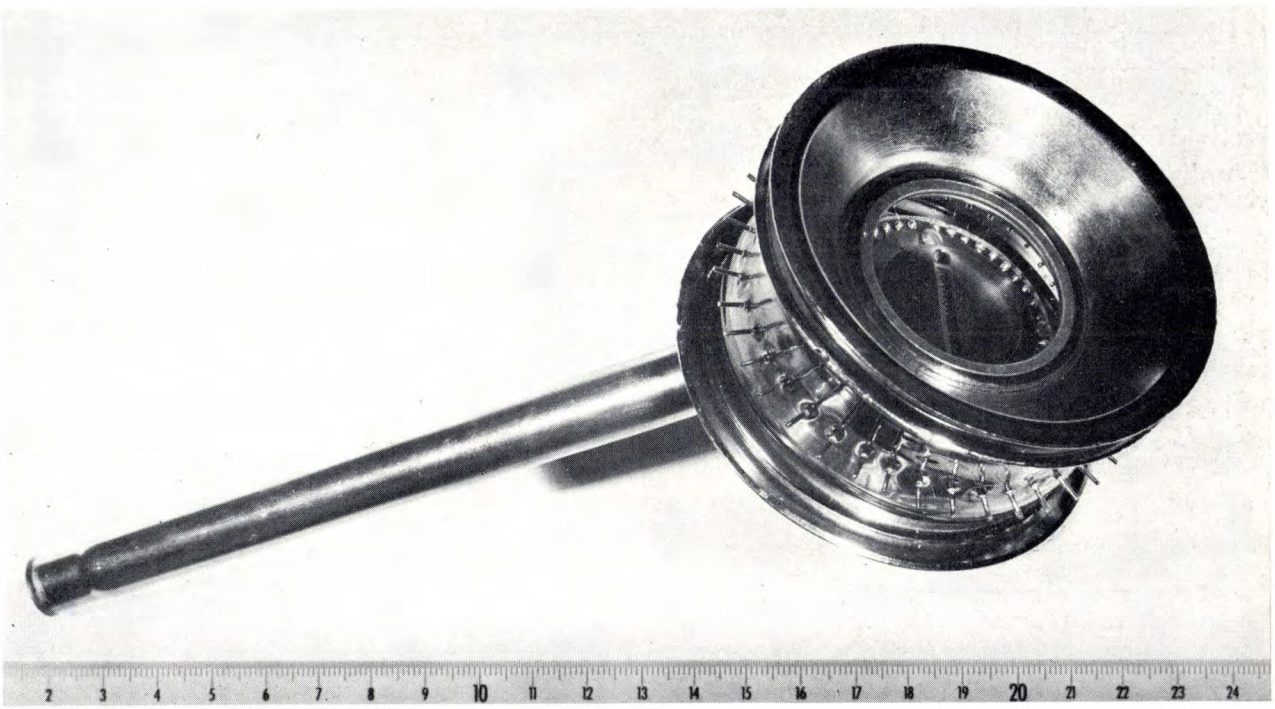


Fig. 9. The pick-up tube. The ring of 40 connectors for the 40 photodiodes of the strip can be seen through the window, and so can the cooled diaphragm, with the strip behind it.

Summary. Pick-up tubes are available nowadays for optical images formed by visible or ultra-violet radiation, but there is no really satisfactory device for scanning images formed by thermal radiation at ordinary temperatures. The various aspects of the problem and the requirements which it imposes are examined. The article describes the work which has been done in attempting to solve the problem by using the photoelectric properties of junctions made of monocrystalline indium antimonide and indium arsenide; a brief description is given of the techniques developed and the results obtained. A method is discussed in detail in which a linear array of a number of InSb photodiodes (say 40) is used. In this case mechanical scanning in only one direction is required; one version of a pick-up tube based on this principle and made in the LEP is described.

Targets for infra-red television camera tubes

M. Berth and J.-J. Brissot

Introduction

Television camera tubes employing the principle of charge storage have nowadays completely superseded opto-mechanical scanning devices in the visible part of the spectrum. Their use in the infra-red however is a more difficult matter.

To clearly illustrate these difficulties, we must first recall the principles underlying the construction and operation of such camera tubes. We shall take as our example a vidicon-type tube with a photoconductive target as shown schematically in *fig. 1*. The target consists of a thin photosensitive layer whose front face, the one which receives the light, is covered with a transparent film which is a good conductor of electricity. The back face is scanned with a low-velocity beam of electrons which is deflected and focused by electric and magnetic fields. On striking this face, the electron beam brings it to the potential $-V_b$ of the cathode (the front face being at zero potential), and the sensitive layer, of thickness d and dielectric constant ε, stores a charge of

$$Q = CV_b = \frac{\varepsilon}{d} V_b$$

per unit surface area. Between two scans by the electron beam, the charge leaks away via the transverse resistance of the layer whose value per unit area is:

$$R = d/\sigma,$$

where σ is the conductivity of the material, which increases in proportion to the incident light level.

For the tube to operate it is essential that in darkness the time constant RC of the discharge is large in relation to the period τ of the scanning cycle. This condition is written:

$$\varepsilon/\sigma \gg \tau.$$

Assuming that the material has a dielectric constant which is equal to that of free space (8.85×10^{-12} Fm^{-1}) and that the scanning cycle takes $\tau = 4\times10^{-2}$ second (frame frequency 25 Hz), it will be seen that σ must have a value less than 2.2×10^{-10} $(\Omega m)^{-1}$. A value as small as this can only be obtained with semiconductors. In these materials the conductivity is related to the effective concentration $\bar{N}$ of the electrons in the permitted bands, the mobility μ of the free charge carriers and the energy gap ΔE by the equation:

M. Berth, Ingénieur C.N.A.M., and J.-J. Brissot, Ingénieur E.N.S.C.P., L. ès Sc., are with Laboratoires d'Electronique et de Physique Appliquée, Limeil-Brévannes (Val-de-Marne), France.

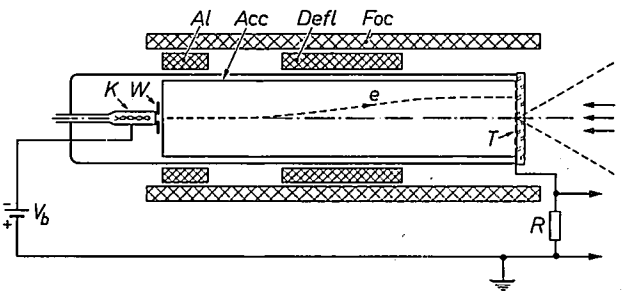


Fig. 1. Schematic diagram of a vidicon. T image electrode on which the optical image is projected. K cathode. W Wehnelt cylinder. e scanning electron beam. *Acc* accelerating electrode. *Al* alignment coil, *Defl* deflection coil, *Foc* focusing and collimating coil. The current from the signal anode of the image electrode flows through the resistance R, and the signal voltage obtained across R is amplified and processed further.

$$\sigma = q\mu\bar{N} \exp\left(\frac{\Delta E}{2kT}\right),$$

where q is the electron charge, k the Boltzmann constant and T the absolute temperature. The parameters μ and $\bar{N}$ do not differ greatly from one material to another, so that to obtain a conductivity lower than a certain value it will be necessary to choose either a fairly high value for ΔE or a fairly low temperature.

Now the wavelength λ_0 of the photoconductivity threshold is related to the energy gap ΔE by the expression:

$$\lambda_0 = hc/\Delta E,$$

where h is Planck's constant and c the velocity of light. A threshold far into the infra-red therefore requires a small energy gap ΔE. In addition, in order to obtain a suitable discharge time constant, a maximum target temperature T is required such that:

$$\lambda_0 T \ll \frac{hc}{2k}\left[\ln\frac{q\tau\mu\bar{N}}{\varepsilon}\right]^{-1}.$$

Numerical calculation using normal values of μ and $\bar{N}$ ($\mu = 10^{-2}$ $m^2V^{-1}s^{-1}$ and $\bar{N} = 2\times10^{25}$ m^{-3}) yields:

$$\lambda_0 T \ll 221 \ \mu m\,°K.$$

This equation only gives a general idea of the form the problem takes: in fact the presence of impurities always leads to an increase in conductivity, so that the target has to be operated at an even lower temperature, and

on the other hand, layers of a microcrystalline or powdered nature result in very low apparent conductivities but at the cost of undesirable side effects (unfavourable rise times, fatigue and hysteresis).

An attempt to overcome the difficulties encountered can be made by going back to a microscopically homogeneous photoconductive substance but in the form of a *P-N* junction mosaic produced on a monocrystalline substrate. The charge-storage effect in such junctions has been studied in detail [1][2], and we shall merely outline its principles.

Under reverse bias a semiconductor *P-N* junction can be represented by the equivalent diagram shown in *fig. 2*. The capacitance C depends on the voltage V.

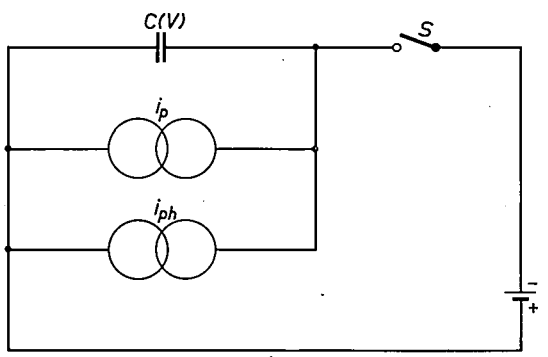


Fig. 2. Equivalent circuit diagram of a *P-N* junction which is charged by an electron beam acting as a switch S.

When the scanning beam passes over the *P-N* junction, it gives to the capacitor a charge Q which, between successive sweeps of the beam, drains off via the two current generators i_{ph} and i_p. These represent respectively the photoelectric effect and the parasitic currents (dark current, surface leakage). If, as a first approximation, we ignore the effect of the current generator i_p, then the current flowing in R during the next scan and corresponding to recharging the capacitor will produce a voltage pulse whose area is proportional to the integral of the illumination received by the junction between the two scans.

The capacitance of the junction per unit surface area is almost independent of the temperature and is determined by the properties of the space charge layer separating the N and P regions. On the other hand, i_p changes rapidly with the junction temperature and with the energy gap ΔE of the semiconductor.

Since the maximum time during which the stored charge is retained is in fact equal to Q/i, there is, as in the case of the photoconductive layer, a condition linking the operating temperature to the threshold of detection. Using relations and data given earlier, it can be expressed by the equation:

$$\lambda_0 T \ll 395\ \mu\text{m°K}.$$

It will be seen that this condition is less severe than that relating to the photoconductive target since it theoretically permits operation at an absolute temperature which is twice as high.

Targets of two types developed at LEP and based on the above considerations, will now be described.

Polycrystalline photoconductive target

Point detectors with a polycrystalline photoconductive layer and operating between 77 °K and 300 °K have been in existence for some time now. The compounds used are chiefly thallium sulphide, the lead sulphides, lead selenide and lead telluride [3]. The sensitive elements generally consist of very thin layers deposited on an insulating substrate and bounded by two electrodes several mm apart; their overall impedance at operating temperature is between 10^4 and 10^8 ohms. However, the use of such compounds in the form of thin layers operating in the direction of their thickness to form electrical images is much more difficult because of their relatively low resistivity. Pure lead sulphide, for instance, has a resistivity of the order of 10^4 Ωcm and would not make a suitable target for a picture-scanning tube since its charge storage times are incompatible with normal scanning frequencies. This drawback will become less and less significant, however, if the resistivity of the microcrystalline photoconductor can be increased to values higher than 10^{11} Ωcm.

Production of the target

One possible way of obtaining photosensitive layers that meet the condition just mentioned is to introduce oxygen into the lead sulphide lattice. This can be done either by oxidizing lead sulphide or by sulphurating lead oxide [4]. The latter process, which has been developed by LEP, is by far the more flexible of the two and enables the resistivity and spectral range of the photoconductive layer to be controlled over a sufficiently wide range. It should however be noted that any increase in the resistivity, which varies with the quantity of oxygen present, is

[1] F. Desvignes and M. Robert, Phénomène d'accumulation dans des jonctions *NP* au germanium refroidies, C.R. Acad. Sci. Paris **252**, 2693-2695, 1961.

[2] M. Robert, Les phénomènes d'accumulation dans les photodiodes. Cas particulier du germanium, Acta Electronica **6**, 341-407, 1962.

[3] J.-J. Brissot and A. Dauguet, Les détecteurs actuels de rayonnement infrarouge, Acta Electronica **5**, 459-516, 1961.

[4] Similar principles have been applied in the "Plumbicon" camera tube with improved red sensitivity.

obtained at the cost of response at the higher wavelengths. A satisfactory compromise consists in cooling the sensitive layer to the temperature of liquid nitrogen. A lightly doped silicon disc forms a suitable substrate for the layer; in fact, being an electrical and thermal conductor, it serves as a signal electrode for the tube, while at the same time it transmits the cold required to cool the layer; moreover, it acts as a filter for the short wavelengths which are generally undesirable for observation in the infra-red.

Preparation of the sensitive layer [5][6] consists in raising the silicon disc to a temperature of 120 °C and depositing on it a layer of pure lead oxide with a thickness of 12 to 15 microns in a vacuum of approximately 10^{-2} torr. The disc with the deposit is then transferred

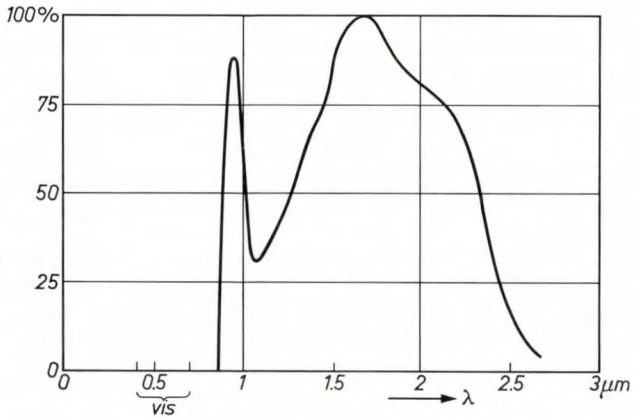


Fig. 3. Relative spectral sensitivity of a camera tube with a lead-oxysulphide image electrode ("SAPHIR" tube), at 77 °K. The bracket indicates the visible part of the spectrum.

Fig. 4. This reproduction of a photograph viewed in infra-red light by a tube equipped with the lead oxysulphide target illustrates the definition obtainable.

to a reaction chamber which is first evacuated and then connected to a source of very pure gaseous hydrogen sulphide. By controlling the temperature, the pressure of the H_2S and the moment at which the reaction occurs, the characteristics of the lead oxysulphide formed are adjusted in such a way that the operation of the target when placed in a vidicon-type tube is optimal.

Characteristics

The spectral sensitivity curve of a photoconductive layer obtained in the above manner is shown in *fig. 3*. The peak in the sensitivity at about 0.9 micron is probably due to the transfer of carriers from the silicon to the layer of lead oxysulphide.

A target of this kind when cooled to 77 °K and scanned in accordance with the conventional 625 line standard is able to detect objects at a minimum temperature of 130 °C (403 °K) with a sensitivity of 2 °C if the camera tube is equipped with a fluorite optical system with an aperture of *f*/4.

The *minimum detectable power* found by progressively attenuating by means of neutral density filters the

radiation from a tungsten filament at 2830 °K, is about 6×10^{-6} W per cm^2. *Fig. 4* gives an indication of the *definition* of the infra-red picture obtained when aiming the camera tube at a photograph.

Owing to the microcrystalline composition of the photoconductive layer, there is a time constant, though not a very large one, that affects the response when the light signal decreases, and a slight after-image is observed with high illumination. This after-effect corresponds to a value much lower than 5% of the residual signal after a 200 ms blocking pulse.

P-N junction mosaic

We now give a description of a target using the photovoltaic effect and based on the considerations outlined in the introduction.

Production of the target

The target consists of a mosaic of *P*-type regions formed upon a thin substrate of *N*-type germanium. The successive stages in manufacture can be summarized as follows:

a) cutting out, grinding, polishing and cleaning of the germanium substrate (diameter 25 mm, final thickness 250 μm);
b) deposition of a mosaic of indium dots on one face of the substrate by evaporating through a metal grid in a high vacuum;
c) alloying the indium to the germanium in a hydrogen atmosphere;
d) electrolytic etching of the junctions thus formed;
e) filling up the grooves separating the junctions with an insulating material so that the electrons of the scanning beam do not have direct access to the *N*-type zone.

The resultant structure is a mosaic of photodiodes at 50 μm between centres (side of cell 35 μm, space between cells 15 μm). With a picture size in the proportion of 4:3 this gives a definition of 300 lines with 400 dots per line.

Fig. 5 shows an example of the electrical reverse characteristics of the junctions as a function of temperature. It is seen that the lower the temperature, the less clearly the breakdown voltage is defined and the smaller the useful voltage range becomes. At very low temperatures, this effect limits the permissible potential difference between the target and the cathode of the tube to a few volts.

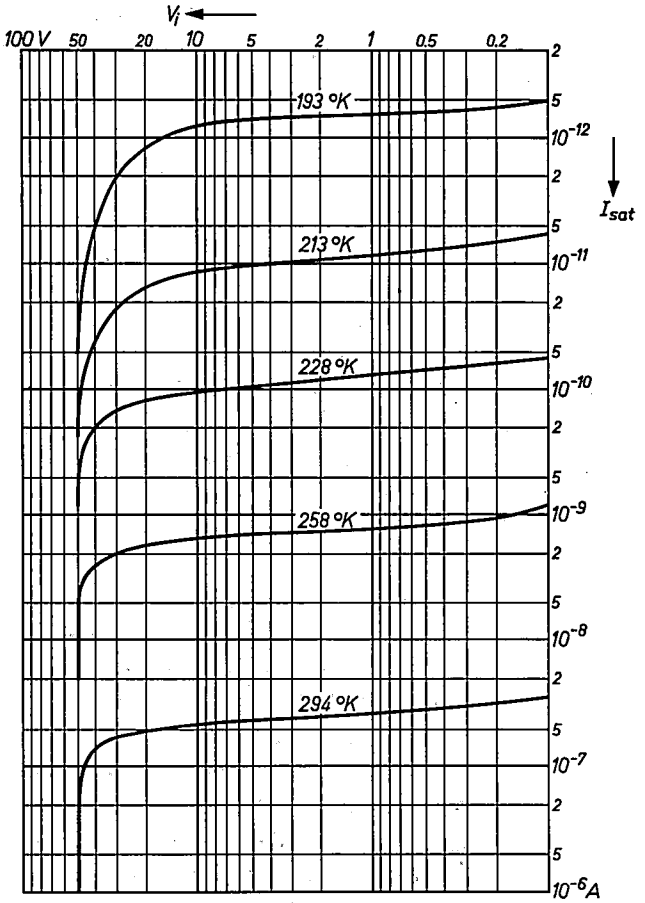


Fig. 5. Static current-voltage characteristics of the germanium photodiodes in the reverse-bias region, for various temperatures. The saturation current I_{sat} is plotted as a function of the reverse voltage V_1.

Characteristics

The detection characteristics of a photodiode operating in a charge-storage mode obviously vary with the operating conditions (i.e. temperatures, frequency and associated optical system). The measured values which are given below were obtained under the following conditions:

Target temperature: 77 °K (boiling point of liquid nitrogen).
Frame frequency: 50 Hz.
Objective: made of fluorite with a maximum relative aperture equal to *f*/4.

It should be mentioned that the operating temperature was chosen for reasons of convenience; in fact, the tube operates satisfactorily at any target temperature below 170 °K.

1) Spectral sensitivity

Fig. 6 shows that with such a mosaic it is only possible to obtain pictures in a fairly narrow spectral zone centred on 1.5 μm. The shape of the spectral sensitivity curve and the position of the detection limits in

[5] French Patent No. 1460817, granted to LEP, Oct. 1965: Procédé de fabrication de couches photosensibles.
[6] G.-A. Boutry, J.-J. Brissot, R. Legoux, J. Périlhou and G. Piétri, Un tube convertisseur d'image pour l'infrarouge moyen, "le Serval", Philips Res. Repts. **20**, 684-706, 1965.

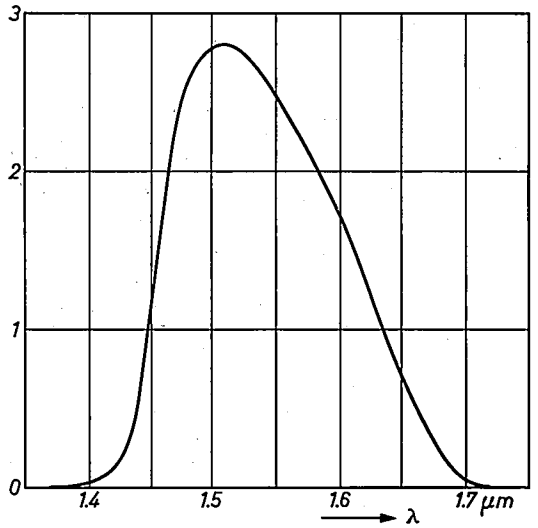


Fig. 6. Spectral sensitivity (in arbitrary units) of the mosaic of germanium photodiodes.

the spectrum are easily explained. At ordinary temperatures the photoelectric threshold of germanium is 1.9 μm; since the energy gap of the material increases when the temperature decreases, the detection threshold drops to 1.7 μm at 77 °K. The shape of the curve on the lower side of 1.5 μm is determined on the one hand by the rapid increase of the absorption coefficient of germanium with photon energy and on the other by the short life-time of the carriers at low temperatures. The diffusion length of the holes is much smaller than the thickness of the germanium layer so that when they are released near the front face by radiation of short wavelength (i.e. by radiation which is totally absorbed in a thin layer of the material), they recombine before reaching the junctions. Consequently, the only carriers which can contribute to the signal are those created near the junctions by photons penetrating sufficiently deeply into the germanium.

When the observed scene is illuminated by a source with a high colour temperature, this absence of sensitivity at short wavelengths is a drawback since the tube only utilizes a fraction of the energy radiated by the source. On the other hand, if the objects concerned are warm objects detectable by their own radiation, the loss of overall sensitivity is insignificant as long as the wavelength of the emission maximum of the radiating body remains much higher than 1.7 μm ($T \ll 1400$ °C or 1673 °K for a black body). It should be noted that in any case this behaviour of the target is conducive to obtaining a clear image whose sharpness depends only on the distance between the centres of the mosaic dots and the diameter of the scanning spot. One reason for this is, that observation in a narrow spectral band makes it easier to design and manufacture achromatic objectives of large aperture, and another is that since carriers created at a large distance from the junction are not collected, there is no danger that lateral diffusion of these carriers will introduce any haziness into the picture.

The selective nature of the spectral sensitivity curve is illustrated by *fig. 7*. The first exposure (*a*) was made in visible light with a normal emulsion; the second (*b*) shows the picture obtained on the screen of the receiver when the subject is "illuminated" by an infra-red source. At 1.5 μm, water in liquid form gives an intense absorption band; it is the quantity of moisture in the skin, the hair or the clothing that determines the contrasts.

2) Minimum detectable energy

The minimum detectable energy was measured by progressively reducing by means of neutral filters the radiation of a tungsten filament heated to 2830 °K. With this light source the target illumination corresponding to a picture which is not distinguishable from its background is about 5×10^{-8} W per cm^2.

3) Minimum detectable temperature

The minimum detectable temperature depends on the objective used. With a relative aperture of *f*/4.5, it is possible to obtain a picture of a black body raised to a temperature of 180 °C (453 °K). This limit is lowered to 150 °C (423 °K) if an objective with an aperture of *f*/1.5 is used, this being the maximum value that the aperture can have in view of the geometry of the tube.

4) Resolving power

To measure the resolution the image of a test pattern consisting of alternate black and white parallel bands is projected on to the target. The resolution limit is a function of the relative orientation of the scan, the bands in the test pattern and the grooves in the mosaic. Generally speaking, interference between these regular structures produces moiré effects which prevent the resolution from being isotropic. The curve in *fig. 8* was obtained by orientating the diagonal formed by the square islands on the mosaic parallel to the line-scan direction and arranging the lines of the test pattern perpendicular to this same direction. If allowance is made for the imperfections of the pattern projector used, i.e. for the contrast transfer function of this instrument, the curve is in good agreement with the predicted result.

Structure of the tubes used

It is useful to give a brief description here of the tube developed at this Laboratory for use with the targets which we have just described.

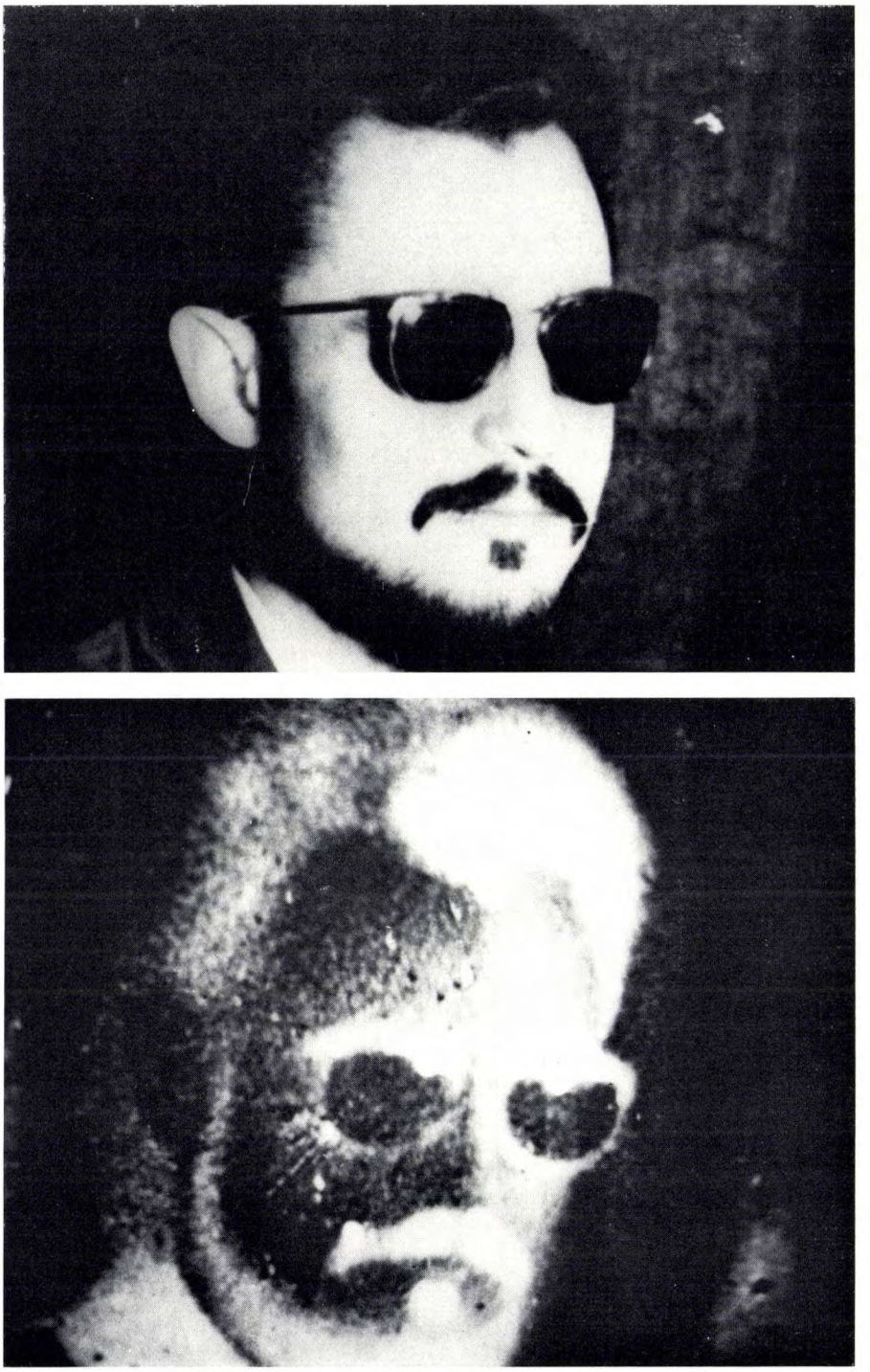


Fig. 7. The same subject, *a*) photographed with visible light on a standard emulsion, *b*) as shown on the screen when illuminated by an infra-red projector and viewed with a television tube equipped with the mosaic target. The information conveyed by method (*b*) is rather different from (*a*).

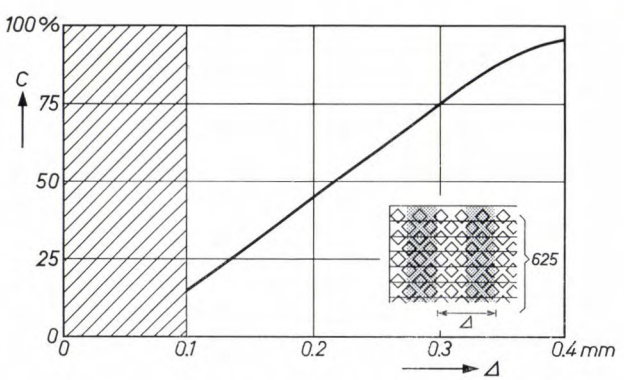


Fig. 8. Resolving power of the mosaic of photodiodes, for reproduction of a test picture with black and white stripes spaced at a distance Δ between centres, the relative orientation of mosaic, test picture and scan being as shown in the inset. The measured contrast C between the light and dark stripes is plotted as a function of Δ. Moiré effects occur in the shaded region.

In general outline, the structure (*fig. 9*) is that of a vidicon, i.e. a thermionic cathode combined with an electronic optical system capable of focusing and collimating low-velocity electrons and a target receiving the infra-red picture on the face away from that exposed to electronic scanning. A set of external coils, which are not shown, supply the fields for focusing and deflection. The differences between the structure of this tube and that of the conventional vidicon are those made necessary by the need to cool the sensitive surface and to shield it from the thermal radiation of the cathode.

The electronic optical system

The electronic optical systems at present used in vidicons are unsuitable when the spectral sensitivity range of the tube extends into the infra-red, since these systems allow the target to "see" the cathode; the energy radiated by the latter would produce a haze on the image which is denser the greater the sensitivity. In the case of a mosaic consisting of germanium photodiodes, the sensitivity is in fact so high that the level of illumination corresponding to saturation would be greatly exceeded.

There are various devices which can be used to over-

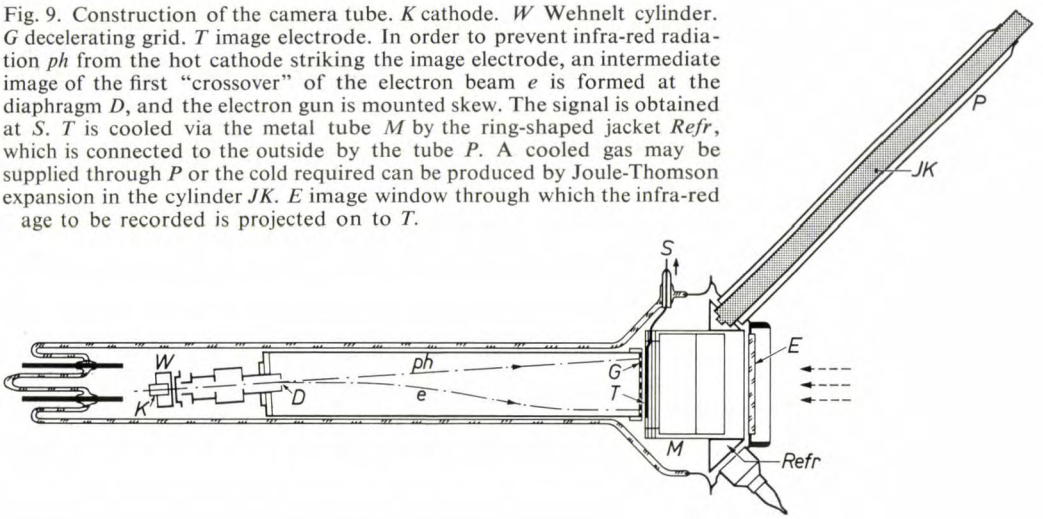


Fig. 9. Construction of the camera tube. *K* cathode. *W* Wehnelt cylinder. *G* decelerating grid. *T* image electrode. In order to prevent infra-red radiation *ph* from the hot cathode striking the image electrode, an intermediate image of the first "crossover" of the electron beam *e* is formed at the diaphragm *D*, and the electron gun is mounted skew. The signal is obtained at *S*. *T* is cooled via the metal tube *M* by the ring-shaped jacket *Refr*, which is connected to the outside by the tube *P*. A cooled gas may be supplied through *P* or the cold required can be produced by Joule-Thomson expansion in the cylinder *JK*. *E* image window through which the infra-red age to be recorded is projected on to *T*.

The cooling device

The detection threshold of the germanium target is not so far into the infra-red as to make it necessary to cool the walls of the tube. The device chosen to cool the target is a vacuum-tight enclosure, annular in shape, connected to the flange carrying the input window by a double-walled tube (see fig. 9). The vacuum in the enclosure acts as a thermal insulator and the metal used has a very low thermal conductivity; thus the heat exchange between the cooling device and the exterior is reduced to a minimum. Since, therefore, the walls of the tube and, more particularly, the input window are practically always at ambient temperature, there is no danger of icing or thermal shock at the glass-to-metal seals when operation of the equipment is started.

Cooling can be effected either by injecting refrigerating fluid (e.g. liquid nitrogen or argon) or by producing the required cold *in situ* by means of a Joule-Kelvin expansion valve inserted in the double-walled tube and connected to a cylinder of compressed gas or to a compressor. The refrigerating power required is low, about 1 W, which corresponds approximately to the evaporation of 30 ml of liquid nitrogen per hour.

come this disadvantage. We have adopted the following. A diaphragm *D* of small diameter (0.4 mm) is placed at the entrance to the collimation chamber, a few centimetres from the cathode; an electrostatic lens placed between the triode assembly of the gun and this diaphragm forms on the latter the image of the crossover (the first focusing point for the electrons from the cathode). Thus the aperture of the light beam emerging from the cathode is limited without affecting the beam current. It is now sufficient to tilt the gun-lens assembly slightly (approximately 4°) in relation to the axis of the tube to ensure that the light beam falls outside the sensitive surface. The electron beam is brought back into the axis of the tube by a fixed transverse magnetic field provided by the alignment coils; then it is refocused on the mosaic by the conventional axial magnetic field.

The infra-red television camera

The images obtained with the infra-red television camera developed at LEP are formed on the target of the tube by means of an objective made of fluorite provided with an anti-reflection coating. The objective is suitable for the 1.2-2.8 μm band, has a focal length of 10 mm and a maximum aperture equal to *f*/4. The

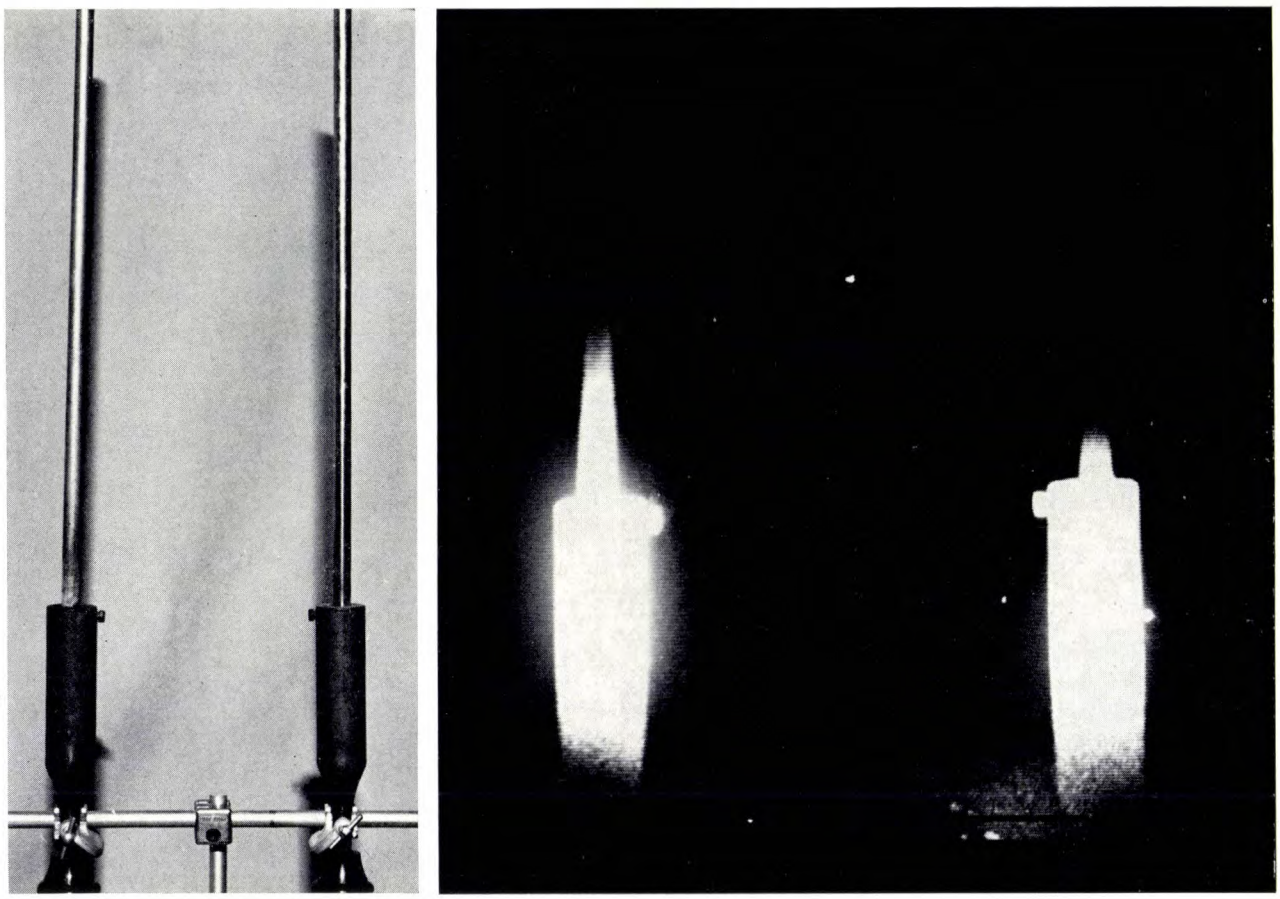


Fig. 10. Experimental demonstration of passive viewing in the infra-red.
a) Two rods, of copper (*left*) and stainless steel (*right*), are heated to 450 °C on the underside.
b) Where the temperature has fallen to about 150 °C along the rod, the luminance of the television picture falls sharply. The difference in thermal conductivity of the rods can be clearly seen.

scanning standard is 625 lines per image and 50 frames per second. This camera can use either of the two tubes described above; its electronic circuits are almost completely identical with those of a light-weight outside-broadcast camera developed at LEP for the O.R.T.F. These circuits, which are completely transistorized, include in particular a crystal-controlled time-base generator and perform all the amplification, scanning and power-supply functions which are necessary for obtaining a complete vision signal satisfying the standards of conventional television. The video signals can therefore be connected to a direct-viewing television receiver or, if desired, recorded on video tape for viewing at a later moment. The target is cooled with the aid of a two-stage Joule-Kelvin expansion valve [7] supplied with gaseous nitrogen (or argon) of high purity (a filter is incorporated in the camera) and at a minimum pressure of 100 bars.

An experiment on closed-circuit television was carried out with this camera equipped with a lead oxysulphide tube. The infra-red picture (*fig. 10b*) of two metal rods (fig. 10*a*) of the same dimensions but of different materials heated to 450 °C (723 °K) at their base brings out the difference of thermal conductivity between these materials. The zone of decreasing luminance corresponds for both rods to a surface temperature of about 150 °C (423 °K).

[7] M. C. Lefranc, Le refroidissement des détecteurs de rayonnement infrarouge, Acta Electronica **9**, 47-90, 1965.

Summary. The vast majority of modern camera tubes used for the visible part of the spectrum operate on the charge-storage principle. In the infra-red, because of the low individual energy of the photons, application of this principle runs into difficulties of both a fundamental and a technical nature. A brief examination of the physical aspect of this problem permits the establishment of relations between the target temperature and the spectral detection threshold; these relations differ depending on whether the photoconductive effect or the photovoltaic effect is considered. This article discusses the technical aspects and behaviour of two types of target which have been developed: one is photoconductive and consists of a microcrystalline, macroscopically homogeneous layer of lead oxysulphide (photoelectric threshold at 2.7 μm); the other is photovoltaic and comprises a mosaic of *P-N* junctions formed on a monocrystalline substrate of germanium (photoelectric threshold at 1.7 μm).

Infra-red thermography equipment for medical applications

M. Jatteau

The work on "passive" infra-red television, originally directed towards military applications, has led to the development of cameras with opto-mechanical scanning. Once instruments of this type were available, the possibilities of — and problems with — this imaging technique for medical and industrial use could be examined. For example, a LEP camera with spiral scanning, operating in the spectral region of 4 μm [1], gave interesting images of the thermal radiation patterns of the human body [2] and allowed us to study methods by which the information gathered could be used to the best advantage.

The interest shown by the medical profession in the first thermography experiments encouraged us to consider the design of an equipment more suited to this type of application, and as a start we wished to investigate the experimental conditions which might be encountered in practice. For this purpose a thermography equipment was built at LEP to a design which had a number of features that enabled doctors to undertake a wide range of investigations. The instrument has a high thermal resolution, it can be used for plotting isotherms (to an accuracy better than 0.05 °C) and it can also operate in different spectral ranges (around 4 μm and 10 μm).

We shall describe this equipment and report on the experiments performed with it by the Radiology Service of the Hospices Civils de Strasbourg, which is specially equipped for thermographic examinations.

Discussion of basic design requirements

Whatever the method used, the infra-red television image is formed by transposing the wavelength of the radiation coming from the different elementary surfaces of a scene into the visible spectrum. Despite a certain analogy with ordinary television the situation is different, for the eye has no way at all of checking the image against the natural scene. The infra-red image can be used mainly in two different ways. In the first, the image should allow a detail in the scene to be identified or localized or just detected; to achieve this, the energy of the radiation originating from this detail and received by the infra-red television unit need only be different from the energy radiated by its surroundings. The detection may then possibly be *active* [3]. In the other — and this is the case in thermography — there is an additional requirement: the luminance variations in the image must represent the differences in surface temperatures. This latter condition determines the direction in which the equipment design should proceed; in this case we must distinguish in the total radiation received from each element of the scene that part which is not due to reflection but which the element emits on its own: only this part is directly dependent on its temperature [4]. The detection must therefore be *passive*.

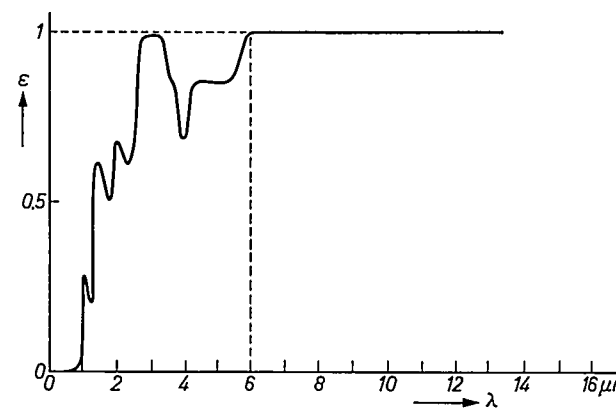


Fig. 1. Variation of the skin emissivity factor ε as a function of the wavelength (post-mortem sample) [5].

The relationship between the energy of the emitted radiation and the surface temperature, i.e. the emissivity ε_λ of the surfaces studied must also be taken into account and this can lead to the determination of a preferred spectral operating range. In the case of medical applications, the surface studied is the skin and according to certain authors [5][6] the healthy skin behaves like a grey body (ε_λ = a constant close to unity) beyond 6 μm (*fig. 1*).

As far as we know, the "instant-view infra-red television" equipment in medical use up to the present operates in the spectral range around 4 μm (indium antimonide radiation detector). We wanted our equip-

M. Jatteau, Ingénieur C.N.A.M., is with Laboratoires d'Electronique et de Physique Appliquée, Limeil-Brévannes (Val-de-Marne), France.

ment to be able to operate at different wavelengths to allow comparison of the infra-red images obtained in each case.

The design requirements of the equipment were also governed by the technical objectives we had in mind: on the one hand, the equipment must permit us to determine the possibilities of thermography for different clinical applications and, on the other hand, it must allow us to define the characteristics required for medical observations.

In order to detect temperature differences of a few hundredths of a degree Celsius near ambient temperature, the measurements of the radiated power must be carried out with a relative precision of better than one thousandth. In practice this corresponds to the detection of very low energy differences (a few nanowatts).

The consequences of these conditions are as follows:
a) The detection must be carried out in the far infra-red. This conclusion is reached by considering the required resolution in space and time, the optical performance that is physically attainable, and the characteristics of presently available radiation detectors in this spectral range.
b) It is preferable to use point detectors. This avoids the disadvantages resulting from local non-uniformity of photosensitive targets or linear arrays of infra-red detectors.
c) The resolutions in space and time are necessarily limited since, for a given scanning system and a given detector, the value of the minimum detectable temperature difference is closely related to the product of the frame rate and the number of elements resolved in an image.
d) When a point detector is used, scanning of the scene is required and this should be achieved by opto-mechanical means.
e) It is preferable to place the mechanical deflection systems in front of any optical focusing system. This avoids aperture modulation of the infra-red beam intensity, which can be detrimental to the required temperature discrimination if the modulation depth is large.

The direct display of the infra-red images is a very useful possibility which we wanted to exploit. Display is obtained on the screen of a cathode-ray tube. For this purpose the frame rate must be at least equal to 1 Hz (because of the limited time constants of the long-persistence phosphors); quantum detectors rather than bolometers, for example, must be used to provide the fast response necessary for this frame rate and also to provide the spatial resolution required for medical applications, which is at least equal to 10^4 elements per frame.

The disproportion between the number of luminance steps which can be detected by an infra-red camera and the number of luminance steps observable on a cathode-ray tube screen [7] suggests that the image should be restored for successive temperature ranges.

The amplitude non-linearity in the conversion of the infra-red radiation to visible light and the selectivity of the quantum detectors imply that the temperature measurements should be performed by comparison with calibrated radiation sources [8]. A very interesting feature of the cathode-ray tube display is that such comparisons can be made directly on the screen.

These last two points led us to develop a device for plotting isothermal zones, like the contour lines of a geographical map [9].

Description of the thermography equipment

We shall now describe in some detail the equipment made at LEP, stressing the points touched upon in the preceding section.

The camera

1) Suitable quantum detectors.

The principal quantum detectors which are available in the far infra-red are summarized in *Table I*. The fourth column gives the specific detectivity D^*(500 °K), which expresses the r.m.s. signal-to-noise voltage ratio obtained per watt of radiant power per unit area of detector and for a bandwidth of 1 Hz, when measurements are made from a black body at 500 °K. The last column permits a comparison of the detectors with respect to their own contribution to the thermal resolution of a given camera in the region of $T = 300$ °K. The significance of the quantity listed here can be explained

[1] J. Cornillault, Télévision infrarouge dans le domaine spectral situé au voisinage de 4 microns, Acta Electronica **7**, 319-360, 1963.
[2] J. Cayzac, Problèmes posés par la détection des faibles différences de température en télévision infrarouge, Acta Electronica **9**, 7-18, 1965.
[3] Active detection is defined by the fact that the objects can be discerned, as in natural vision, by the radiation which they reflect or transmit and which originates from one or several sources illuminating the scene.
[4] G. Bruhat, Cours de physique générale. Thermodynamique, 5th edition, revised by A. Kastler, publ. Masson, Paris 1962, pp. 591-677.
[5] R. Elam, D. W. Goodwin and K. Lloyd Williams, Optical properties of the human epidermis, Nature **198**, 1001-1002, 1963.
[6] E. Hendler, J. D. Hardy and D. Murgatroyd, Skin heating and temperature sensation produced by infra-red and microwave irradiation, in: Temperature, its measurement and control in science and industry, Vol. 3, Part 3, editor J. D. Hardy, Reinhold Publ. Co., New York 1963, pp. 211-230.
[7] F. Kerkhof and W. Werner, Television, Philips Technical Library, Eindhoven 1952.
[8] G.-A. Boutry and M. Jatteau, Sur la mesure des températures de surface par télévision dans l'infrarouge, C.R. Acad. Sci. Paris **266 B**, 214-217, 1968 (No. 4).
[9] M. Jatteau, Dispositif pour tracer les isothermes dans un appareil d'exploration optique au moyen de rayons infrarouges, French Patent No. 122249, 1967.

as follows. The minimum temperature difference that can be detected between two black bodies near a temperature T by a camera using opto-mechanical scanning and a point detector is proportional to the product:

$$\sqrt{\frac{\Delta f}{A}}\,\frac{1}{\Omega\tau_1}\,\frac{T}{4\tau_2 D_s^* \int_{\lambda_1}^{\lambda_2}\left(\frac{dL}{d\lambda}\right)_T d\lambda}. \quad \ldots \quad (1)$$

α β

The first factor α characterizes the camera: A is the sensitive area of the detector, Ω the aperture angle of the diaphragm-detector system, Δf the bandwidth of the amplifier and τ_1 the transmission factor of the internal optical path of the camera. The second factor β, whose value for $T = 300$ °K is given in the Table, characterizes the spectral sensitivity range. Here τ_2 is the transmission factor in the space between black body and camera, D_s^* the "effective" specific detectivity (defined from D^*(500 °K) taking account of the effective spectral response of the detector for a 300 °K black body) and $dL/d\lambda$ the monochromatic luminance of the black body.

2) Opto-mechanical scanning.

The scanning of the scene in two orthogonal directions is achieved, on one hand, by means of a rocking mirror (period 0.5 s) and, on the other hand, by means of an octogonal prism with plane reflecting surfaces rotating at constant speed (1500 rev/min). The rapid line scanning (the image being swept past the stationary point detector) is thus obtained from a uniform rotary motion. *Fig. 2* is a schematic diagram of the camera with detector D, rotating prism P and rocking mirror F. The relatively slow discontinuous motion of F is achieved by means of a cam C and only involves equipment which is relatively light and of small dimensions.

This system led to the use of a relatively large line-deflection prism, but this enabled us to place the drive motor inside the prism. The engineering of this element, achieved by mechanical milling of optical quality [10], is all the more remarkable in view of its size. The inclination of the "line" scanning prism indicated in fig. 2 is necessary to ensure that the angular vertical field scanned is divided into two equal parts by the horizontal plane. With this arrangement there is no occultation of the field of view by any fixed or mobile mechanical part and, therefore, no spurious modulation of the incoming radiation. The scanning distortion connected with this inclination is acceptable.

Table I. Data for the principal available detectors. The materials are used as photoconductors when not stated otherwise. The thermistor is listed solely for the sake of comparison since its time constant ($\approx$ 1 ms) is unsuitable for the applications envisaged. Maximum values of D^* for a field of view greater than 120° are given; D^* may be increased for certain detectors by limiting the field of view by means of a cooled diaphragm (see P. W. Kruse, L. D. McGlauchlin and R.B. McQuistan, Elements of infra-red technology; generation, transmission and detection, Wiley, New York 1963).

Detector	Operating temperature in °K	Specific detectivity D^*(500 °K) in m W^{-1} Hz$^{1/2}$	Limits of "effective" spectral band		Factor β in (1) for $T = 300$ °K in m °K Hz$^{-1/2}$
			λ_1 in μm	λ_2 in μm	
InSb (photovoltaic cell)	77	13×10^7	3.85	5.45	1.48×10^{-8}
InSb	77	3×10^7	5	8.6	2.98×10^{-8}
Ge(Au)	65	7×10^7	5	8.6	1.27×10^{-8}
Cd(Hg)Te	77	3×10^7	8.24	13	1.57×10^{-8}
Ge(Hg)	$\leqq$ 30	9×10^7	8.1	14	0.6×10^{-8}
Ge(Cu)	4.2	9×10^7	7.8	23	0.4×10^{-8}
Thermistor	ambient	1.5×10^7	0	∞	3.4×10^{-8}

The choice of the optical focusing system is partly determined by the following considerations.

The low temperatures required for the operation of the infra-red detectors for medical cameras are generally obtained by means of liquefied gases. Indeed, for these applications present cooling systems, however simple, are still too cumbersome (Joule-Thomson expansion [11] requires high-pressure compressed gas of high purity, and Stirling-cycle cryogenerators [12] are not entirely free of noise and vibration, while both systems require a certain amount of maintenance). Thus, the detector has to be placed in a cryostat. The dimensions of detector cryostats of sufficient capacity to ensure several hours of independent operation are generally

incompatible with the use of centred mirror systems with wide apertures, unless high light losses are accepted. Preference has therefore been given to an "off axis" telescope comprising a parabolic mirror (*O* in fig. 2) which, by sliding along its optical axis, permits focusing at different distances; the cryostat *Cr* of the detector, a Mullard Ge(Hg) cell, is in a fixed vertical position.

respectively on the outside and inside of a ring *R* which is mechanically linked to the prism and provided with slits.

The rocking motion of the frame scanning mirror is followed by a high-precision potentiometer *Pot* with a plastic track.

A complete view of the camera arrangement is given in *fig. 3*.

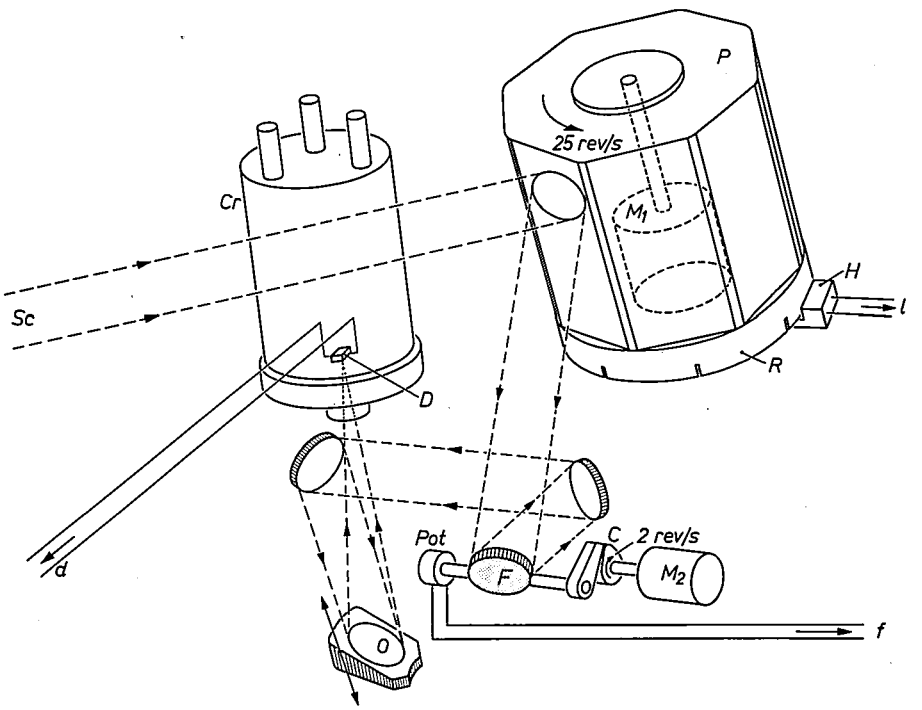


Fig. 2. Schematic diagram showing the principle of the experimental LEP camera for thermography. The line scanning prism *P* (eight plane reflecting faces), made to rotate by motor M_1, and the frame scanning mirror *F*, given a rocking motion by cam *C* and motor M_2, sweep an image of the scene *Sc* past the detector *D*, which is housed in a cryostat *Cr*. The infra-red image is focused on to *D* by the parabolic mirror *O*. The infra-red image conveyed by the detector signal *d* is displayed on the screen of a cathode-ray tube, whose time-base is synchronized by signals *l* and *f* from the optical reading head *H* with rotating slit ring *R* and from the potentiometer *Pot*.

Since the opto-mechanical scanning is performed before the infra-red beam is focused, the system always operates on the optical axis of the telescope whose parameters are determined in such a way that the dimensions of the circle of confusion remain below those of the sensitive area of the detector even at the minimum focusing distance (0.8 m).

4) Synchronizing the scanning signals.

The synchronization of the line scanning signals applied to the image-display tube is provided by pulses supplied from an optical reader head *H* consisting of a collimated light source and a photoelectric cell placed

Image-display device

The image-display device mainly consists of an oscilloscope of classical type (*fig. 4*) equipped with a cathode-ray tube with a large flat screen (diameter 180 mm). A mobile track has been designed to carry the

[10] A. Charles-Georges and A. Salmon, Usinage de cavités de pompage pour lasers "solides", Acta Electronica **10**, 315-330, 1966.

[11] H. T. Hayward, Joule-Thomson liquefier, Cryogenics Symp., Frankfort 1966, pp. 9-12.

[12] A. Daniels and F. K. du Pré, Closed cycle cryogenic refrigerators as integrated cold sources for infra-red detectors, Appl. Optics **5**, 1457-1460, 1966.

scanning signal generator and amplification circuits, the video amplifier, the d.c. power supplies and a monitoring oscilloscope. The isothermal-plotting device described in the following paragraph is located almost completely inside the camera.

$T + \Delta T$) undergo a high, perfectly linear amplification (of the order of 10^7 times). The average level $S_T + \frac{1}{2}\Delta S$ can be freely chosen between the extreme levels. To preserve the accuracy of the isothermal zones which, at their limit, occupy the entire field being scanned, very

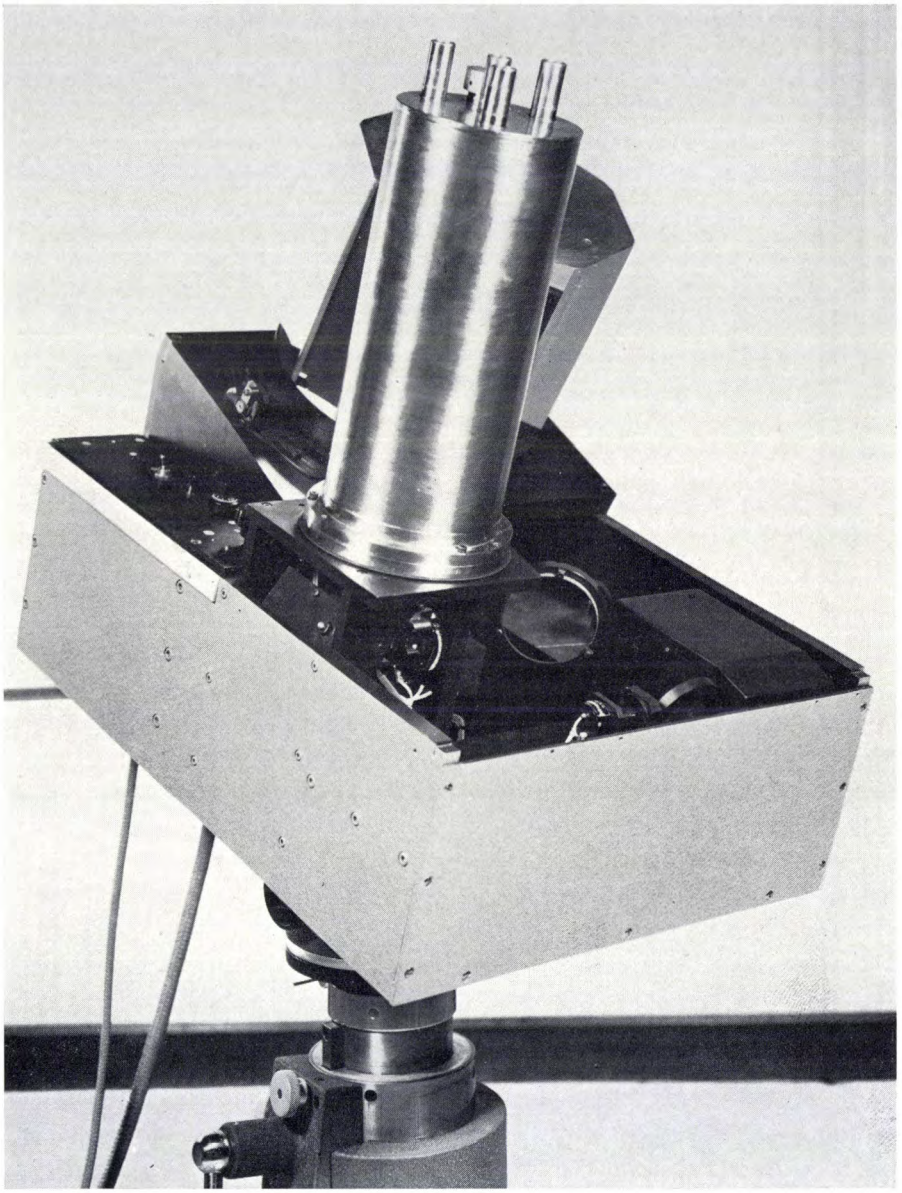


Fig. 3. Camera equipped with a Ge(Hg) detector cooled to the temperature of liquid helium in a double-walled cryostat.

Isothermal-zone plotting device

The principle of the isothermal-plotting device developed at LEP is as follows [9]. The signal supplied by the detector is progressively amplified, then amplitude-limited throughout the electronic chain (*fig. 5*) so that only the amplitude variations ranging between levels S_T and $S_T + \Delta S$ (corresponding to T and

low frequency signals must be amplified without distortion. The upper limit has been chosen equal to the maximum frequency used in television broadcasting. For the device connected to the camera the values obtained are $f_{\min} = 0.18$ Hz (which led us to use an electronic chain with direct coupling) and $f_{\max} = 38$ kHz.

The amplitude ΔS is finally selected at the output S

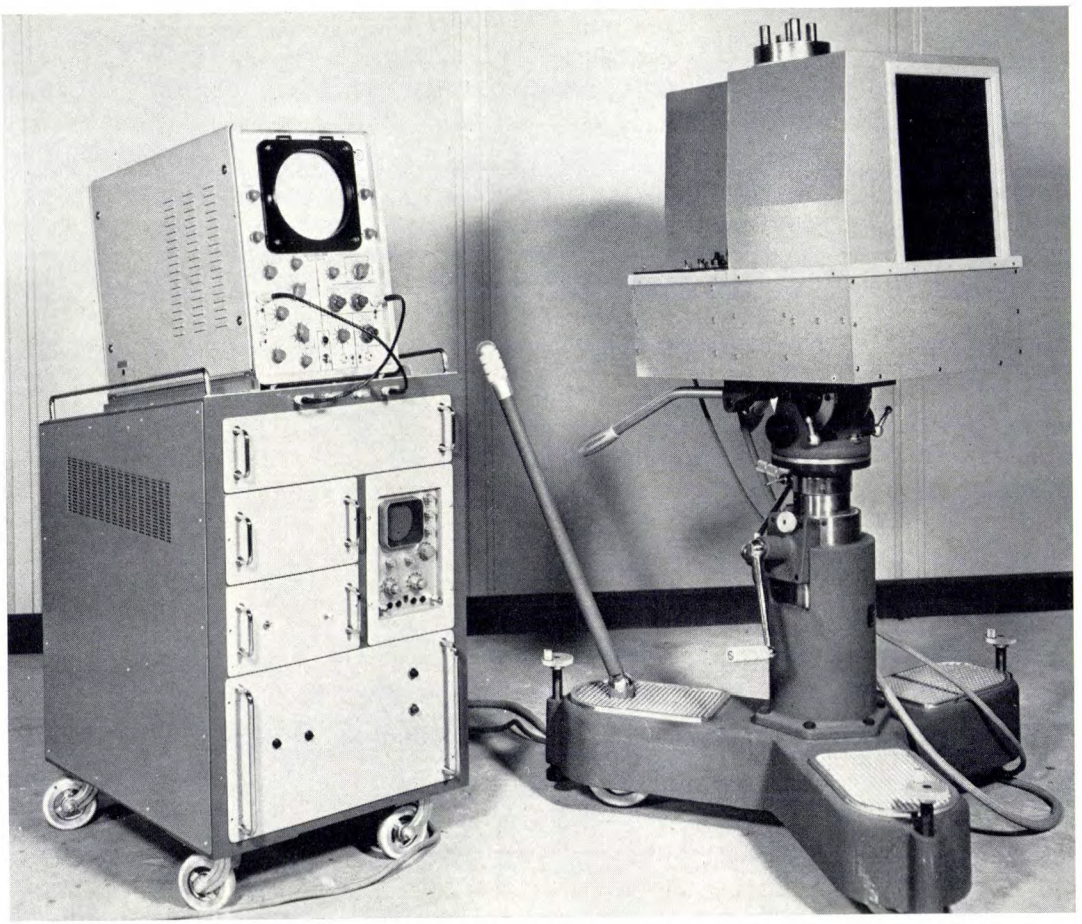


Fig. 4. Overall view of the experimental LEP thermography equipment. The camera is mounted on a mobile tripod with cradle head for easy manoeuvering without sacrificing stability. The image-display device, mounted on a mobile trolley, is shown on the left.

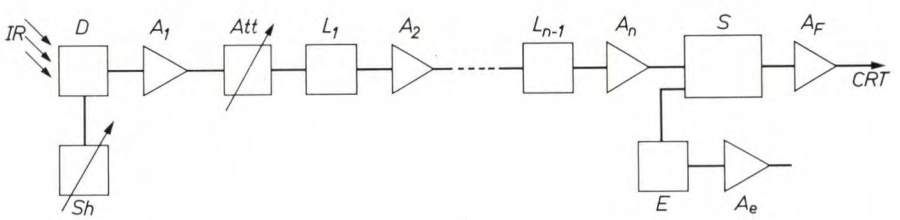


Fig. 5. Block diagram of the isothermal-zone plotting device.

of the electronic chain, whose gain can be adjusted by means of an attenuator *Att*. Each amplifier $A_2 \ldots A_n$ is protected from saturation effects by a limiter $L_1 \ldots L_{n-1}$ and the gain of the preamplifier A_1 is determined such that it will not be saturated by the maximum signal corresponding to the extreme temperatures chosen (0-60 °C). In order to obtain an image which represents only the isothermal zones, every signal whose level reaches at least the upper threshold of selection is fed back to the black level by an inversion signal originating from the selection circuit *S-E* and amplifier A_e. The output amplifier A_F applies the resulting signal to the grid of the cathode-ray tube.

The reference level maintained throughout the amplifier chain corresponds to the optimum bias of the detector. (This is obtained either by means of a compensating circuit when photoconductors are used or directly in the case of photovoltaic detectors.) The different levels are scanned by continuously "shifting" the signals supplied by the detector *D* along the entire scale. This is done by means of circuit *Sh* in fig. 5 which varies the resistance of the bias circuit or the bias voltage.

Technical characteristics

The main technical characteristics of the LEP equipment are listed in *Table II*. The isothermal-plotting device offers the facilities described below and illustrated by some results obtained in the spectral range around 10 μm.

a) Accurate examination of the content of the infra-red images. The isothermal zones can be plotted at any temperature between 0 °C and 60 °C, regardless of the extent of these zones (although of course limited to the element resolved or the scanned field) to a maximum precision of 0.1 °C using a 4 μm detector and better than 0.05 °C using a 10 μm detector. This allows a detailed examination of the images to be made: differences in radiated energy equal to the minimum detectable difference are clearly brought up; precise geometrical location of the thermal details is obtained (*fig. 6a, b, c*).

b) Thermal mapping by superimposing isotherms obtained at different levels on one photographic film or on the screen of a storage tube (*fig. 7*).

c) Continuous variation from 0.05 °C to 60 °C of the temperature range corresponding to the range of luminances presented by the picture. This variation allows the accuracy to which the isotherms are plotted to be freely chosen (fig. *6a, c*). Moreover, the contrast of the displayed images can be improved by adjusting the width of the amplitude range selected as a function of the extreme temperatures of the scene, having regard to the properties of the phosphor of the cathode-ray tube screen (or the photographic film). This is demonstrated by *fig. 8a, b*.

d) Superimposition of the isothermal zones on a normally contrasted image (fig. *8b*).

e) Rapid discrimination of regions higher in temperature than the isotherm displayed. This possibility is useful when, during the examination, the isotherm delimits several dark zones each of which may be warmer or colder (*fig. 9*).

f) Possibility of differential and absolute temperature measurements by comparison with reference sources. With a Ge(Hg) detector the maximum precision of the comparisons made directly on the image corresponds to a temperature difference between two "black" sources of less than 0.05 °C.

Table II. Technical characteristics of the LEP thermography equipment.

Repetition rate	2 frames per second	
Definition	100 lines per frame 135 picture elements per line	
Angular field scanned	about 420 × 560 milliradians (24 × 32 degrees)	
Angular resolution	4 milliradians	
Focusing distance	from 0.8 m to infinity	
	Operation at 10 μm (detector Ge(Hg) at 4 °K)	Operation at 4 μm (detector InSb at 77 °K)
Thermal resolution around 300 °K	< 0.05 °C	≈ 0.1 °C
Width of isothermal zones plotted	from 0.05 °C to 60 °C	from 0.1 °C to 60 °C
Temperature range of isothermal-zone plotting	0 to 60 °C	0 to 60 °C

Conclusions derived from medical experiments using the equipment

Since the end of 1967 the LEP equipment has been used experimentally in France at the Central Radiology Service (Director: Professor Gros) of the Hospices Civils de Strasbourg where a team of doctors and physicists are studying the fundamental problems of medical thermography. Several thousand examinations have been performed by this team with a Barnes thermograph in various clinical services [13] and various types of equipment have been tried out. Every examination is performed in a thermography room kept at a constant temperature (21 ± 1 °C) [14], approximately 15 minutes after the patient has been prepared.

An examination made with the LEP equipment consists in photographing an image with the best possible contrast and then plotting a series of four to six isotherms. These plots are made for the temperature levels considered to be of interest during a previous analysis made by direct observation of the screen of the display

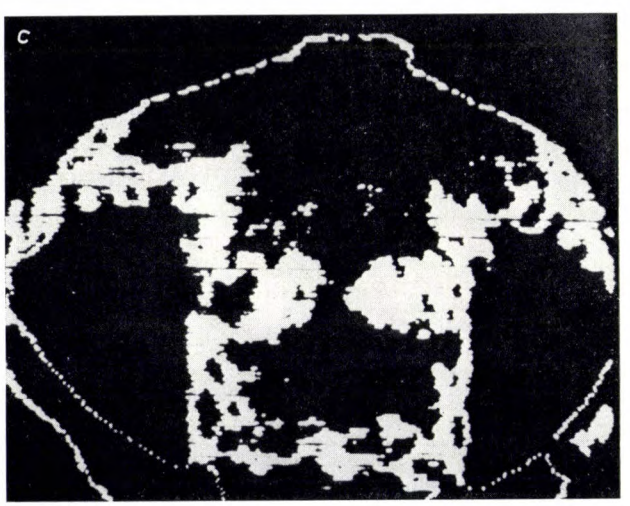


Fig. 6. Image of isothermal regions of the breasts. Isotherms are plotted with different temperature resolutions ΔT and for different temperatures. *a*) $\Delta T < 0.05$ °C at a certain temperature T_1. *b*) $\Delta T < 0.05$ °C at a temperature $T_2 < T_1$. *c*) $\Delta T \approx 0.2$ °C at $T_3 \approx T_1$.

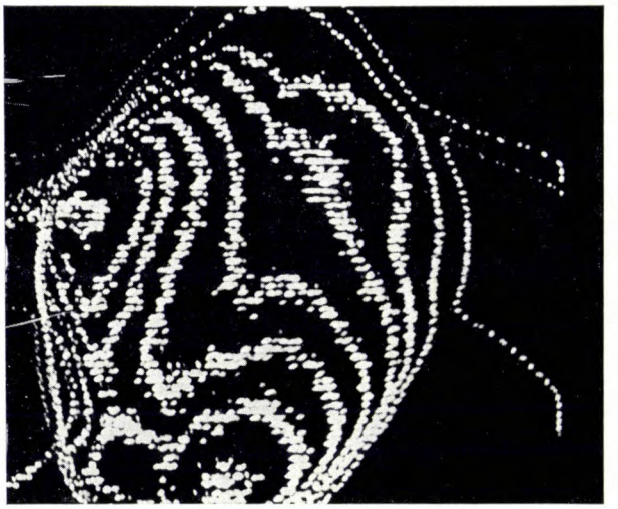


Fig. 7. Thermal map of a leaf of Ficus elastica. Width of the isotherms less than 0.05 °C, temperature difference between isotherms $\approx$ 0.1 °C, total temperature range $\approx$ 1.2 °C. The temperature increases from the edge of the leaf to the circular region at the left.

Fig. 8. Infra-red images of a face. *a*) Demonstration of luminance contrasts in the hair. *b*) Demonstration of contrasts in the temperature of the skin. The contour of the face and one isotherm are superimposed on this image.

It would not be possible to observe the contrasts reproduced in (*a*) and (*b*) in a single picture.

[13] Colloque international de thermographie médicale, Strasbourg 1966, J. Radiologie **48**, 1-98, 1967.

[14] Ch. Gros, M. Gautherie and P. Bourjat, Conception d'une salle de thermographie médicale, J. Radiologie **49**, 371-372, 1968 (No. 5).

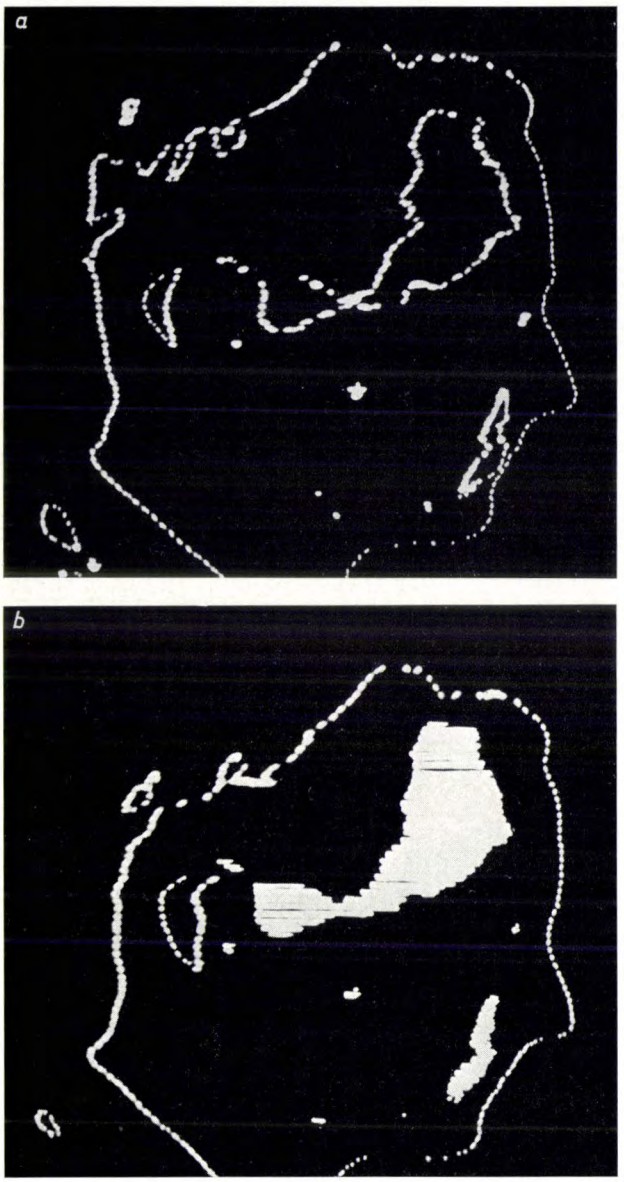


Fig. 9. Infra-red images of a face as recorded by the isothermal plotter. *a*) Isotherms for $T = 34$ °C. *b*) Temperature regions above 34 °C.
The outline of the face was recorded by the same equipment but at a different adjustment.

tube (thermoscopy). *Figs. 10* and *11*, which have kindly been provided by the Strasbourg Radiology Service, show the results of examinations made with the LEP equipment operating in the 10 μm region.

Usefulness and problems of medical thermography

Thermography is considered by medical experts [13][15] to be an investigation technique complementary to the more conventional scientific methods (radiology, scintigraphy, histology, etc.). It is attractive not only because of the originality of the type of information provided but also because it is harmless both to the patient and the practitioner. The main special problems to which it seems thermography can contribute effectively are, in particular, mammography (the detection and location of breast cancer), the study of vascular diseases, rheumatology, placentography, laryngology, the study of the effects of drugs and vaccines, the study of the development of skin grafts, etc.

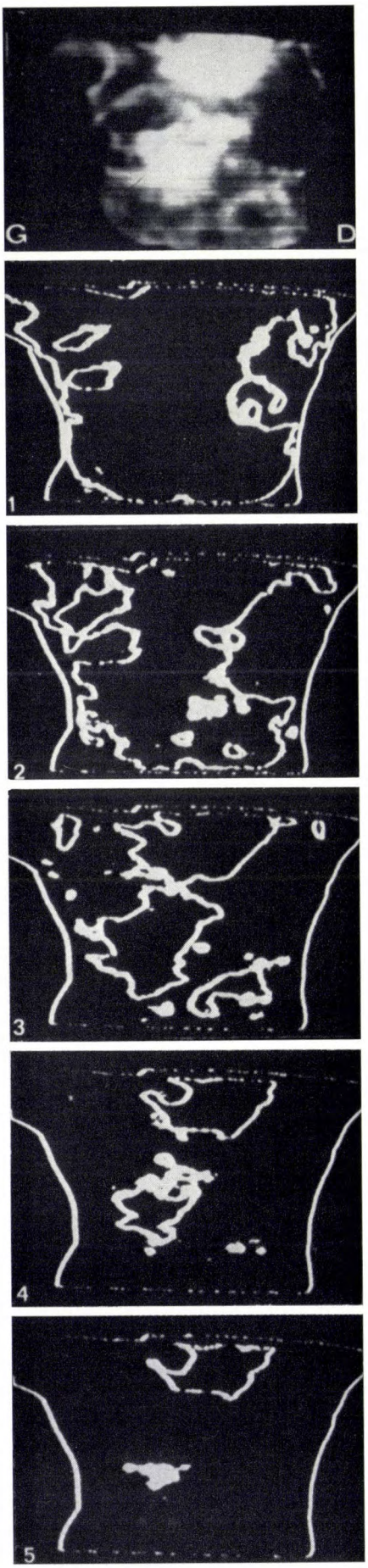
Nevertheless, even today it seems that the development of medical thermography is linked up with the study of the fundamental problems posed by the exact interpretation of the images from the physical and medical points of view [16].

In particular, these problems relate to the determination of differential and absolute temperatures from the infra-red radiation, the knowledge of the thermo-stabilization mechanisms which result in a variation of skin temperatures with time, the understanding of the processes by which internal pathological phenomena can "project" themselves so as to be revealed by thermal anomalies, and the improvement of thermographic equipment for a better adaptation to medical applications.

A number of detailed results

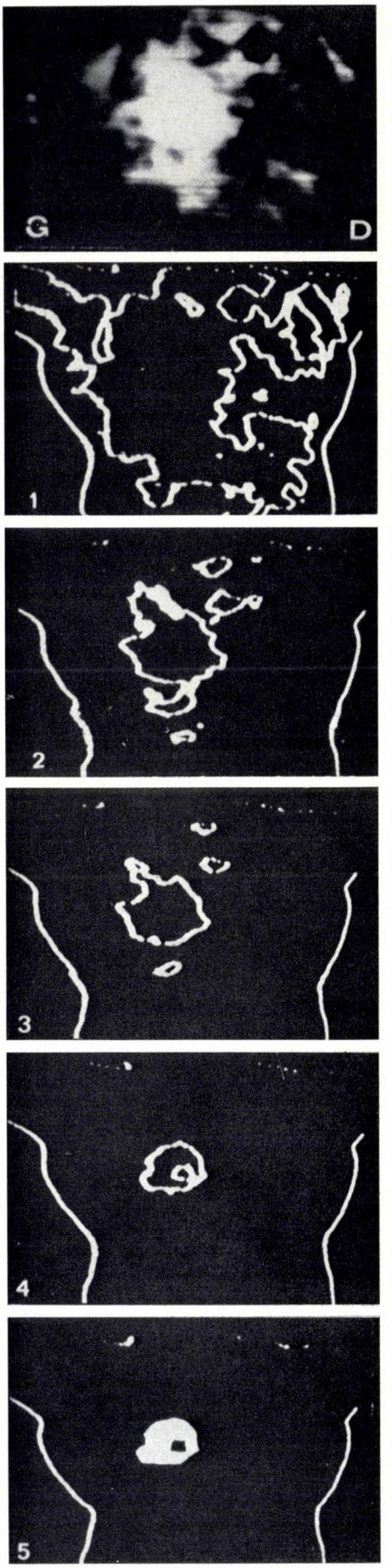
1) Thermography corroborates the radiological and clinical examinations in approximately *80% of cases of breast cancer.*
2) *The plotting of precise isotherms* permits better location of pathological sites than on a normal image. This should lead to a closer correlation between the thermographic and the radiographic examinations.
3) As a result of its *high temperature resolution* and the possibilities offered by the isothermal-plotting device, the LEP equipment has enabled some 20% more pathological cases to be detected than when a thermograph using a bolometric detector was used. It appears, however, that the sensitivity of the equipment would be sufficient even if the minimum detectable temperature difference is as large as 0.1 °C [17][18] and corresponds to the minimum width of the isotherms.
4) After a short period of adaptation, the user is not troubled by the low frame rate for *direct observation of the images.* This direct observation permits the study of dynamic thermal phenomena such as the natural and artificial cooling of the skin surface and the effect of drugs and vaccines.

[15] Thermography and its clinical applications, Ann. New York Acad. Sci. **121**, art. 1, 1-304, 1964.
[16] R. Genève, Introduction à la thermopraphie médicale, Acta Electronica **12**, 7-19, 1969, (No. 1).
[17] Ch. Gros, M. Gautherie, P. Bourjat and C. Vrousos, Thermographie des affections mammaires, in: Medical thermography, Proc. Boerhaave Course Postgrad. Med. Educ., Leiden 1968, pp. 68-81.
[18] Ch. Gros, M. Gautherie, P. Bourjat and C. Vrousos, Les applications médicales de la thermographie infrarouge, Acta Electronica **12**, 63-119, 1969, (No. 1).



←

Fig. 10. Infra-red images of female thorax showing an encysted abscess in the left breast.
At the top an image with normal contrast.
1-5: five recordings of isothermal zones with temperatures increasing from *1* to *5*.
The isothermal zones are superimposed on the trace representing the outline of the body.



→

Fig. 11. Infra-red images of female thorax with epithelioma, left breast ulcerated.
At the top an image with normal contrast.
1-4: four recordings of isothermal zones with temperatures increasing from *1* to *4*.
5: image of entire region with temperatures above that represented by the isotherms in *4*.
The isothermal zones are superimposed on the trace representing the outline of the body.

In some cases, the thermal contrasts can be improved by artificial cooling of the body region studied; this is due to the fact that regions of healthy skin have a different cooling rate from pathological areas [18].

5) With regard to the *spectral range*, the comparison of images of normal contrast obtained simultaneously from one subject with LEP equipment operating in the 10 μm range (Ge(Hg) detector) and with the Barnes thermograph tends to confirm the remarks concerning the spectral emissivity of the skin (see p. 278). The infra-red radiant power emitted by healthy and pathological areas as compared with that emitted by the black body, when viewed at different angles of incidence, wavelengths, and temperatures, has still to be investigated.

6) In view of the variations in skin temperature as a function of time [18][19] *the duration of observation* of the body region concerned must be sufficiently short to ensure that errors due to such variations are smaller than the minimum temperature difference which can be detected by the equipment. This duration depends particularly on the body region concerned, the efficiency of the artificial cooling or, if the cooling is natural, the adaptation time of the patient, and the ambient temperature (*fig. 12*). The recording of a certain number of preselected isotherms must be done in a time which is, at most, equal to this period of observation. This limits the duration of image formation. For example, for a patient-adaptation time reduced to five minutes and an examination which consists of recording ten isotherms (precision 0.1 °C) the duration of image formation should not exceed 4 seconds.

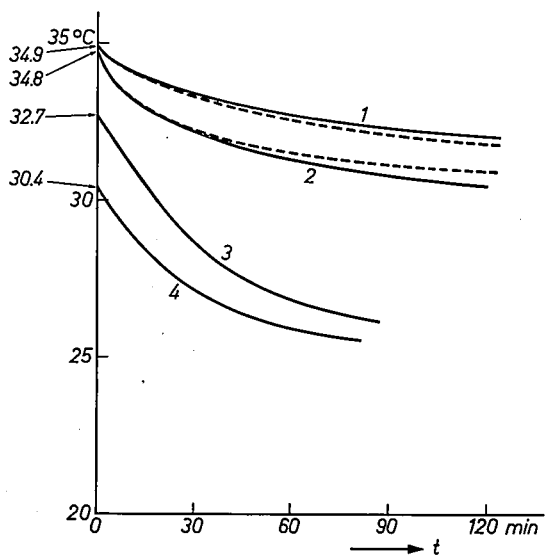


Fig. 12. Experimental curves (solid lines) and theoretical curves (dashed lines) of the variation of the skin temperature of various parts of the body as a function of the time t. These curves depend on the ambient temperature T_a. (Taken from Ch. Gros, M. Gautherie, P. Bourjat and C. Vrousos [18].) *1* forearm, $T_a = 23.5$ °C. *2* abdomen, $T_a = 22.1$ °C. *3* hand, $T_a = 21.2$ °C. *4* foot, $T_a = 21.8$ °C. The precise initial temperatures are indicated for each case.

7) It is convenient to *record* isotherms of different levels in different colours on the same photograph for an easier interpretation of the results.

8) The dimensions of the element resolved on a breast may be a few millimetres at a working distance of 1.5 to 2 metres; this value is the result of a compromise between the ideal conditions of observation and the technical requirements.

Improvements in thermography equipment

The present conclusions of the medical profession, and the technical suggestions resulting from the development and practical use of the experimental equipment described, have enabled us to define the principal characteristics of a more suitable thermography equipment for medical applications.

Returning now to the design requirements formulated on the basis of the different objectives established by ourselves, particularly with respect to the method of image display on a cathode-ray tube screen, it is possible to define the technical characteristics of a new equipment in which the angular resolution would be increased with respect to the previous version (1.5 milliradians) and the scanned field would be adapted to the medical "scenes" (about 210×280 milliradians, i.e. 12×16 degrees), as would the thermal resolution (0.1 °C). The 10 μm spectral range would be preferred (using Ge(Hg) point detectors or, in the near future, CdHgTe detectors cooled to the temperature of liquid nitrogen). The possibilities offered by the isothermal-plotting device would be retained and even increased by semi-automatic recording, recording in colour, and the superimposition of a normally contrasted picture.

A different arrangement of the opto-mechanical scanning systems (pyramidal "line" analysing prism) would permit the development of a more compact camera than the one used in the experimental equipment.

It is possible that the introduction of picture tubes employing an electro-optical effect and eliminating all flicker (such as TITUS, developed at LEP [20]) will soon lead to the development of a new class of thermography equipment. Up to now the duration of a frame has been limited by the phosphor time constant, but

[19] A. C. Burton, The pattern of response to cold in animals and the evolution of homeothermy, in: Temperature, its measurement and control in science and industry, Vol. 3, Part 3, editor J. D. Hardy, Reinhold Publ. Co., New York 1963, pp. 363-371.

[20] G. Marie, Un nouveau dispositif de restitution d'images utilisant un effet électro-optique: le tube TITUS, Philips Res. Repts. **22**, 110-132, 1967. See also this issue, page 292.

these new tubes will remove this restriction — the only restriction that will remain being due to the phenomenon studied. Among other things, this will permit high-definition images to be obtained without reducing the temperature resolving power.

Extending the applications of thermography

Thermography may also have numerous applications in the scientific and industrial fields. Many phenomena are, in fact, closely related to thermal dynamics, and the knowledge of surface temperatures can, for example, permit the location of internal heat sites or the conditions of heat exchange with the environment.

The qualities required for the equipment will vary according to each field of application. The following are the most important characteristics: a high geometrical resolving power for the inspection of miniaturized or integrated electronic circuits, rapid analysis for dynamic thermal phenomena on models for supersonic wind-tunnels, a high sensitivity for the fields of plant and animal biology, a high observation distance for meteorology and aerial thermal mapping (where problems will be caused by the transmission of the atmosphere). Insofar as the range of temperatures is close to the ambient, the fundamental problems of thermography will, however, be similar to those mentioned above with regard to medical applications.

For this reason the equipment described should, in our opinion, be suitable in many cases for studying and determining the experimental conditions particular to one application or another and can thus contribute to defining the required technical characteristics of more sophisticated equipment.

Summary. LEP have developed a thermography equipment whose thermal resolution is better than 0.05 °C at 300 °K and which can operate in different spectral ranges (particularly 4 and 10 microns). It includes a special device for plotting isotherms to an accuracy better than 0.05 °C. The author first outlines the chief technical objectives which were used to determine the main requirements for the equipment design. He then describes the camera, which is equipped with a single-cell quantum detector (InSb for 4 μm, Ge(Hg) for 10 μm) and an opto-mechanical scanning system. The electronic device which allows the isothermal zones of the surfaces examined to be viewed directly on the screen of the display tube is briefly outlined. The possibilities which this method of representation opens up are discussed and illustrated with a few examples. The last part of the article is devoted to applications of thermography, particularly in the medical field. The results of numerous examinations carried out by the Hospices Civils de Strasbourg with a LEP thermography equipment allow a number of conclusions to be drawn. On the basis of this experience and the technical improvements that can still be made in the present camera, it is possible to draw up a specification for an improved version of the equipment.

Objective lenses for infra-red image formation

C. Hily

The objective lenses that have been developed at LEP are intended to be used for forming infra-red images on photosensitive surfaces. In this application the spectral region in which the optical system has to operate is determined by the wavelength regions for which the atmosphere is transparent, and by the threshold wavelength of the detector in use. Consequently, three spectral bands have to be considered:

1) Wavelengths from 1.2 to 2.8 μm. The detectors that can be used in this band are germanium detectors (threshold wavelength 1.8 μm), lead sulphide detectors (threshold wavelength 2.7 μm), indium arsenide detectors (threshold wavelength 3.4 μm) and a vidicon-type television camera tube [1].

2) Wavelengths from 3.3 to 5.5 μm. Many semiconductors, such as InSb, PbTe or PbSe, can be used in this band.

3) Wavelengths above 8 μm. Here thermal detectors like the bolometer can be used, or photodetectors (mercury-doped germanium, or mixed tellurides of cadmium with mercury, lead or tin).

The selection of the spectral region determines the choice of material for the objectives.

Generally speaking, there are very few materials which are transparent to infra-red and also suitable for use in optical systems; their chief characteristics have been given by Billard and Cornillault [2]. Conventional types of optical glass, all based on silica, are transparent up to 2.8 μm. Only glasses which have a high lead content and a very low silica content and those whose chief vitrifying constituent is germanium oxide (e.g. the SOVIREL glasses G0018 and VIR3) are transparent beyond 4.5 μm. Calcium aluminate glasses are transparent up to 6 μm, glasses based on arsenic trisulphide up to 11 or 12 μm, and the sintered glasses Irtran 1 and Irtran 2 (Kodak) are transparent up to 8.5 and 14 μm.

Crystals are also used. These include MgO, CaF_2, Si and Ge, whose transparency limits are 8.5, 12, 15 and 23 μm respectively.

1.2-2.8 μm band

The objective for the 1.2-2.8 μm band was designed for use with a vidicon-type television camera tube [1] and the image of the field considered has to cover the photosensitive surface. The optical system must therefore have an image field with a diameter of 25 mm and a fairly high resolution, of the order of 600 points per line, in television terms.

For the intended application (focal length 50 mm), the angular field had to be wide, and in addition to spherical aberration and coma, astigmatism and field curvature also had to be corrected. The objective therefore had to be composed from several elements. The solution found at LEP is based on Taylor's triplet: its two outer elements are convergent and the one between is divergent; in fact there are five lenses in the objective. The front element is made up of two lenses, one of VIR3 and the other of MgO; the middle element is formed by two fused silica lenses and the final element is a single lens made of a very dense flint glass. A photograph of the objective is shown in *fig. 1*, a diagram

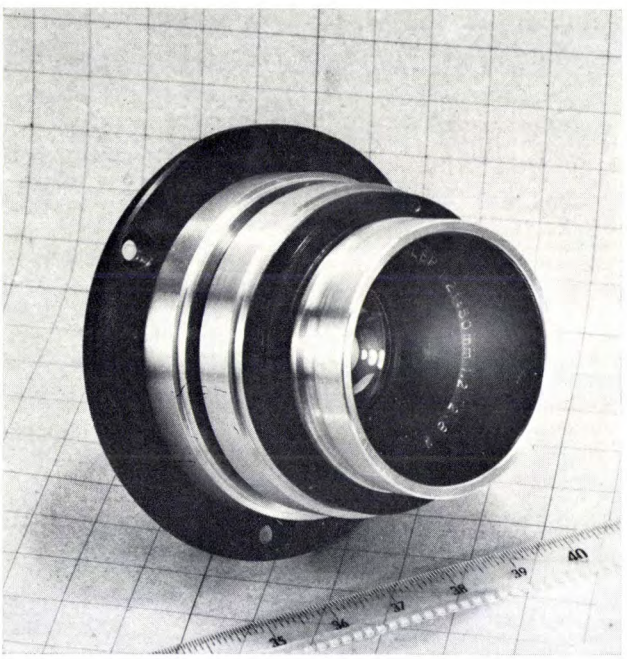


Fig. 1. Anastigmat consisting of five lenses, for the 1.2-2.8 μm band.

C. Hily, Ingénieur E.S.O., is with Laboratoires d'Electronique et de Physique Appliquée, Limeil-Brévannes (Val-de-Marne), France.

[1] M. Berth and R. Petit, Acta Electronica **10**, 123-135, 1966; see also M. Berth and J.-J. Brissot, page 270 of this issue.
[2] P. Billard, Acta Electronica **6**, 7-64, 1962.
P. Billard and J. Cornillault, Acta Electronica **6**, 73-169, 1962.
[3] See the article by F. Desvignes, J. Revuz and R. Zeida, page 264 of this issue.

in *fig. 2*. The curve representing the contrast transfer factor M as a function of the spatial frequency ν measured in the focal plane at the image side is given in *figs. 3a* and *b*.

3.3-5.5 μm band

The detector used for the 3.3-5.5 μm band is a linear array of cells of low resolving power [3]; each cell has a width of 200 to 500 μm.

This arrangement makes the problem much easier: a cementless Si-Ge doublet is adequate, and any curvature of the field can be corrected by a negative lens placed in the image plane.

A lens of this type has been developed at LEP; it has a focal length of 100 mm and an image-field diameter of 20 mm. The lenses are made from single crystals of Si and Ge and have a diameter of 75 mm.

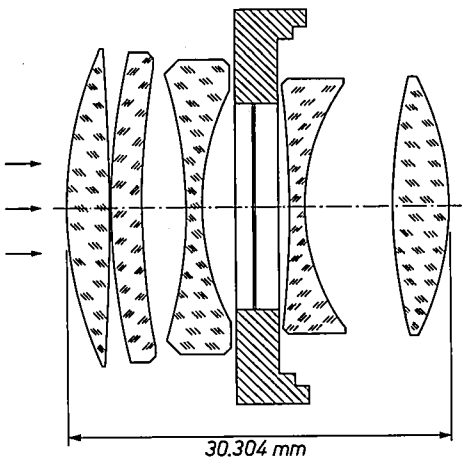


Fig. 2. Diagram showing the arrangement of the objective for 1.2-2.8 μm.

Band above 8 μm

LEP are at present developing refractive, reflective and mixed systems for the 14-16 μm band for incorporation in horizon sensors or infra-red simulators.

At present only germanium appears to be suitable for refractive and mixed optical systems in this band. Various types of optical systems are being made, such as Mangin eccentric mirrors, and collimators with

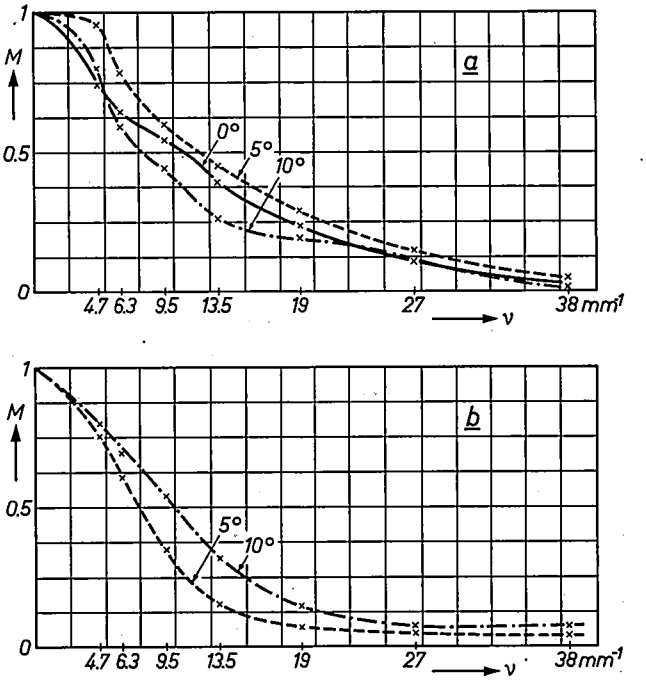


Fig. 3. Contrast transfer factor M of the objective as a function of the spatial frequency ν in the image plane, with focusing in the plane giving the highest value of M on the axis at $\nu = 25$ mm^{-1},
a) with the lines in the test picture perpendicular to the plane containing the optical axis and the central ray, for fields of view of 0°, 5° and 10°,
b) with the lines in the test picture parallel to the plane containing the optical axis and the central ray (field of view 5° and 10°).

lenses, made of single-crystal or polycrystalline germanium (some of these combinations have aperture diameters as large as 200 mm). Nevertheless, the performances of these optical systems are not so high as those of systems operating in the visible spectrum because high-quality correction of the various aberrations in the infra-red beyond 14 μm would require a combination of various materials which are not at present available.

Summary. The choice of materials for various infra-red radiation bands is examined, and the limitations imposed by the detectors with which the optical systems will be used are considered. Some examples of actual solutions are given: a five-lens anastigmat based on Taylor's triplet for the 1.2-2.8 μm band, a cementless doublet lens made from single crystals of germanium and silicon for the 3.3-5.5 μm band, and refractive and mixed (refractive-reflective) combinations using germanium lenses for the 14-16 μm band.

Large-screen projection of television pictures with an optical-relay tube based on the Pockels effect

G. Marie

The only solutions at present offered for the problem of reproducing television pictures in dimensions comparable with those of the cinema screen — all of them involving projection — are complicated and often very costly.

In the first part of this article, we shall examine a number of solutions proposed or already in use. In the second part, a real-time system will be described which is based on a new type of optical relay using an electro-optic effect. Its chief advantages are ease of operation and absence of flicker in the reproduced images.

Comparison between some existing projection systems

In this section we shall examine the following television picture projection systems, stating their performances and comparing their efficiencies:

1) The optical relay, which does not operate in real time, the images being recorded on film and then projected.
2) Projection from cathode-ray tubes.
3) Optical relays operating in real time and in which scanning is achieved either by a moving light beam (laser system) or by an electron beam (Eidophor).

Cinema-type projection

Projection of a television picture via a recording on film offers the advantage of a high efficiency which makes it possible to use very large screens (e.g. 150 m^2). This, however, is a costly solution because it employs special film material which cannot be erased and because the exposed film, generally speaking, will never be used again; moreover, it inevitably involves a time delay between the moment of transmission of the picture and the moment at which that picture is projected.

This second disadvantage can be overcome and the time delay reduced to less than a minute by using films, either of the conventional type (e.g. Kelvin-Hughes [1]) or of the vesicular type (e.g. Kalvar [2]), which can be developed rapidly. The cost of these films does not normally allow them to be used except in the case of short sequences. However, they are often used for the reproduction of slow-scan pictures, such as those obtained with radar.

In order to compare the performance of the various projection systems, we shall calculate the electrical power per unit of screen surface area necessary to obtain a maximum luminance of 28 cd/m^2, the average maximum value found in cinemas. If we assume that the screen reflects light in accordance with Lambert's law and with an efficiency of 80%, the necessary illumination is then 110 lux. The light source generally used is a high-pressure xenon lamp whose luminous efficiency can be as much as 33 lumens per watt. Since the individual efficiencies can be evaluated as 50% for the light-collection system, 66% for the maximum transparency of the film and the objective and 50% for the relative projection time, we obtain a consumption of 20 W/m^2 (*Table I*).

Projection by cathode-ray tube

Cathode-ray-tube projection was the first high-resolution system for the projection of television pictures [3]. The performance of this system, however, is limited by the maximum permissible dissipation on the tube screen. In one of the most highly developed solutions [4], a tube is used which can dissipate a mean electrical power of 25 W. The fluorescent screen of the tube radiates in accordance with Lambert's law, with a luminous efficiency close to 8 lm/W. The projection system, which uses large-aperture optics (Schmidt system), has an overall efficiency of the order of 25%. Taking into account the fact that the power consumption is proportional only to the mean luminance of the picture, which is generally about a third of the maximum luminance, we can assign a value of 3 to the "modulation efficiency" (Table I). From this, we deduce a consumption of 18 W/m^2 for a screen luminance of 28 cd/m^2. As the power which the tube can dissipate is limited (25 W in the example quoted), images can be projected only on screens with a surface area of a few square metres. With an arrangement comprising three tubes and three optical projection systems and using a directional screen giving a gain of approximately 3, it has been possible to project pictures in colour having a luminance of 20 cd/m^2 on a surface area of 7 m^2 [4]; however, it should be pointed out that the use of a directional screen obviously limits the useful field of observation and hence the number of viewers.

G. Marie, Ingénieur E.N.S.T., Dr. Ing., is with Laboratoires d'Electronique et de Physique Appliquée, Limeil-Brévannes (Val-de-Marne), France.

Optical relay using a laser beam

Optical relays have been developed in the earliest days of television; a light beam from a source was amplitude-modulated with the aid of a Kerr cell, then deflected by a rotating mirror before reaching the screen. With an arrangement of this type, however, it was not possible to reproduce high-resolution and large-dimension images because the light modulators and light deflectors could only accept beams of limited solid angle and cross-section.

The recent advent of lasers concentrating the entire radiation in a beam of very small solid angle and cross-section (e.g. diameter 1 mm, divergence 10^{-3} rad) now makes it possible to reconsider the use of such a device to obtain high-resolution pictures. Modulation can be achieved without difficulty by means of a Kerr or Pockels cell and, so far as the problem of deflection is concerned, it can be solved either by digital deflection or by continuous scanning using ultrasonic waves or rotating or vibrating mirrors [5].

Let us compute the efficiency of the system, choosing as a source the continuous-wave laser operating in the visible region for which the efficiency and the power output are the highest, i.e. the ionized-argon laser. When a strong magnetic field, greater than 0.1 tesla (1000 gauss), is applied to its plasma, the efficiency η can reach a value of 10^{-3} [6]. We list below the values of η and the relative eye sensitivity factor $\bar{y}$ as a function of the wavelength of the emitted spectral lines:

λ (nm)	η	$\bar{y}$
488	5.5×10^{-4}	0.18
514.5	3×10^{-4}	0.60
498.7	0.8×10^{-4}	0.28
476.5	0.7×10^{-4}	0.12

As the value $\bar{y} = 1$ corresponds to a luminous efficiency of 680 lm/W, the sum of the light energies of the various emission lines gives an overall luminous efficiency of 0.20 lm/W. Furthermore, we must take into account the efficiency of the modulator, which is about 75%, and also the relative projection time — 0.8 times 0.92 (20% loss due to line flyback and 8% loss due to frame flyback) — with the 625-line standard. The consumption necessary to obtain a luminance of 28 cd/m² then becomes 1000 W/m² (Table I).

The efficiency of this system is therefore much too low to permit projection on large screens. In the case of reproduction of colour pictures, one might consider the use of the ionized-argon laser for the green primary (emission line 514.5 nm) and the blue primary (line 476.5 nm and, though less suitable, 488 nm) and the He-Ne laser for the red primary (line 632.8 nm, efficiency 3×10^{-4} approximately); it will then be seen that the consumption is about 2000 W/m², i.e. 100 times higher than that for cinematographic projection.

Table I. Power requirements of different projection systems for a maximum screen luminance of 28 cd/m² (illumination 110 lux).

	Light efficiency of source in lm/W	Efficiency			Consumption in W/m²	
		optics	modulation	scanning	black-and-white	colour
Cathode-ray tubes	8	0.25	3	1	18	50
Cinema (with xenon lamp)	33	0.5	0.66	0.5	20	20
Eidophor (with xenon lamp)	33	0.5	0.5×0.5	0.5	53	80
Lasers (ionized argon)	0.2	1	0.75	0.8×0.92	1000	2000
TITUS (with xenon lamp)	33	0.25	$0.5\times0.75\times(0.8)^2$	1	55	80

Optical relay with electron-beam scanning (Eidophor)

The operating principles of the Eidophor were proposed by F. Fischer as early as 1939. The first practical experiments were performed in 1944 and development

[1] L. Corben, R. Lindstrom, A. Shepp and S. Hersh, in: Unconventional photographic systems, SPSE Symp. Washington D.C. 1964, pp. 166-168.
[2] R. T. Nieset and N. T. Notley, J. SMPTE **74**, 786-788, 1965.
[3] P. M. van Alphen and H. Rinia, Philips tech. Rev. **10**, 69-78, 1948/49.
[4] T. Poorter and F. W. de Vrijer, Philips tech. Rev. **19**, 338-355, 1957/58.
[5] S. M. Stone, 8th Nat. Symp. on Information display, San Francisco 1967, pp. 161-168.
[6] E. F. Labuda, E. I. Gordon and R. C. Miller, IEEE J. Quantum Electronics **QE-1**, 273-279, 1965.

of the system was completed in about 1950 [7][8].

Its operating principle is shown in *fig. 1*. The light from a cinema-type projection lamp *A* is directed to a concave mirror *C* coated with a film of oil *D*, after reflection on a mirror *B* made up from strips and placed at the centre of curvature of the concave mirror (Schlieren optics). When the film of oil is perfectly uniform, the concave mirror gives an image of the strip mirror

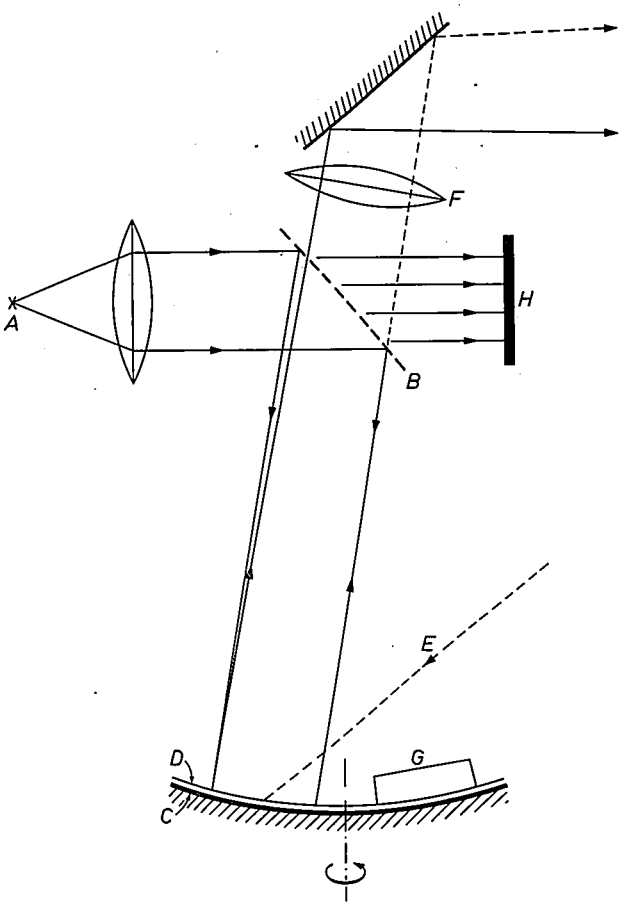


Fig. 1. Operating principle of the Eidophor optical relay; the Eidophor utilizes the deformations caused in a film of oil by the charges deposited by an electron beam. *A* light source. *B* strip mirror. *C* rotating concave mirror. *D* oil film. *E* electron beam. *F* projection lens. *G* knife-edge to smooth the oil film. The part of the light from *A* which is not reflected by the strips of mirror *B* is completely absorbed in *H*.

which is practically superimposed on itself. All the light which reaches the film of oil is then reflected towards the source. When the oil film is distorted, however, part of the light is transmitted through the strip mirror towards the objective and the projection screen. The distortion of the oil film is induced by electric charges distributed with spatially variable density and deposited on the film by an electron beam whose density is modulated at a relatively high frequency (e.g. of the order of 10 MHz). One can thus project a television picture, using the video signal to control the modulation depth of the electron beam. Erasure is achieved automatically with a suitable time constant by choosing a good compromise between the conductivity and the viscosity of the oil film, which has to be kept at a well defined temperature. This device shows the same flicker effect as conventional cathode-ray tubes and the laser projection system. To avoid permanent deformations, it is necessary to rotate the concave mirror slowly and to smooth the surface of the film with a knife-edge (*G* in fig. 1).

To calculate the consumption of the system, we may assume that the same light source and the same optical collection system are used as in the case of cinematographic projection. We must then take into account the fact that only 50 % of the light from the lamp is usefully reflected by the strip mirror *B* towards the oil film [9] and that, in the return direction, this mirror transmits at most 50 % of the light diffused by the oil film; besides, we may assume that the relative projection time is approximately 50 % because of the exponential decay of the deformations of the film, this decay being necessary to reproduce movement. The consumption necessary to obtain a luminance of 28 cd/m^2 is then 53 W/m^2 (Table I).

In order to project colour-television pictures, a triple projector has been constructed by adding to the device shown in fig. 1 two more electron guns and placing dichroic mirrors between the strip mirror and the concave mirror; the dichroic mirrors split the incident luminous beam into three primary components, red, green and blue, which are directed towards three different zones of the concave mirror; these mirrors reconstitute the beam from the three reflected light components. In practice, the power consumption of this arrangement is about 1.5 to 2 times larger than that of the black-and-white system (Table I).

The Eidophor is the only real-time optical relay at present in use which is able to project television pictures on to a screen with an area of several tens of square metres. Pictures exhibit a flicker effect that becomes increasingly disturbing as the picture area is increased. It has the further disadvantage of being expensive, bulky and not easy to operate. In fact, it usually has to be demountable to permit replacement of the cathodes, whose life-time is limited by the presence of a residual oil-vapour pressure, and it incorporates a pumping system; it also includes rotating and cooling elements in vacuum. Attempts are at present being made to build a sealed-off model. Nevertheless, the bulkiness and complexity of the system remain and this is why a number of laboratories have tried to develop devices of smaller size based on other physical phenomena, in particular on electro-optic effects.

Relay using the Pockels effect

Operating principle

In the optical relay, which will now be described, modulation of the beam is performed by using the Pockels longitudinal electro-optic effect exhibited by a monocrystalline plate belonging to the potassium dihydrogen-phosphate family. It should be noted that the idea of utilizing the Pockels effect for the projection of television pictures is not new, having been put forward as long ago as 1936 by L. S. Kaysie [10], M. von Ardenne [11] and V. A. Babits [12]. An attempt at the realization of such a device was described in 1960 by S. Rissmann and H. Vosahlo [13] and we know that work along these lines is at present in progress in several American laboratories. To our knowledge, this work uses crystals operating at ambient temperature and it seems at the present time not to have resulted in the development of an optical relay capable of projecting high-definition images.

It is known that when a crystal of KH_2PO_4 is placed in an electric field E parallel to its optical axis c, a difference Δn in refractive index occurs for light components whose directions of polarization are at right angles to each other. These components, initially in phase, show a phase shift φ upon leaving the crystal of thickness l. Since the effect is linear, Δn is proportional to the field E and φ is proportional to the product El, i.e. to the potential difference V between the input and output faces of the crystal. When the crystal is placed between two crossed polarizers, the resulting transmission is equal to:

$$T = \sin^2 \tfrac{1}{2} \varphi = \sin^2 kV;$$

in this expression the parameter k does not depend on the thickness of the plate. As long as T remains below 75%, the characteristic is practically quadratic and similar to that of conventional picture tubes.

The sensitivity of a crystal can be characterized by the voltage $V_{\lambda/2}$ for which the phase shift is 180° and therefore the transmission is maximum. This voltage reaches approximately 8 kV in the case of ordinary potassium dihydrogen phosphate and about 3.5 kV in the case of the deuterated crystal (KD_2PO_4). If light goes through the crystal twice and if transmission is limited to 75%, in order to remain within the quadratic region of the characteristic, the maximum voltage which is necessary for KD_2PO_4 is reduced to 1.2 kV. If the dimensions of the crystal plate are of the order of 30×40 mm and if one wants to obtain a resolution of 600 lines with 800 elements per line, two points which are 50 μm apart must be able to hold potentials which differ by approximately 1 kV. It seems that this result has not as yet been reached, either because of the occurrence of leakage currents at the surface of the crystal or because of the effect of these potential differences on the trajectories of the electrons, even in the presence of high acceleration voltages (e.g. 20 kV).

The TITUS tube

A special tube based on the Pockels effect, called TITUS ("Tube Image à Transparence Variable Spatio-temporelle"), has been developed at LEP. The operating principle has been defined in cooperation with Y. Angel, professor at the Conservatoire National des Arts et Métiers in Paris, and R. Genève at LEP, and already described in a previous article [14].

When the variations with temperature of the Pockels effect in a ferroelectric crystal are examined, it is observed that the induced birefringence is not proportional to the electric field but to the electric polarization, i.e. to the product εE when the crystal is operated above the Curie point. As the crystals of the KD_2PO_4 family are ferroelectric, we overcame the previous difficulties by operating the crystal at a temperature near its Curie point t_C. *Fig. 2* shows the variation with temperature of the dielectric constant of KD_2PO_4

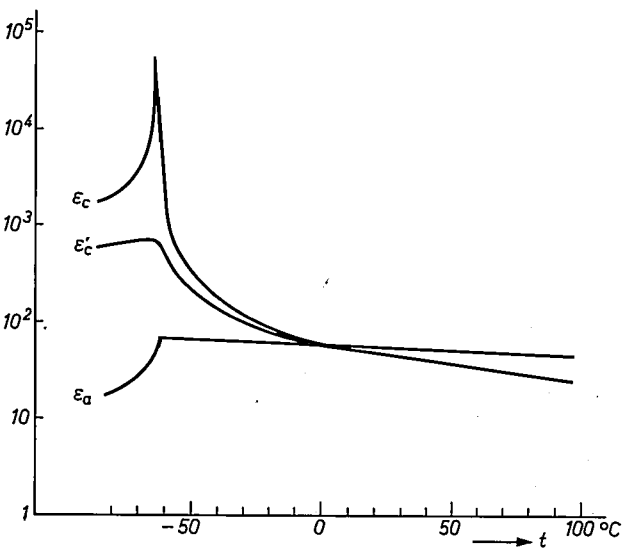


Fig. 2. Curves showing the variation of the dielectric constant of a KH_2PO_4 crystal with 90% of its hydrogen replaced by deuterium, as a function of the temperature t: ε_a in a direction perpendicular to the optical axis; ε_c and ε_c' in a direction parallel to the optical axis for a free crystal and a clamped crystal.

[7] E. Baumann, J. Brit. Instn. Radio Engrs. **12**, 69-78, 1952.
[8] E. I. Sponable, J. SMPTE **60**, 337-343, 1953.
[9] In the present stage of development the strip mirror has a "Venetian blind" structure permitting a gain of a factor of 2 in the efficiency. The power consumption for large projection screens will, however, be about the same as indicated in Table I, as the apparent larger size of the image of the light source results in a lower efficiency of the condenser system.
[10] L. S. Kaysie, in: Television and Short-Wave World, May 1936.
[11] M. von Ardenne, Telegraphen-, Fernsprech-, Funk- und Fernseh-Technik **28**, 180-184, 1939.
[12] V. A. Babits, Television Engineering, textbook, Rensselaer Polytechnic Institute.
[13] S. Rissmann and H. Vosahlo, Jenaer Jahrb. 1960/I, 228-244.
[14] G. Marie, Philips Res. Repts. **22**, 110-132, 1967.

for a field in a direction perpendicular to the optical axis (ε_a) and in a direction parallel to the optical axis; in the latter case it is shown both for a free crystal (ε_c) and for a clamped crystal (ε_c'). In fact, since the Pockels effect in a free crystal is accompanied by a piezoelectric effect which might cause coupling between the different points of the picture, it is necessary to clamp the crystal plate, e.g. by gluing it; one then observes that the constant ε_c' reaches a maximum of approximately 650 at a temperature near t_C and that no domains occur in the crystal for temperatures below t_C. The operating conditions may therefore extend from $t_C - 4$ °C to $t_C + 2$ °C, approximately, which makes temperature stabilization relatively easy. The maximum potential difference required for modulation of the light beam is now reduced to 80 V for double-transit operation [15]. It is also found that ε_a is about 10 times lower than ε_c', which makes it possible to choose a target thickness exceeding the distance separating two elements in the picture, without loss of resolution.

The choice of the operating temperature not only modifies the sensitivity of the Pockels effect but also makes it possible to give the optical relay an important additional feature: it can reproduce television pictures without flicker, however low the scanning speed. Indeed, the discharge time constant of the crystal (equal to the product of the dielectric constant and the resistivity) becomes about 1 hour at —60 °C, whereas it is about 0.1 second at ambient temperature. In our case, this phenomenon raises the problem of removal of the charges deposited by the electronic beam, while erasure can be considered to be automatic in the case of operation at ambient temperature. For the removal of the charges we have made use of secondary emission from the target by employing an arrangement already used in a few storage tubes. This arrangement is shown in *fig. 3*, which also illustrates the principle of double-transit operation with the aid of a multilayer dielectric mirror *M*. The target is bombarded by an electron beam whose acceleration potential is of the order of 500 to 1000 V, so that the secondary-emission yield at saturation of the bombarded face is greater than 1; a grid *G* is placed in front of the target, a short distance away from it — about 50 μm. It can be shown that, under these conditions, the equilibrium potential reached by the point of impact of the electron beam is a few volts higher than the potential of the grid; the electron beam, which has a constant intensity, then works practically as a "flying" short-circuit between the grid and the point of impact on the target. It is therefore sufficient to apply the video signal between the grid *G* and the transparent conducting layer *E* to ensure that the various points on the target get charged to the corresponding video voltage when they are hit by the beam. The erasure and writing functions are therefore combined and the result is an entirely flicker-free operation; indeed, the charge at each point remains constant between two successive scans of the electron beam as well as during these scans, for points which have not moved on the image.

Power requirements of the projection system

The KD_2PO_4 crystals belong to the tetragonal system and have an optical axis; therefore, when they are placed between crossed polarizers, one can obtain a total extinction of the transmitted light only for beams parallel to the optical axis. Operation with a light beam that has a certain amount of divergence introduces a level of unwanted light which limits the contrast. It has been calculated that, with a crystal plate having a thickness of 0.2 mm and operating in a double-transit system, it is possible to obtain a contrast of 50 for a beam whose aperture angle is about 10°. The solid angle of the light beam which can be used is therefore more limited here than in the case of cinematographic projection. For a xenon-arc type of light source, whose efficiency is 33 lm/W, and a target area of 30×40 mm², the efficiency of the collecting system is thus reduced to about 25%. One must also take into account:

a) the fact that we are working with polarized light, introducing a 50% loss of light (when polarizing prisms, without dissipation, are used);
b) the maximum transmission of 75% if it is desired to keep within the quadratic region of the modulation characteristic;
c) the transmission coefficient of about 80% of the transparent conducting layer, crossed twice by the beam.

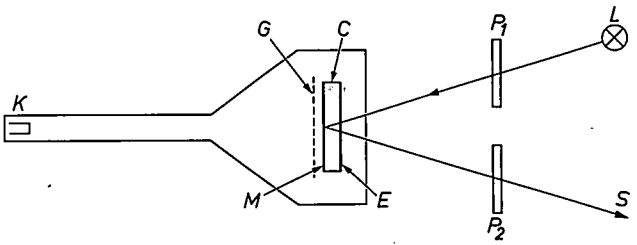


Fig. 3. Basic structure of the TITUS tube. *C* monocrystalline plate of KD_2PO_4 exhibiting Pockels effect. *M* multi-layer dielectric mirror. *G* ground-potential collector grid placed a short distance away from the crystal. *E* transparent conducting electrode receiving the video signal. *K* cathode at a potential between —500 and —1000 V, which delivers an electron beam of constant intensity. For practical purposes the electron beam can be regarded as a flying short-circuit between the grid and the target. Also shown in the figure: external light source *L* and crossed polarizers P_1 and P_2; the light is projected on to the screen *S*.

[15] This value does not include the potential-dividing effect introduced by the presence of dielectric coatings deposited on the faces of the crystal.

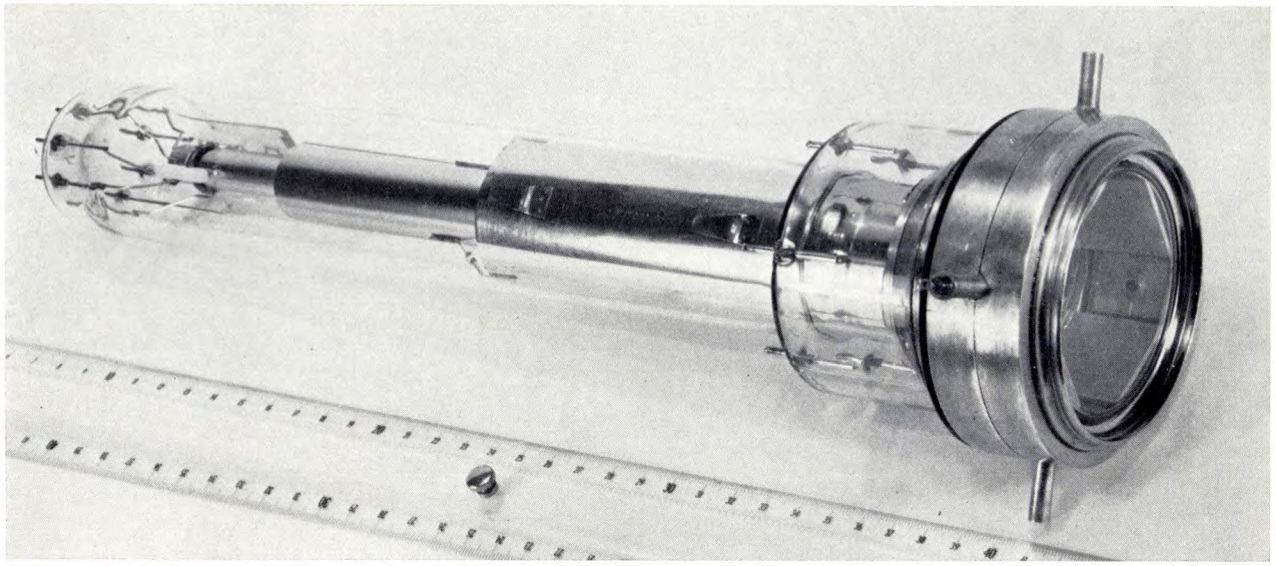


Fig. 4. Experimental TITUS tube incorporating a cooling system consisting of two stages of Peltier cells; the latter are partly visible through the window.

For an illumination of 110 lux of the screen (Table I), we thus obtain a consumption of 55 W/m^2, a value very close to that found in the case of the Eidophor. As in that case, the consumption is increased by a factor of 1.5 to 2 when colour pictures are projected, which can be done very simply using three relay tubes and two dichroic mirrors (as shown by the dashed portions in fig. 5).

First experimental results

The experimental realization of the optical-relay tube gave rise to many technological problems, related in particular to operation with polarized light and the need to cool the target uniformly to a temperature close to —60 °C. These problems were solved in cooperation with J. Donjon and R. Le Pape at LEP. *Fig. 4* shows a tube having the dimensions of a 3-inch image orthicon in order to enable conventional deflection and focusing coils to be used; cooling of the target is achieved with two stages of Peltier cells incorporated into the tube.

The experimental arrangement is shown in *fig. 5*. The light from the xenon lamp is condensed by means of a concave mirror on to a polarizing calcite prism of the Glazebrook type which transmits to the tube only the light component whose electric vector is parallel to the plane of the figure. The projection lens is placed between this prism and the tube so that the luminous beam incident on the target has a mean direction which is normal to the latter; when the light beam is reflected and passes through the lens and the prism again, only the light component with its electric vector perpendicular to the plane of the figure is transmitted to the screen.

The first laboratory experiments consisted in pro-

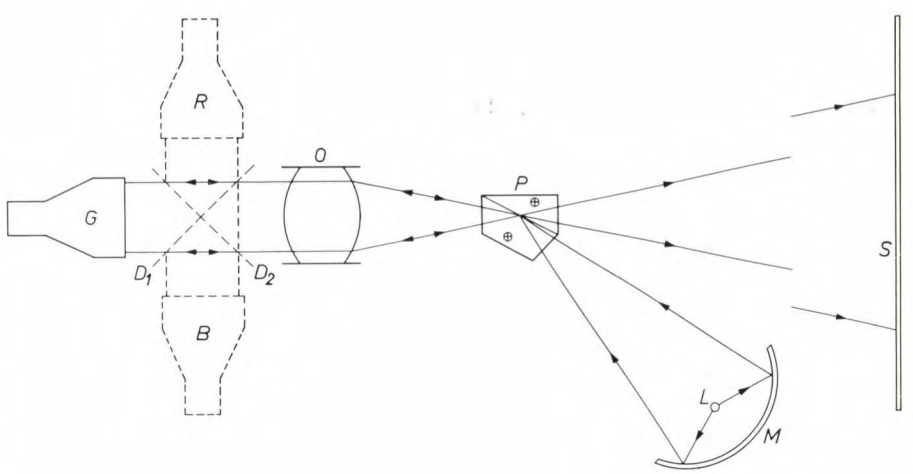


Fig. 5. Projection system with TITUS tube for black-and-white pictures (solid lines). The figure also shows, in dashed lines, the extra components that have to be added for the projection of colour pictures. *R*, *G*, *B* three identical TITUS tubes for red, green, and blue. D_1, D_2 dichroic mirrors. *O* objective. *P* polarizing beam-splitter. *L* light source. *M* concave mirror. *S* screen.

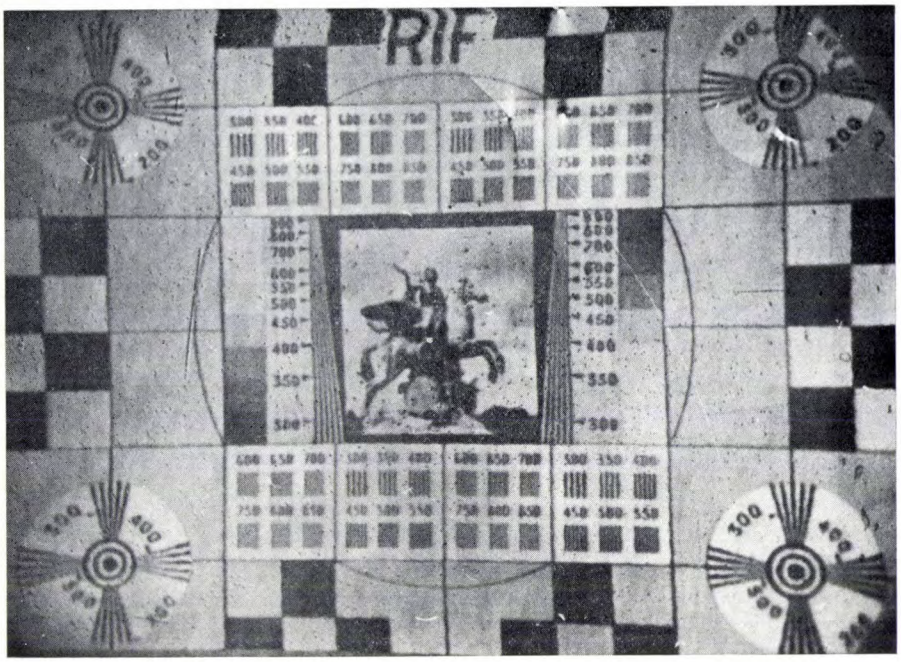


Fig. 6. Photograph of a picture projected on a screen with a surface area of about 1 m^2. Printed reproduction does not show the real resolution of the tube, which is in fact as high as 750 elements per line.

jecting a picture on to a screen with a surface area of approximately one square metre; the resolution (750 elements per line) and the contrast obtained are satisfactory, as can be seen from the photograph in *fig. 6*.

Advantages and applications

The optical relay we have just described is a simple sealed-off tube of relatively small dimensions. Compared with existing systems of the Eidophor type, it has the advantage of being very easy to operate. Moreover, the pictures do not flicker, which is particularly important in the case of large-screen projection and which might result, for professional applications, in a reduction by almost a half of the passband needed for a television channel.

The applications of this optical relay are not restricted to conventional television. Indeed, the absence of flicker makes it particularly well suited to the reproduction of slow-scan images such as those obtained in infra-red television with mechanical scanning (one image per second approximately) or in the case of radar pictures (a few pictures per minute). Finally, it should be noted that the memory features of the tube (storage time of the order of one hour) and the feasibility of simultaneous writing and erasing — localized if needed at precise points of the image — will also make it suitable for application as a computer peripheral unit.

Summary. After recalling the problems involved in the projection of television pictures on large screens the author describes the solutions known at present: recording on film and projection, projection from cathode-ray tubes, use of lasers, and the Eidophor. The latter is the only real-time optical-relay system at present in use, although it has some drawbacks due to its bulkiness and complexity. Simpler solutions are possible by using a new type of optical relay based on the Pockels effect in a KD_2PO_4 crystal. Different varieties are under study in Europe and the United States. The characteristics of a relay tube made at LEP and known as TITUS are examined in greater detail. In this tube the crystal is brought to its Curie temperature (—60 °C) and one of its most important features is that it provides a flicker-free picture. First results show that satisfactory resolution and contrast can be obtained.

Transparent windows for a wide spectral range

R. Legoux

Modern photodetector elements of both the photoconductive and photoemissive types are often used in evacuated tubes, and for their efficient use in the detection of electromagnetic radiations it is often necessary to have windows whose optical transmission extends considerably on both sides of the visible spectrum.

Although glass, which is inexpensive and easy to use, is employed on a large scale in the manufacture of electronic tubes, it has a rather limited transparency region. It is therefore replaced by materials, generally refractory, which are transparent in the spectral range to be transmitted.

These materials, usually single crystals, are chosen from natural or synthetic crystals such as corundum, quartz, fluorite, lithium fluoride, periclase, etc., or even from substances with a simpler crystalline structure such as germanium and silicon. However, special techniques have to be used if these materials are to be used for vacuum-tight windows in enclosures where the residual pressure has to be maintained at 10^{-10} torr or less for periods that may extend to several years.

In fact, unlike glass, most of these materials cannot be directly sealed to the glass or metal walls of enclosures. The techniques that have to be employed for sealing such windows are very much like those employed in modern industry for brazing metals or ceramics. The same rules and requirements apply, and we can quote some of them:

a) The solders must not contain metals with high vapour pressures such as cadmium, zinc, manganese, and lead. This is because the soldered parts will be in a vacuum and will be subjected to thermal processes at relatively high temperatures after sealing.

b) For the same reasons, etchant fluxes which are usually soluble in the solder and improve the wetting of the workpieces cannot be used. The soldering procedure must therefore be carried out either in a controlled (neutral or reducing) atmosphere, or in a vacuum. Soldering in a vacuum has the advantage that the pieces to be joined are degassed to a certain extent and therefore certain undesirable reactions between the workpieces and the gas (or impurities in it) are minimized.

In practice, the window to be soldered in is mounted flat on its support or, more generally, located in a recess machined in the support and which it fits exactly without any play (*fig. 1*).

R. Legoux is with Laboratoires d'Electronique et de Physique Appliquée, Limeil-Brévannes (Val-de-Marne), France.

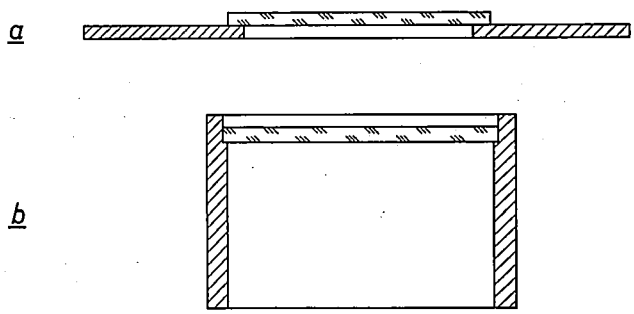


Fig. 1. Two types of window seal. *a*) Window mounted flat on support. *b*) Window mounted in an accurately machined recess.

For a given type of window the material and exact shape of the support are determined by technological considerations based on some of the physical properties of the pieces to be assembled, in particular the coefficient of expansion, the modulus of elasticity, and the thermal conductivity. Other factors that are important are the brittleness of the window material and the metallurgical compatibility of the solder with the window and its support.

The coefficients of expansion are usually matched as closely as possible because this gives seals which are more reliable, free from stress and easier to make. In cases where it is impossible to match the coefficients of expansion it is often advisable to use a support with an expansion coefficient which is slightly higher than that of the window and to effect the seal by means of a machined recess. The resulting compression seal is vacuum-tight but has the disadvantage that the window is under permanent mechanical stress. Apart from making the soldered parts more liable to shatter, this stress can introduce undesirable polarization in radiation transmitted through the window.

In these extreme cases the designer is obliged to make use of the elasticity of the support, which is given a suitable shape, e.g. a feathered wall, to make it sufficiently flexible to remain within the allowable tolerance.

Finally an important point for obtaining a good seal is the choice of the solders, which are either simple metals or alloys. The choice is determined by the various physical and chemical properties of the solders and the materials to be joined; by the various metallurgical reactions which develop between the solder in the liquid state and the materials (such as solubility of the materials in the solder, diffusion of the solder into these materials and formation of intermetallic com-

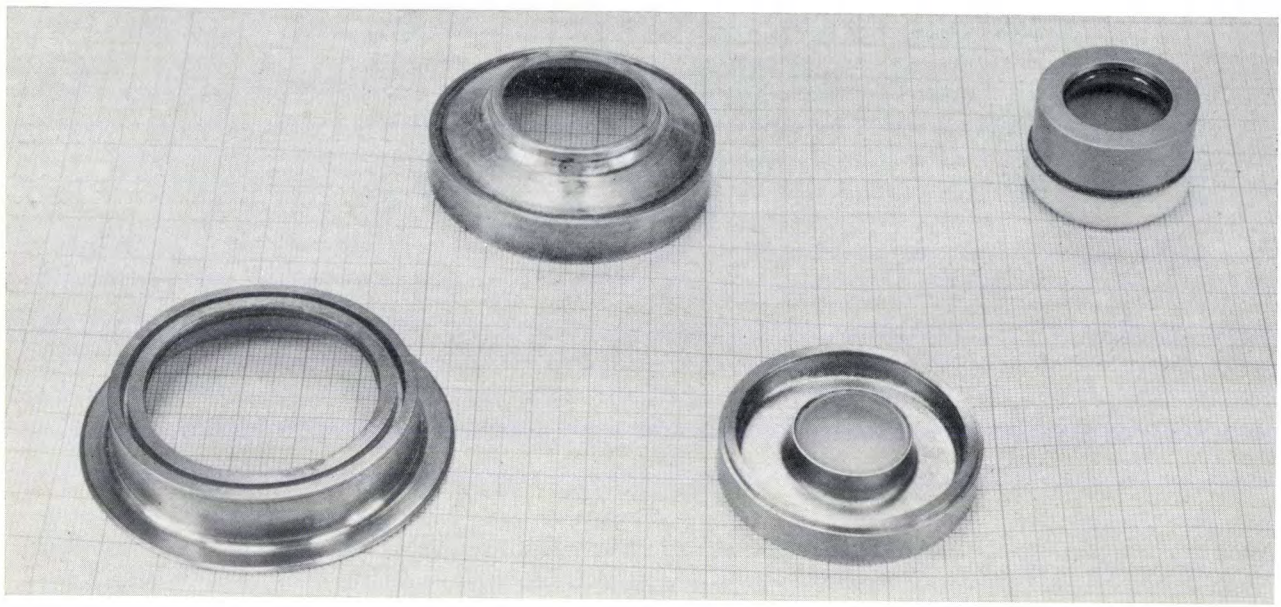


Fig. 2. Sealed-in windows of different materials. From left to right: sapphire, CaF_2, MgO, SiO_2.

pounds); and by the operating conditions of the windows: temperature, atmosphere, etc.

A consideration of all these factors gives two main rules which the solder should satisfy. First, its mechanical properties should be at least as good as those of the parts to be joined. Secondly, the wetting of these parts should be adequate: this is a very important factor which mainly depends on the choice of solders and on the physical-chemical state of the parts to be joined.

Sealing processes have been developed at LEP for windows for a very wide range of applications. *Fig. 2* shows some windows that have been made with various kinds of the more commonly employed materials. These seals have good mechanical characteristics and can readily withstand the heat treatments normally applied in making vacuum devices. The seals can be fitted to the devices or equipments for which they are intended by hard or soft soldering.

Summary. The transmission of electromagnetic radiation often necessitates the use of transparent windows with an extensive transmission range on either side of the visible spectrum. The optical transmission of glass is limited, and it has to be replaced by materials such as corundum, quartz, fluorite, silicon, germanium, etc. LEP have developed sealing processes for such windows. These seals, with matched coefficients of expansion, have considerable mechanical strength and can readily withstand the heat treatments used in making vacuum devices.

An outside-broadcast equipment with mobile television cameras

J. Cayzac

Introduction

In 1960 the Laboratoires d'Electronique et de Physique Appliquée, in close cooperation with the technical department of the French Television and Radio Broadcasting Organization O.R.T.F., developed a prototype portable equipment comprising a light-weight television camera and a completely self-contained UHF transmitter.

J. Cayzac is with Laboratoires d'Electronique et de Physique Appliquée, Limeil-Brévannes (Val-de-Marne), France.

Operation of this equipment by O.R.T.F. provided the knowledge and experience on which to base the design of a second and more highly developed system. This consists of a number of identical independent mobile units linked with a fixed unit installed in an outside-broadcast van. This system, with the exception of the aerial control unit, is produced by T.R.T. (Télécommunications Radioélectriques et Téléphoniques). The principal differences between the first and second equipments are as follows:

1) For a considerable part of the operating time the camera will be used alone, connected by coaxial cable to a video tape recorder or to an outside-broadcast van. The new camera has therefore been designed as a completely self-contained unit incorporating its own power supply. This simplification created one difficulty, that of increased weight. This has been overcome by making almost exclusive use of integrated circuits; the new camera weighs 7 kg.

2) Because it is necessary to take simultaneous shots of the same event from different positions, several mobile units must be in use simultaneously, and it must be possible to superimpose pictures on each other. These requirements make it necessary to synchronize the time-bases of the mobile units. A special device for this function has been developed which synchronizes the time-base generators of the various mobile units with that of the central relay station.

3) Distortion of the signal is introduced by the UHF links between the mobile units and the relay station and this must be corrected. The distortions are most noticeable when a mobile unit changes its position. An improvement in performance is obtained by using automatically trained directional aerials.

The UHF link transmitting the vision signal produced by a mobile unit to the outside-broadcast van operates in the 1500 MHz band. The fixed unit installed aboard the van receives the vision signals and generates the error signals required for synchronization of the various mobile units. A multichannel control link operating at 32 MHz is used to transmit the signal for automatic training of the aerials and synchronization signals from the van to the mobile units.

The new equipment comprises three mobile units and a three-channel unit, installed in the outside-broadcast van, which synchronizes the mobile units.

Each mobile unit is made up of the following four self-contained units: the camera, the 1500 MHz transmitter, the "intersynchronization" unit, the aerial control unit surmounted by the automatically aimed directional aerial. The last three units can be fitted in a harness worn on the back, while the camera, which is linked to the equipment in the harness by a cable, is held on the operator's shoulder (*fig. 1*).

Fig. 1. Portable equipment for mobile unit. The camera, which is held on the operator's shoulder, weighs 7 kg. The equipment in the harness comprises the UHF transmitter, the intersynchronization unit and the aerial control unit with the encased directional aerial on top of it.

The fixed equipment installed aboard the outside-broadcast van comprises three vision receivers, a three-channel comparator, a three-channel signal encoder, the multichannel 32 MHz transmitter for the return link with the mobile units and a three-channel aerial control unit.

Overall design

We shall now examine the operation of the system with reference to the block diagram in *fig. 2*.

Each mobile unit has its own time-base generator and produces a complete television signal. When the camera is used without the "intersynchronization" device, the time-base is derived from an internal crystal-controlled circuit. When the "intersynchronization" device is in operation, the time-base is driven by an external signal controlled in frequency and phase, the line and frame phases being locked to those of the signals from the external time-base generator. The output signal from the camera is at the standard level (1 volt into 75 ohms) and is used to frequency-modulate the UHF transmitter which operates in the 1500 MHz band. The maximum frequency deviation is 8 MHz peak-to-peak and the output power is 1 watt.

The video signals are received in the outside-broadcast van on three receivers. The line and frame phases are compared with those of a reference signal (*synchr. ref.* in fig. 2). The comparison process results in the generation of error voltages which are then encoded. The resultant digital signals are mixed with the return signals, including the three aerial-control signals. The 32 MHz control link transmits the following signals to the mobile units:

1) The four line and frame-phase correction signals together with an aerial-connection signal for switching an indicator lamp ("tally light", individual to each mobile unit).

2) Instructions to the cameramen on an audio channel (common to the three mobile units).

3) The aerial control signal (individual to each mobile unit).

Since the mobile section of the system is made up of four completely self-contained units, it may be adapted to suit various operational modes other than that shown in fig. 2. The most important ways in which the equipment can be used are as follows. The camera can be used alone, the video signal provided by the camera for example being recorded on video tape. The synchronized mobile equipments can be used with a coaxial-cable link; for this purpose a device is incorporated which conveys the video signals in one direction and the intersynchronization signals in the reverse direction on the same coaxial cable. Finally, the intersynchronization device can be used alone to synchronize the time-base generator in the outside-broadcast van with that of a central relay station.

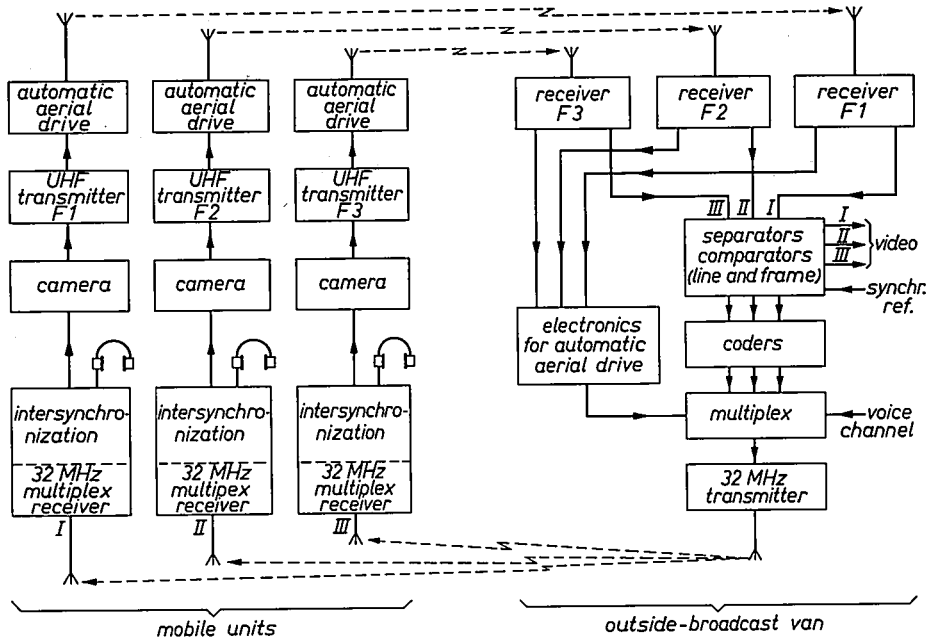


Fig. 2. Block diagram of the entire system, comprising a fixed unit in the outside-broadcast van and three mobile units.

Operating principle of the intersynchronization device

The equipment for synchronizing the signals from the various mobile units must meet the following specification.

a) The line and frame signals must be locked in frequency and phase at the point where comparison with the reference signal takes place.

b) The line locking has to be controlled accurately to within $\pm$ 200 ns under the most exacting operating conditions.

c) The pass-band necessary for the transmission of the intersynchronization signal for one camera must not exceed a few kilohertz.

The solutions adopted are now briefly described.

Line intersynchronization

In the solution chosen to achieve line intersynchronization, the initial frequency difference is corrected relatively slowly. Small line-phase fluctuations occurring in the course of operation are corrected by a high-speed digital control. Correction increments or decrements of 100 ns are made with a maximum frequency of 50 per second.

This device is controlled by means of three signals: "advance", "retard" or "stop". These are transmitted over the return link in the form of bursts of two subcarriers, each of which can assume the "0" or "1" state, at a rate of 50 per second. These same signals are used to keep the frequency control in its balanced position. To begin with, before frequency synchronization has been achieved, an inhibiting device makes it possible to lock the system on the "advance" or "retard" signal, depending on the sign of the frequency error, until the error has been sufficiently reduced to be corrected by the high-speed system.

We will assume that the oscillator of the mobile unit is in synchronism with the central reference oscillator (*synchr. ref.* in fig. 2) but that the phase is not correct. If T_1 represents the line cycle, it is possible to correct the phase by introducing an extra delay t such that the total delay is a multiple of T_1. The extra delay required will change when the mobile unit changes position. Thus, if $\tau_1 \ldots \tau_n$ are the fixed delays in the system, t must be adjusted so that $NT_1 = \tau_1 + \ldots + \tau_n + t$, where N is an integer, and when the mobile unit is moving, t must be continually adjusted to keep N integral. To be able to adjust the value of t at any moment the system uses a digital phase shifter with a permanent memory which can vary t in steps of (approximately) 100 ns (corresponding to one picture element in the television picture). This device retains the previously chosen value of t indefinitely when the phase comparator in the outside-broadcast van transmits the stop signal or if the link is cut. The error is measured every 20 milliseconds (frame cycle T_f) in this comparator. If it is less than 100 ns, no correction is applied; if it exceeds 100 ns, a correction signal is sent to the mobile unit and t is changed 100 ns in the appropriate sense. Since this correction occurs every 20 milliseconds, any error in the total delay τ equal to or greater than 100 ns is corrected within a time $(\tau/100)\times T_f$. To avoid over-correction the threshold of sensitivity of the system has been adjusted to a value slightly above 100 ns.

This method of correction is stable provided the UHF link is not too long and the velocity of the mobile unit is not too high. In fact the time for propagating the correction signal must not exceed 15 milliseconds, which is equivalent to a UHF link more than 2000 km long; and the value of 100 ns chosen for the correction increment or decrement makes it theoretically possible to counteract the phase variation due to the Doppler effect for velocities of the mobile unit up to 5400 km/h; in practice, owing to frequency jitter of the oscillators, this value is reduced to about 2500 km/h.

In the event of momentary failure of the 32 MHz link, the phase shifter interprets the absence of a signal as a "stop" order. The controlled oscillator has a very long response time and it retains an acceptable phase error for 10 to 20 seconds.

The permanent-memory phase shifter operates in the following way (*fig. 3*). A series of phase-shifting networks giving increasing phase shifts φ_0, φ_1, φ_2, ... φ_n are incorporated in the output circuit of the reference oscillator of a time-base generator whose frequency — as usual in television transmission — is double the line frequency ($2F_1$). These circuits are connected to an electronic commutator whose direction of rotation can be reversed. Each phase step corresponds to (approximately) 100 ns. When the switch has completed a whole cycle, it returns to φ_0, having effected an advance or

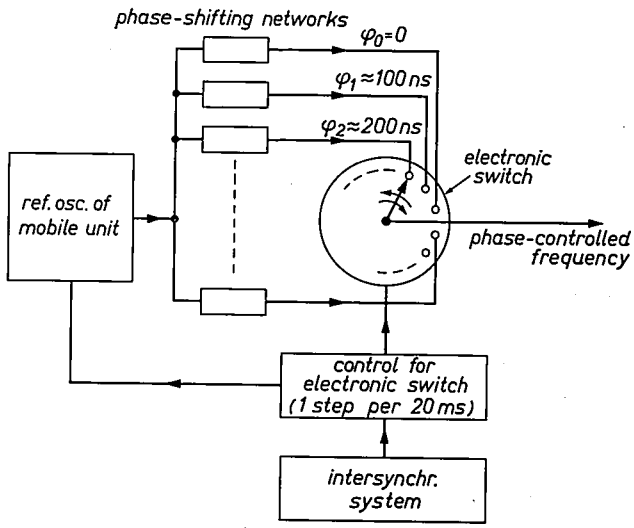


Fig. 3. "Phase wheel" system for high-speed digital control of the time-base frequency in each mobile unit. The electronic switch driving the "wheel" advances at the rate of 1 step per 20 ms. Instructions from the intersynchronization system cause the phase wheel to turn in the required direction. Phase correction of the frequency of the time-base reference oscillator at a rate of (approximately) 100 ns per 20 ms is thus achieved.

For the usual reference frequency $f = 2F_1 = 32$ kHz a number of 320 phase-shifting networks would be required to produce the 100-ns phase steps. The number is reduced to 3 networks by using a reference frequency f about 100 times higher as explained in the text.

retardation of 2π. If it goes on turning continuously, it thus modifies the frequency of the control oscillator continuously.

If this "phase wheel" system really operated at a reference-oscillator frequency of $2F_1$ as mentioned above, with a 32 μs cycle time, as many as 32/0.10 = 320 phase-shifting networks and a 320 position commutator would be necessary to cover all possible retardations with an accuracy of 100 ns. This would be very impractical. However, since the system merely checks the phase error every 20 ms and provides 100-ns steps of retardation or advance at this rate, the oscillator frequency can be chosen anywhere within a certain range. It was therefore decided to multiply this frequency by a factor k. This factor divides the number of networks which are necessary to define a cycle and multiplies the number of rotations which the reversible commutator must make in order to correct a given transit-time variation. The maximum value of k is that which gives only *three* networks, with a 120° phase difference relative to each other (for two networks the direction of rotation would be ambiguous).

This phase-correction device is capable of correcting frequency differences but an automatic *frequency precorrection system* has to be combined with it because the phase correction which is limited to about 100 ns every 20 milliseconds is insufficient to correct the local-oscillator drift. (This could be improved by increasing the sampling frequency to give more rapid correction, but the allowable propagation time would be reduced in the same ratio.) Moreover, to correct a frequency error permanently, the reversible switch has to turn continuously, which introduces an annoying 100 ns jitter. Owing to the precorrection system a 100 ns jump needs to take place only every 20 or 30 seconds, which is not very noticeable.

It should be noted that incorporation of the frequency precorrection system also increases the drift speed limit tolerated by the high-speed corrector. This changes from about 2500 km/h to 12 000 km/h, a velocity which is only likely to be of interest in very exceptional cases.

Frame control

Since the frame-recurrence signals are obtained by division of a frequency which is double the line frequency F_1, a device has been designed which corrects the shift of the frame-recurrence signals, at the rate of two lines per frame, by appropriately changing the factor by which the counter divides.

The counter control data are obtained with the aid of a comparator which provides two binary signals X and Y for each intersynchronization channel as a result of the comparison with the reference signal (*synchr. ref.* in fig. 2). The combination of these two binary signals determines the factor by which the counter should divide in accordance with the following code (Tm = frame-recurrence signals of the mobile camera, Tr = frame-recurrence signals of the reference source):

X	Y	counter divides by:	direction of shift of Tm in relation to Tr
1	1	625	in coincidence
0	0	625	in coincidence
0	1	623	Tm lagging Tr
1	0	627	Tm leading Tr

The stop condition for frame synchronization is defined by the states (0,0) and (1,1). The difference between these two states is used to transmit the "aerial connection" signal from the fixed to the mobile unit. Thus in state (0,0) the system divides by 625 and a "tally light" is switched on to indicate that the signal is going out on the aerial, while in state (1,1) the system again divides by 625 but the tally light is extinguished and the signal is not applied to the aerial. For any other combination the tally light is extinguished because the equipment is unable to transmit synchronized pictures.

Principle of aerial training device

It has been mentioned above that the transmitting aerial is trained automatically in order to reduce the reflected signals. This is achieved in the following way. Imagine a directional transmitting aerial which may be turned so that the lobe of the radiation pattern may point in two different directions A or B, as shown in *fig. 4a*. Consider a receiver sited successively at R_1, R_2,

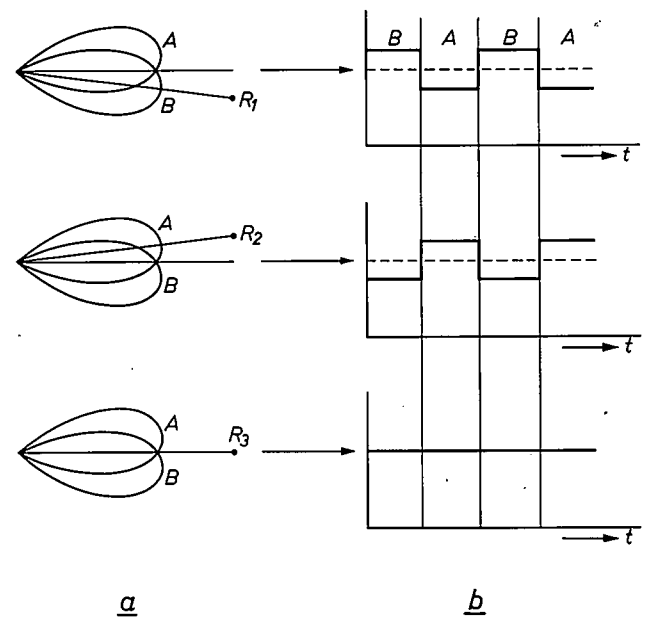


Fig. 4. Diagram explaining the device for automatically training a transmitting aerial. *a*) Positions R_1, R_2 or R_3 of a receiver with respect to two alternating orientations A and B of the lobe of the aerial's radiation pattern. *b*) Time-variation of the amplitude of the signals received with the three different positions.

R_3. The received signal at R_1 is stronger when the beam is in position B. At R_2 reception is stronger in position A. Finally, at R_3, since the receiver is on the axis of symmetry of the figure, the received signal is the same for beam direction A and beam direction B. Let us now suppose that the beam is switched at some low frequency between directions A and B. The strength of the signal received at the three receiver positions is shown as a function of time in the diagram in fig. 4*b*. It will be seen that the received field strength is constant if the receiver is on the axis of the figure formed by the two beams, but undergoes amplitude modulation if the transmitter is aimed so that the receiver is not on this axis. Moreover, the phase shift of this amplitude modulation in relation to the beam switching signal differs by 180° depending on the direction of misalignment.

When the received signal is detected in the outside-broadcast van, we therefore obtain an alternating signal which disappears when the aerial is correctly aimed and whose phase is reversed when it passes through zero. The training signals to the transmitting aerial are derived from this signal and transmitted over the 32 MHz circuit. A synchronous demodulator incorporated in the mobile unit provides the correction voltage required to drive the motor which adjusts the position of the aerial. To provide the two alternating beams the aerial consists of two radiators, one above the other, fitted with deflectors orientated along different axes. A switching diode allows each radiator to be excited in turn.

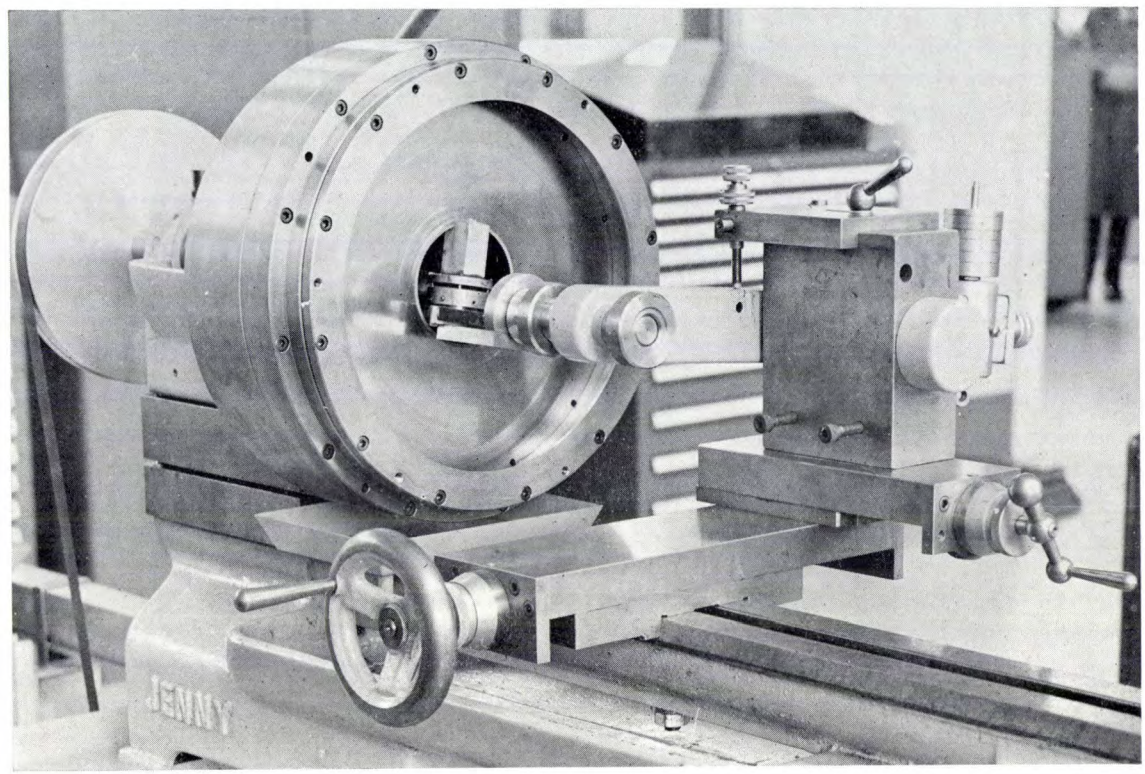
A final important point in the design of this system is the method by which the most favourable direction is found when the link is first set up. To select the most suitable path the control system causes the aerial to rotate through an angle of 360°; during this rotation it analyses the amplitude characteristics of the received signals and registers the angle at which the best reception is obtained. On the next rotation it causes the servomechanism to lock the aerial to the direction selected.

Final remarks

The usefulness of these self-contained units, whose standardized outputs enable them to be connected to a wide range of equipments, has been clearly demonstrated in the course of the first few months of operation by O.R.T.F. By using integrated circuits and micro-miniature techniques a system has been designed which is much more versatile than the first-generation equipment and is no larger or heavier. For example, the time-base generator of the original equipment occupied a volume of 1000 cm^3, whereas the new generator only occupies 170 cm^3. An additional feature of this equipment is that some of its units (the transmitter, the receiver and the "intersynchronization" device) can be used in a colour-television link.

Summary. A new self-contained television outside-broadcast equipment is described, with particular emphasis on the novel features of operation. The equipment consists of one or more mobile camera units which are linked to the outside-broadcast van by a 1500 MHz transmitter. The system employs portable self-contained cameras; an entirely digitalized intersynchronization device for locking the time-bases; an automatic system for training the aerials which virtually eliminates the effect of reflections; and a transmitter operating in the 32 MHz band with a number of channels for transmitting control signals from the van to the mobile units.

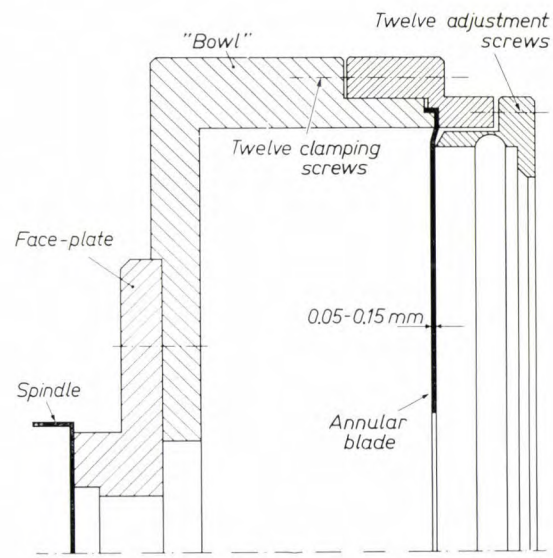
Bowl cutter machine



The cutting of brittle materials like semiconductors sometimes presents problems if the loss of material has to be kept to a minimum and the crystal lattice has to be disturbed as little as possible. To meet these requirements, A. Charles-Georges and A. Salmon of LEP have perfected a bowl cutter machine [1].

An annular metal blade (see diagram) is clamped by a number of screws to a "bowl" support. The blade can then be stretched and made warp-free by means of adjustment screws. This process allows the blade to be set exactly perpendicular to its axis of rotation; the maximum error measured near the cutting edge can be made less than 5 μm. The cutter carries either a fixed abrasive or a suspension of abrasive in oil.

The material to be cut is clamped to the end of a pivoting arm supported by two slide systems, one moving parallel to the axis of rotation and the other at right angles to it. At the other end of the arm a system of adjustable counterweights allows the cutting pressure to be adjusted. By using a low speed of rotation (cutting speed between 0.25 and 0.40 m/s) and a very constant cutting pressure (it can be kept constant within less than 2 grammes by the counterweight system), and by the use of a free abrasive the bowl cutter machine can cut thin specimens with faces parallel within very small tolerances. For example, a germanium slice



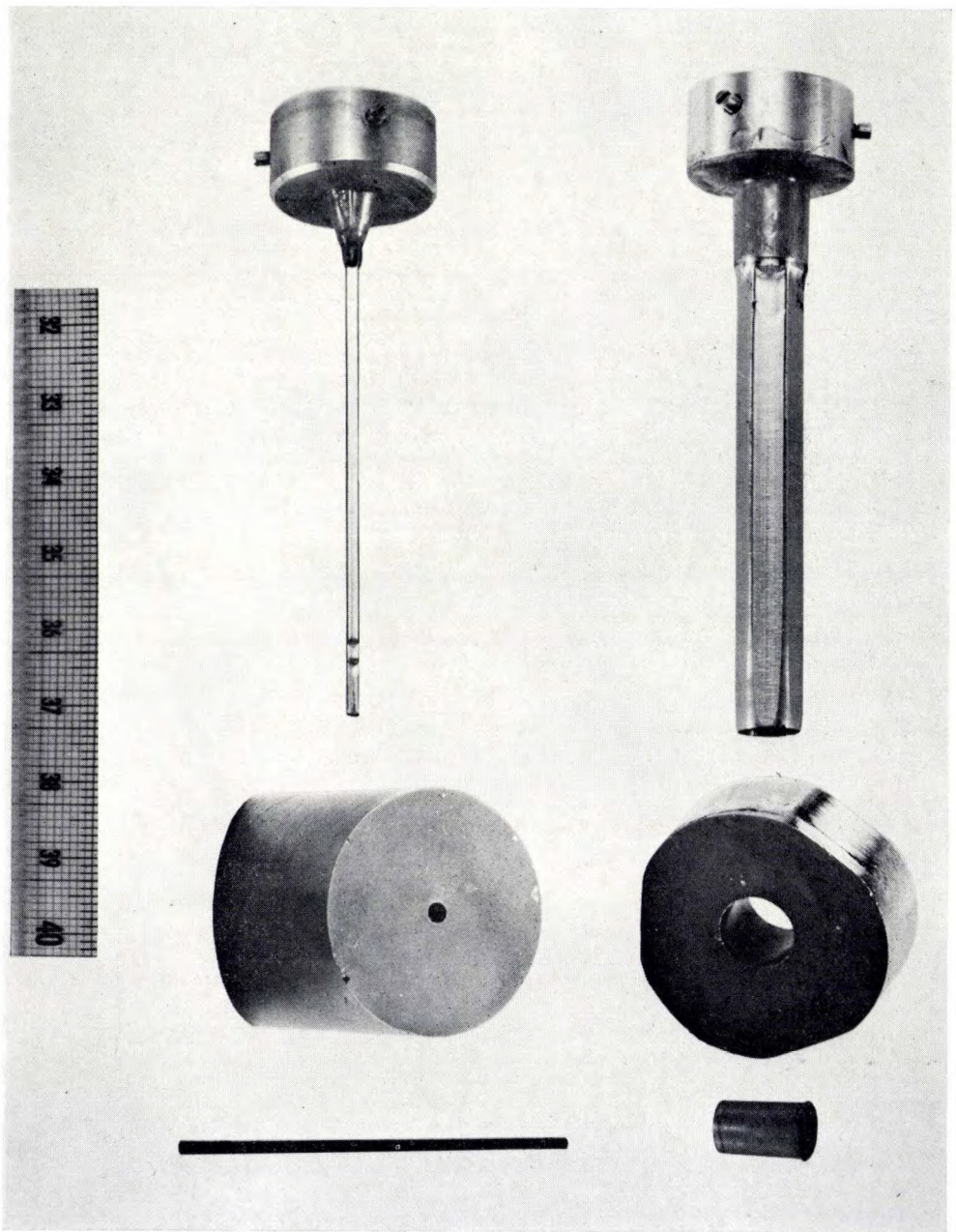
80 mm in diameter and 0.5 mm thick can be made with its faces parallel to within 20 μm.

A set of cutters with different internal diameters allows larger pieces of material to be cut, e.g. rectangular wafers of germanium 150 × 45 mm and 1 mm thick, with faces parallel to within 30 μm.

[1] French Patent No. 1500037, of 13th June 1966.

A method for small-diameter core drilling

Existing methods for abrasive machining of brittle materials have recently been supplemented by a new core drilling method, developed at LEP [1] by A. Charles-Georges and A. Salmon.



The chief problem in core drilling is that of ensuring a continuous flow of fresh abrasive slurry to the cutting face during machining. This becomes more difficult when the hole to be bored has a small diameter (about 2 mm in certain cases of interest to us) compared with its length (80 to 100 mm).

The difficulty is overcome by providing the drilling bit with two sets of channels, one set for the supply of fresh slurry, the other for the removal of the dirty slurry. A bit in the form of a *tube of polygonal cross-section* will automatically create such channels as it penetrates into the material. The rotating bit cuts into the material to form a circular "ditch"; the outer diameter of the ditch

[1] French Patent No. 1568837, of 22nd Dec. 1967.

corresponds to the circumscribed circle of the polygon and the inner diameter to its inscribed circle. As the ditch becomes deeper a cylindrical hole is bored through the material, while a cylindrical solid core remains at the centre of the hole. With this special configuration one set of channels is formed between the polygonal bit and the wall of the hole, and another set between the bit and the core.

The holes drilled in this way are accurately straight, and since the abrasive is free there is very little disturbance of the crystal lattice. This new method has been used for drilling holes 30 mm deep with a diameter of 0.9 mm, and also for drilling holes 150 mm deep with a diameter of 5 mm. Since the core can easily be retrieved from the hole, the method is also useful for producing very thin cylindrical rods.

Volume 30, 1969, No. 8/9/10 pages 213-308 Published 19th December 1969

PHILIPS TECHNICAL REVIEW

VOLUME 30, 1969, No. 11/12

"The larger grows our sphere of knowledge, the larger becomes its contact with the unknown". Although many will feel that this quotation, which was originally coined for science in general, seems a particularly apt description of the situation in the field of high-energy physics, the search for ever new facts and data in this field continues, as does the effort to devise ever more powerful machines for making this search. The proton synchrotron is the acme of the development of such machines, and Philips — while not exactly involved in building them — are gratified to have made an important technical contribution to their design. This contribution is the theme of an article published in this issue: it deals with the ferrite materials developed for five major proton synchrotrons. The unique interest of these machines prompted us to commission another article which takes a closer look at all five of them. This article was written for us by Dr. Gouiran of CERN, Geneva. An introduction to both articles, by Professor Casimir, gives some thought to the historical background of their subject and to the role industrial activities can play in this field.

Valuable cooperation has been given by members of the different proton-synchrotron laboratories in providing us with data and illustrations and in checking the parts of the articles about their machines. On behalf of the authors, we wish especially to acknowledge the help of Mr. M. Plotkin of the Brookhaven National Laboratory; Dr. C. Bovet, Dr. H. Fischer, Mr. W. Pirkl and Mr. P. H. Standley of CERN; Dr. J. H. Martin and Dr. T. de Parry of the Argonne National Laboratory; Mr. D. A. Gray, Mr. T. Hyman and Dr. D. J. Thomson of the Rutherford High Energy Laboratory at Chilton; and Mr. J. L. Kirchgessner of the Princeton-Pennsylvania accelerator at Princeton.

Electrostatic machines, particle accelerators and industry

H. B. G. Casimir

Electrical discharge phenomena at high voltages have been known to man since ancient times: thunder and lightning filled him with wonder and awe. Centuries went by before modest sparks could be produced with the Leyden jar and a piece of catskin. Soon after that, it was realized that these sparks were essentially manifestations of the same effect or, from the opposite point of view, that the atmosphere is a great electrostatic machine. Benjamin Franklin (1706-1790) supplied the proof of this by flying a kite. This was at once an important piece of physics and a striking example of a rational explanation of a phenomenon hitherto feared as a capricious demonstration of supernatural power.

But the physicists did not rest content with little sparks. The great frictional electrostatic machine of Van Marum (1750-1837), which can still be admired in Teyler's Museum in Haarlem, is an impressive apparatus. Many more electrostatic machines based on friction and electrostatic induction were built in the course of the nineteenth century.

After Franklin the electricity in the atmosphere was hardly used at all as an aid in performing systematic laboratory experiments, although an attempt was made in this direction as late as 1930 on the Monte Generoso with cables suspended along the mountainsides. On the other hand the Van de Graaff generator, a well engineered form of the electrostatic machine, continues to be an important tool in modern physics.

Discharges in rarefied gases were widely and systematically investigated towards the end of the nineteenth century. The discovery of the electron and X-rays — or as we might say today, of two elementary particles — was a direct result. The theoretical investigation of the electromagnetic field of a moving electron led to the conclusion that its mass is not independent of its velocity (Abraham, Lorentz), a conclusion which has also been confirmed by observation. The special theory of relativity gave a simple, general formulation of this relation between mass and velocity. The mass m of the

Prof. Dr. H. B. G. Casimir is a member of the Board of Management of N.V. Philips' Gloeilampenfabrieken and is responsible for all research activities in the Philips Group of Companies.

moving electron (charge e) is given by the relation:

$$mc^2 = eV + m_0c^2$$

and the velocity v by:

$$m^2v^2 = m^2c^2 - m_0^2c^2.$$

Here m_0 is the rest mass of the electron, c the velocity of light and V the voltage accelerating the electron. The magnitude of m_0c^2 is roughly 500 000 electron volts. It is thus only at voltages of half a million volts and more that the "relativistic effects" are very considerable. Such voltages were not yet available when the theory of relativity was established.

Much greater particle energies, of several million electron volts, had already been found in nature, in particles in the radiation emitted by radioactive nuclei. These high-energy particles were the first tools of nuclear physics. In 1919, Rutherford and his co-workers found a first indication that alpha-particles emitted by "radium C" and having a range of 7 cm in air (corresponding to an energy of 7.7 MeV) can cause a nitrogen nucleus to disintegrate. If it were possible to generate beams of particles with comparable energy but of greater intensity, the way would be open for an entirely new attack on the nucleus of the atom. Georg Gamow, a physicist of Russian descent, who died in America on 20 August 1968 — more widely known for his witty writings on modern physics and its history — pointed out that energies of less than a million electron volts could also be adequate to produce nuclear reactions. Particles with such energies — too low for them to be able to get over the potential barrier around the nucleus — would be able to tunnel *through* this barrier and, once they reached the interior of the atom, would cause disintegration. ("Ein Teilchen in Kern eindringt und ausruft Zertrümmerung", he would say in his not entirely faultless German.) This was confirmed in 1932 by Cockcroft and Walton, who bombarded lithium with protons and found that the protons could react with a lithium nucleus even at an energy of 125 keV, the end-product being two alpha-particles of very high energy. The protons used for this purpose were artificially accelerated by means of a new type of high-voltage generator in which an a.c. voltage was converted into a very much higher d.c. voltage by a cunningly devised circuit of capacitors and rectifiers. Though Cockcroft and Walton did not achieve more than 800 kV with this arrangement, it marked the entry of nuclear physics into the technical stage.

Soon after Cockcroft and Walton's first experiments a number of generators based on the same principle were built in Eindhoven at Philips Research Laboratories in a more industrial version. The 1.25 MV generator which was installed in the Cavendish laboratory in Cambridge in 1937 [1] can boast a remarkable record of service. In 1965 I saw it in Johannesburg where it was still rendering valuable services. The original mercury-vapour rectifiers had meanwhile been replaced by solid-state rectifiers and the capacitors were gradually beginning to fail. The capacitors had not, however, been designed to operate at an altitude of 1900 metres so that after 30 years of operation our judgement should not be too harsh.

Interesting nuclear physical reactions can be produced by particles of widely differing energies. On the one hand for example, the reaction of the hydrogen nucleus with an atomic weight of 3 (tritium) with the hydrogen nucleus with an atomic weight of 2 (deuterium) occurs even at energies lower than 100 keV. On the other hand, when the energy of the bombarding particles is increased, new reactions, new excited states, new phenomena are found each time. Although voltages of 5 MV and higher were later successfully achieved by using Van de Graaff machines, it was at first difficult to exceed 1 MV either with Van de Graaff or with Cockcroft-Walton generators. Then physicists struck upon the idea of accelerating particles not by passing them once through a very high voltage, but by subjecting them to repeated kicks obtained with a.c. voltages. This is done both in the linear accelerator and in the cyclotron. Both principles have been described in this journal on previous occasions [2]. I do not intend here to dwell on the amazing, and sometimes bewildering, expansion which nuclear physics has subsequently undergone.

Side by side with nuclear physics and partially utilizing similar methods of observation (ionization chambers, Wilson chambers and Geiger counters) there went another branch of research — that of cosmic rays. Fast electrons and other forms of radiation reach the earth from space. In these rays, there are particles with energies very much higher than were known from radioactive processes or than could be produced with accelerators at that time. It was to some extent the same situation as with the thunderstorm in the days of Franklin, but now the source of power was not the atmosphere but space. A surprising discovery made during this investigation was that of the positron (or positive electron) by Anderson in 1932, although, in a certain sense, this had already been predicted by Dirac's theory. It was found that in the vicinity of a nucleus a gamma quantum of at least twice half a million electron volts can be transformed into an electron-positron pair. It is

[1] A. Kuntke, A generator for very high direct current voltage, Philips tech. Rev. **2**, 161-164, 1937.

[2] Cyclotrons are discussed in Philips tech. Rev. **12**, 65, 241, 349, 1950/51 and **14**, 263, 1952/53, the linear accelerator in Philips tech. Rev. **14**, 1, 1952/53, and **15**, 1, 1953/54.

difficult for us to realize today that when Anderson made his discovery technology was not yet far enough advanced to generate X-rays of the required energy. However, pair formation was not the end. Great "showers" of positrons and electrons can occur in cosmic rays, requiring many millions of electron volts. Shortly afterwards, artificial radioactivity was discovered and it was then found that an artificial radioactive nucleus which has an excess of protons emits positrons. Just as important was the discovery of the "heavy electrons" or mesons, which was largely due to the work of Powell, Occhialini and their co-workers in the forties; their technique was to observe particle traces in photographic emulsion. These particles had also been predicted: Yukawa had postulated them in 1935 to account for the short range of the specific nuclear forces which keep the protons and neutrons together in the nucleus. The fact that the inspired but certainly qualitative theory of Yukawa was so strikingly confirmed by experiment still remains for me one of the most remarkable events in modern physics. Incidentally, work in the field of cosmic rays was particularly attractive to physicists with an inclination for globe trotting and mountaineering, for the properties of these rays depend to a significant extent both on the degree of latitude on earth and on altitude.

The intensity of the cosmic rays is, however, slight. If one wished to make a proper investigation of the new particles and the completely new phenomena which emerged on their appearance, it would be necessary to generate them artificially. Similarly, nuclear physics only advanced successfully when the investigator was no longer solely dependent on radioactive elements. In order to generate the simplest of the "heavy electrons" an energy of no less than 500 MeV is required. And just as in nuclear physics new excited states and new nuclear reactions were constantly being found when the energy increased, so in this domain of typical high energy physics each gain in energy leads to new particles or new reactions. Every advance in accelerator technology has so far been of interest for fundamental physics. Thus the negative protons were promptly discovered as soon as an energy of more than $2m_pc^2$ was available (here m_p is the mass of the proton; m_pc^2 is roughly 1000 MeV).

In an after-dinner speech at the conclusion of a CERN conference I once pointed out that the physics of the higher energies could be regarded as the highest form of science since one not only evolves one's own experiments, but even creates one's own particles. What then must be the attitude of an industrial laboratory towards this field which, in its fundamental aspects, is so fascinating? As yet, no technical application is in sight. No new energy sources are emerging, nor are new means of communication or very special effects on living matter, although therapy by means of "pions" has been considered. This does not mean that applications will never be found, nor do I wish to belittle the intrinsic philosophical value of this research work. The industrial laboratory, however, cannot as yet afford to enter this domain independently.

Unfortunately, this is also true of the construction of accelerators. We could build Cockcroft-Walton generators, synchrocyclotrons and isochronous cyclotrons with our own resources, and have done so on our own account. With the big accelerators this is no longer possible. The project is too complicated, the market too small, the investments are too great. Thus the role of industry is restricted to supplying components for large projects which are outside our sphere of influence both in design and in purpose. For a laboratory such as ours, which, from time to time, has been accustomed to making a modest but independent contribution to the development of instrumentation on the frontiers of physics, it is to some extent disappointing to have to recognize that it is no longer in a position to do this. It is all the more gratifying that materials developed by us for completely different purposes play a useful and even almost essential part in such projects, as the articles contained in this issue will demonstrate.

Summary. As introduction to the following articles about five major proton synchrotrons a historical sketch is given of the efforts to apply ever higher voltages and ever higher energies to the investigation of elementary particles. A few steps in the related growth of our knowledge of these particles are indicated.

Ferroxcube for proton synchrotrons

F. G. Brockman, H. van der Heide and M. W. Louwerse

The core of the accelerating station of the Brookhaven Cosmotron — the first proton synchrotron to be completed (1952), accelerating protons to 3 GeV — consisted largely of ferrite material provided by Philips (ferroxcube) [1]. Since that time Philips have provided similar materials for the construction of five other proton synchrotrons: the 28 GeV machine for CERN at Geneva (Switzerland), which was put into operation in 1959, the 33 GeV machine at Brookhaven, Upton, N.Y., U.S.A. (1960), the 12.5 GeV machine at Argonne near Chicago, Ill., U.S.A. (1963), the 7 GeV machine at Chilton, England (1963), and the 3 GeV machine at Princeton, N.J., U.S.A. (1963).

This article is intended to explain how ferroxcube is used in these machines and to discuss the characteristics of the required material. In the treatment that we shall give here the principles of a proton synchrotron will be briefly recapitulated, the properties of ferroxcube that are of interest will be reviewed, and it will be shown that for each of the five machines it was possible to manufacture a grade of ferroxcube that was matched to the specific requirements.

As we shall often refer to these machines individually, it will be convenient to make use of the abbreviations that are commonly used. Abbreviations and full titles are as follows:

1) CPS: the CERN Proton Synchrotron at Geneva.
2) AGS: the Alternating Gradient Synchrotron at Brookhaven.
3) ZGS: the Zero Gradient Synchrotron at Argonne.
4) Nimrod, the 7 GeV Proton Synchrotron at Chilton.
5) PPA: the Princeton-Pennsylvania Accelerator at Princeton.

The ferroxcubes treated in this article were developed when these accelerators were designed and built. It should perhaps be noted here that since then new developments in ferrite technology have taken place at Philips that are of great importance for the modernization of the synchrotrons. In fact, completely new accelerating stations incorporating the newly developed ferrites have been designed for the CPS and the AGS. Up to the present there have been 15 accelerating stations in the CPS and 12 in the AGS, but now the higher quality of the new materials makes the new stations so much more effective that fewer of them are required, leaving more valuable space for experiments.

Dr. F. G. Brockman is with Philips Laboratories, Briarcliff Manor, N.Y., U.S.A.; H. van der Heide is with Philips Research Laboratories, Eindhoven; M. W. Louwerse is with the Philips Electronics Components and Materials Division (Elcoma), Eindhoven.

The principles of the proton synchrotron

In a proton synchrotron (see *figs. 1* and *2*) the protons are injected into an annular *vacuum chamber*, and are kept in a given closed orbit in this chamber by a guiding *magnetic field* which must increase during the acceleration. At one or more positions around the ring the particles are accelerated by an *electric field*. When the particles have reached their maximum energy, they are ejected from the ring or made to strike an internal target, the magnetic field is brought down to a low value again, and a new cycle begins.

The trajectory of a particle of charge e moving with momentum p perpendicularly to a magnetic field of flux density B will be curved, and the radius of curvature ϱ is given by the relation [2]:

$$p = \varrho eB. \quad \ldots\ldots \quad (1)$$

In a proton synchrotron ϱ is kept constant by increasing B in proportion to p.

As will be discussed more fully in another article in

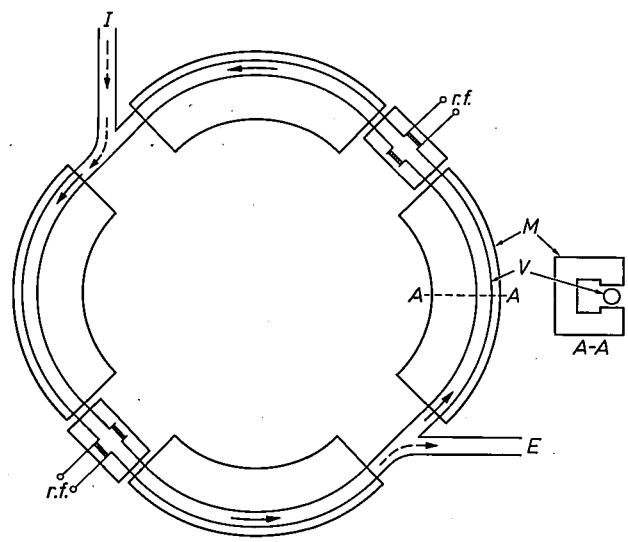


Fig. 1. A proton synchrotron, schematic. The protons are kept in a closed orbit in the vacuum chamber *V* by guiding magnets *M*. Between magnets the orbit is straight. Straight sections provide room for accelerating the protons (*r.f.*), for injecting them at the beginning of a cycle (*I*) and for ejecting them at the end (*E*).

this issue [3], eq. (1) relates the size and the final energy of the accelerator. In fact, since the maximum value of B is limited by saturation of the magnet steel, and the final energy is (in the relativistic limit) proportional to p, eq. (1) indicates that the final energy is proportional to the radius. Thus, for the AGS the final energy (33 GeV) and the radius (128.5 m) are both about ten times as large as the corresponding values for the PPA (3 GeV, 12.5 m).

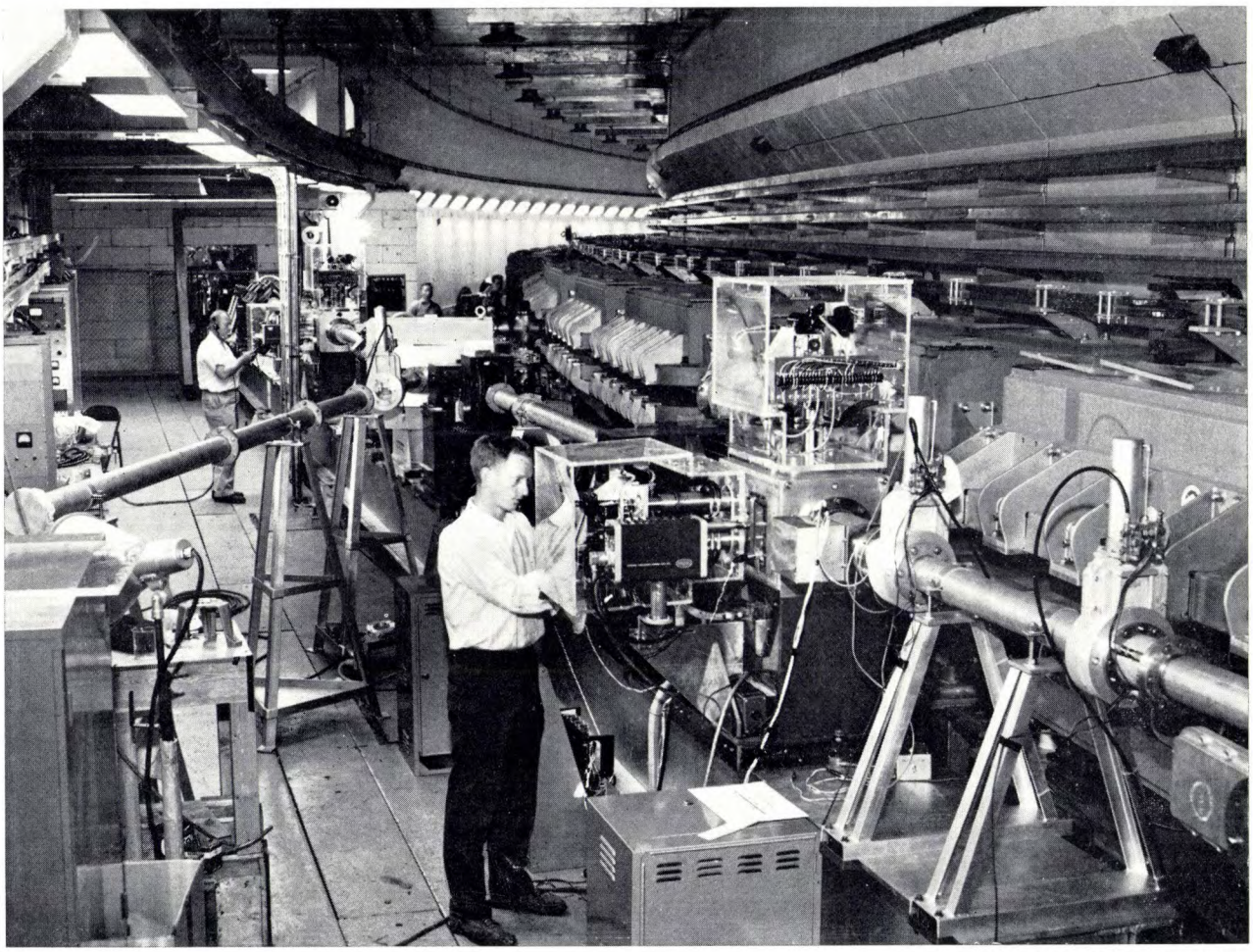


Photo Brookhaven National Laboratory

Fig. 2. An interior view of the 33 GeV Alternating Gradient Synchrotron (AGS) at Brookhaven, at the junction of the injection line and the main synchrotron ring.

The particles are accelerated by giving them repeated kicks by means of *oscillating electric fields* that appear across insulating sections in the metal vacuum-chamber wall. An essential characteristic of a synchrotron is the *synchronization* of these oscillating fields to the particle motion: this ensures that the particle is always exposed to a field of favorable phase when it passes the gap. Consequently, the particles are accelerated in bunches, the positions in between corresponding to unfavorable phases of the field. As the particle is accelerated, its velocity and therefore its orbital frequency increases steadily, and the frequency of the r.f. voltage across the gap must be made to follow the variation exactly.

The orbital frequency is the velocity of the particle divided by the circumference. In the Cosmotron, for instance, the orbital frequency came out to be 0.36 MHz at injection increasing to 4.2 MHz at ejection. In the higher-energy accelerators the radius is larger and therefore, at least at ejection when the velocity is approximately the velocity of light c, the orbital frequencies are lower (e.g. 0.375 MHz at ejection in the AGS). In the five machines under discussion the frequency of the

[1] F. G. Brockman and M. W. Louwerse, The ferroxcube transformer core used in the Brookhaven Cosmotron, Philips tech. Rev. **15**, 73-83, 1953/54.

[2] The formulas in this article are adapted to the use of SI units.

[3] R. Gouiran, Five major proton synchrotrons, p. 330-365 of this issue.

r.f. voltage is (for reasons to be discussed in [3]) not equal to the orbital frequency but an integral multiple of it. In the AGS this multiple is 12, and the final r.f. frequency is 4.45 MHz; at injection the velocity is such that the r.f. frequency comes out to be 1.4 MHz. We note that the ratio of the two frequencies in the AGS, 4.45/1.4 = 3.2, is smaller than with the Cosmotron: 4.2/0.36 = 11.7; this is due to the injection energy at the AGS being higher, the injection velocity thus being closer to c.

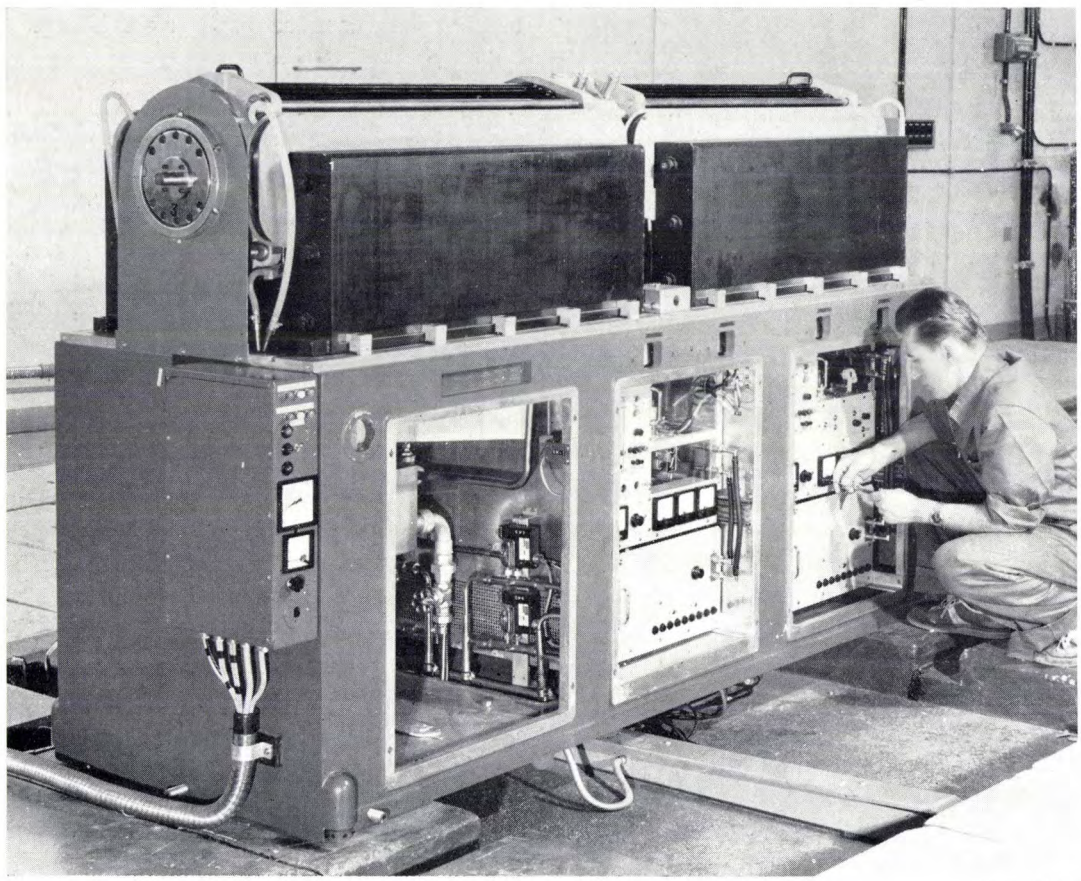


Photo CERN

Fig. 3. An accelerating station of the CERN Proton Synchrotron (CPS) at Geneva, before being mounted in the synchrotron ring.

It is in the accelerating stations in which the accelerating fields are produced that the ferroxcube materials are used in large quantities. (Lesser quantities are used in other components, such as coupling transformers, etc.) An accelerating station for the CPS is shown in *fig. 3*, and one for the AGS in *fig. 4*.

The accelerating unit and the function of the ferroxcube

An accelerating station is shown schematically in *fig. 5*. It is a shielded structure, cylindrical around the proton beam, allowing high r.f. electric fields to be produced across the accelerating gap without introducing any other effects on the beam, or elsewhere. It can be considered as a single-turn inductor. An r.f. current fed into the inductor (i.e. into the wall of the structure) produces an r.f. magnetic flux in the inductor around the axis, which in turn induces an r.f. voltage across the ends of the inductor, i.e. across the gap. The amplitude $\hat{V}_\sim$ of the gap voltage $V_\sim$ is related to the amplitude $\hat{B}_\sim$ of the r.f. flux density $B_\sim$ by:

$$\hat{V}_\sim = \omega A \hat{B}_\sim, \quad \ldots\ldots \quad (2)$$

where ω is the r.f. angular frequency and A is the area of the inductor that is traversed by the flux.

To give the particles the required acceleration, $\hat{V}_\sim$ must have a value of thousands of volts. On the other hand, the current feeding the inductor is limited, since it is derived from vacuum tubes in the r.f. power supply which have a limited power dissipation. In order to minimize the current required for the given gap voltage, the *impedance* of the inductor should be *as large as possible*.

The impedance can be made large in two steps. In the first place the inductor can be loaded by a ferromagnetic

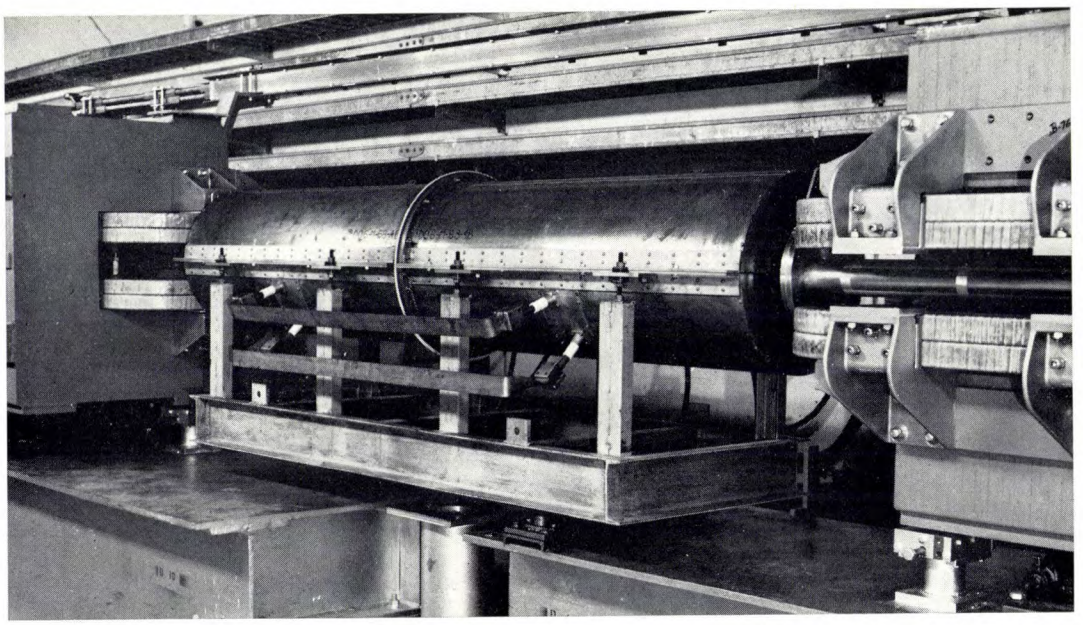


Photo Brookhaven National Laboratory

Fig. 4. An accelerating station of the AGS in Brookhaven.

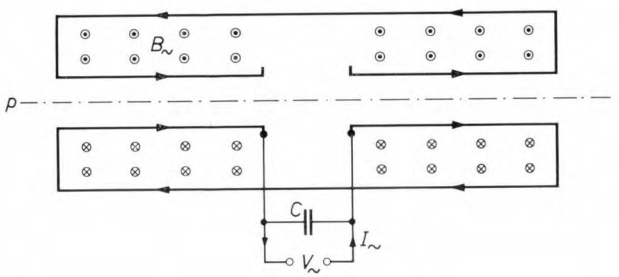


Fig. 5. An accelerating cavity (schematic). The protons move along the axis of the cylindrical structure in a beam p. The cavity can be considered as a single-turn inductor. An r.f. current I fed into the cavity wall produces an r.f. magnetic field in the cavity around the axis, and the corresponding r.f. flux (of density $B_\sim$) induces an r.f. voltage $V_\sim$ across the gap. An external capacitor C may be used to obtain an LC circuit resonant to the operating frequency.

core. If the relative permeability of the core is μ_r, then, for given linear dimensions of the inductor, its impedance is increased by a factor μ_r. Or, putting it the other way round, a given required impedance can be obtained with smaller linear dimensions of the inductor if it is loaded with a ferromagnetic core. The core should not introduce undue losses and, at the frequencies involved, the only ferromagnetic materials that can function effectively are the ferrites. Secondly, by making the inductor part of a resonant LC circuit tuned to the operating frequency, the impedance is still further increased. This is accomplished by connecting a suitable capacitance across the terminals (cf. fig. 5). The impedance Z in the two cases is:

$$\text{non-resonant case: } Z \approx j\omega L, \quad \ldots \quad (3a)$$

$$\text{resonant case: } Z \approx \omega Q_c L. \quad \ldots \quad (3b)$$

L is the inductance and Q_c the quality or Q factor characterizing the losses of the circuit. The equations (3a) and (3b) are only valid if Q_c is sufficiently larger than 1. The magnitude of the impedance Z is Q_c times as large in the resonant case as it is in the non-resonant case.

Eq. (3a) and (3b) are easily derived from the equivalent circuits for both cases (*fig. 6a* and *b*). R is the equivalent series loss resistance of the inductor L. The Q factor is $Q_c = \omega L/R$ and the angular frequency at resonance ω_0 is given by $\omega_0^2 = 1/LC$. Case (*a*) is a special case of case (*b*) with $C = 0$ ($\omega_0 = \infty$). The impedance Z is given by:

$$1/Z = j\omega C + 1/(R + j\omega L). \quad \ldots \quad (4)$$

Introducing ω_0 and Q_c instead of C and R, then approximately, for $Q_c \gg 1$:

$$1/Z \approx (1/j\omega L)\,(1 - \omega^2/\omega_0^2 - 1/jQ_c). \quad \ldots \quad (5)$$

For $\omega_0 \to \infty$ and neglecting $1/jQ_c$ compared to 1, we obtain eq. (3a); for $\omega = \omega_0$ eq. (3b) is obtained.

It may be noted that the power dissipated *in* the circuit, which for a given voltage is proportional to the *real* part of $1/Z$, is *in both cases* given by $1/\omega Q_c L$ (see eq. 5). Thus, for a given inductor, this power is the same whether the inductor is resonated or not.

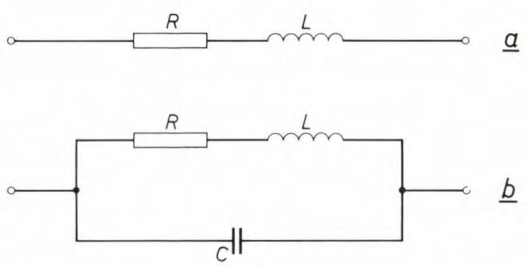


Fig. 6. The equivalent circuits of a non-resonant cavity (*a*) and of a cavity resonated with an external capacitor C (*b*).

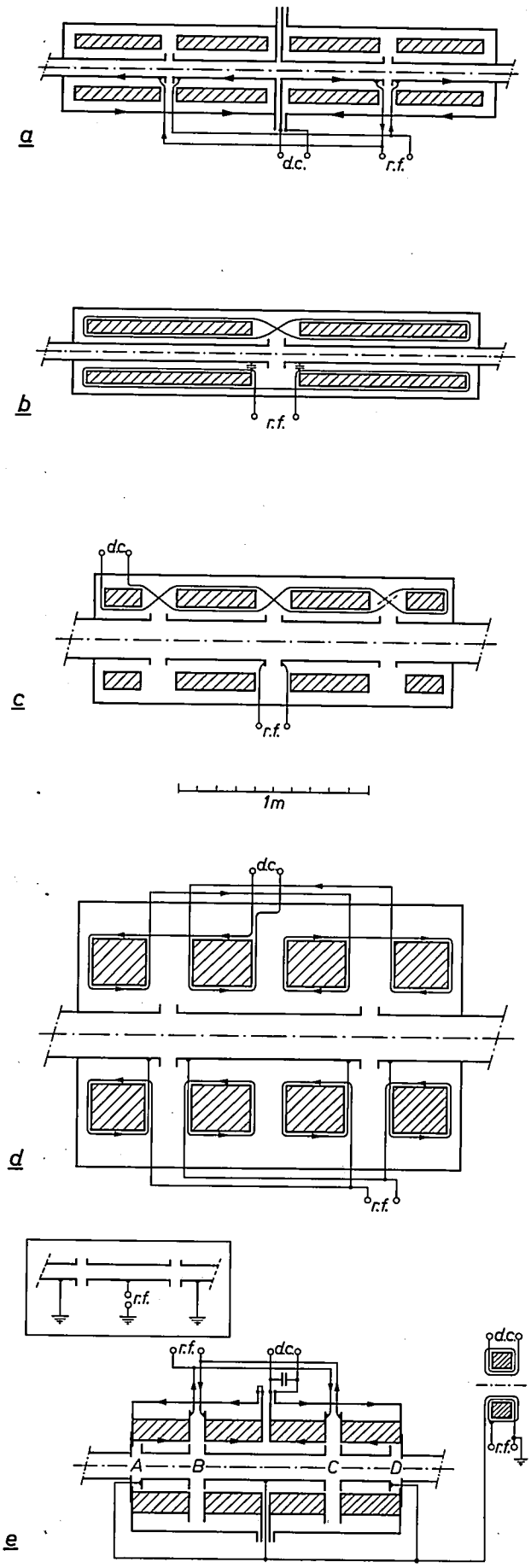
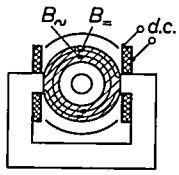


Fig. 7. An accelerating cavity of each synchrotron, in longitudinal section. *a*) AGS (Brookhaven), *b*) CPS (Geneva), *c*) ZGS (Argonne), *d*) Nimrod (Chilton), *e*) PPA (Princeton). The ferrite cores (hatched) are drawn to scale; cavity walls (heavy lines) and wiring (thin lines) are schematic. In comparing sizes it should be noted that the cavities of the AGS (*a*), the CPS (*b*) and the PPA (*e*) are circular in cross-section, whereas those of the ZGS (*c*) and of Nimrod (*d*) are rectangular, with the long side perpendicular to the plane of the drawing. (Cross-sections of the cores are compared in fig. 20.) In the first four cavities the pattern of fig. 5 is easily recognized. If more than one gap is required, foreshortened $\frac{1}{2}\lambda$ sections can be introduced between the foreshortened $\frac{1}{4}\lambda$ end-sections. The r.f. source is connected to "*r.f.*", the polarizing current source to "*d.c.*". For (*b*) flux lines of the r.f. field and of the polarizing field are shown in the cross-section ($B_\sim$ and $B_=$ respectively).



In the PPA (*e*) the acceleration is in two steps. In the first part of the cycle the particles are accelerated with the aid of a drift tube. In the inset of (*e*) the drift-tube principle is indicated: the section between the two gaps changes the sign of its voltage in the time that the particle travels through this section, so that the particle is accelerated at both ends. In the PPA accelerating station the inner part acts as the drift tube. The r.f. voltage is applied to the three sections in parallel and the accelerating field appears at the gaps *A* and *D* only. During the second part of the cycle the protons are accelerated by the cavity surrounding the drift tube, and the accelerating field appears at the gaps *B* and *C*. The drift tube is resonated with an external inductor, and the resonant system is again tuned to the changing frequency by polarizing a ferrite in the inductor. It may be noted that in a multiple-gap cavity the voltages across the gaps are in phase; a particle sees virtually the same phase when passing these gaps as the time of passage between the gaps is small compared to an r.f. period. In a drift tube, on the contrary, the change of phase during passage is essential.

Resonant circuits, often in the form of cavity resonators, have been widely applied to solve similar problems, in electron synchrotrons for instance and also in many other high-frequency devices, klystrons, magnetrons, linear accelerators, etc. In these applications the frequency is constant. In a proton synchrotron, however, the frequency is changing rapidly, and therefore a choice has to be made between: a) a non-resonant system, in which the advantages of resonance are lost and an inductance is incorporated that is sufficiently large; and b) a resonant system that, at any moment, is tuned to the rapidly varying orbital frequency of the particles. In the first case the demands on the r.f. source will be relatively high, in the second case special provisions must be made for tuning the system. The solution (a) was used in the Cosmotron [4]. In the five accelerators discussed in this issue, on the other hand, a resonant system has been chosen. In

[4] Whereas the accelerating station of the Cosmotron was essentially untuned, resonances due to distributed capacitances did occur and had to be dealt with. See J. P. Blewett, M. Plotkin and E. J. Rogers, Rev. sci. Instr. **24**, 795-800, 1953, in particular pp. 798 and 799.

[5] In practice, the gap-voltage amplitude is often varied over wide ranges, to improve particle capture, acceleration efficiency, etc.

these five accelerators the ferrite is also used for *tuning the system.* The resonant frequency of the system, $\omega_0 = 1/\sqrt{LC}$, is varied by varying the inductance L. This is achieved by applying a variable d.c. magnetic bias field to the ferrite. This field polarizes the ferrite, thereby reducing its permeability for oscillating fields of relatively small amplitude; as saturation is approached the relative permeability approaches unity.

The resonant frequency could, in principle, also be varied by mechanically varying the capacitance C in fig. 5, but the former method is more amenable to feed-back control, since it is non-mechanical.

The accelerating unit of fig. 5 has a close resemblance to a resonant cavity consisting of two $\frac{1}{4}\lambda$ coaxial resonators connected "face-to-face" and resonating in push-pull. A cavity of this type can also be tuned by means of a ferrite. The resemblance, however, is not complete since in the five synchrotrons the cavity sections are much shorter than $\frac{1}{4}\lambda$, to ensure a relatively uniform flux distribution in the axial direction, and the cavity is not resonant by itself, but is resonated by an external capacitor. The two halves of the accelerating unit may be referred to as *foreshortened* $\frac{1}{4}\lambda$ coaxial resonators.

The function of the ferrite, then, may be summarized as follows:

a) It enables a sufficiently high impedance to be obtained with an accelerating unit that is small enough for practical use.

b) It provides a convenient method for tuning the acceleration station to the varying operating frequency.

To fulfil these functions properly, the ferrite must meet the following requirements:

1) Its permeability must be sufficiently high.

2) The magnetic losses under r.f. excitation must be sufficiently low, i.e. its Q factor must be sufficiently high.

3) A sufficient reduction of its permeability for r.f. fields must be obtainable with a bias field of practical magnitude.

It can be seen from eq. (3b) that the first two requirements can be combined in the one that the instantaneous value of $\mu_r Q\omega$ should be sufficiently high under all operating conditions.

From the expression for the resonant frequency of the system, $\omega_0 = 1/\sqrt{LC}$, it follows that the bias has to be capable of reducing the permeability by a factor of

$$\frac{\mu_{max}}{\mu_{min}} = \left(\frac{f_f}{f_i}\right)^2, \quad (6)$$

where f_i and f_f are the initial and the final orbital frequency respectively (any corrections for filling factors, etc. are neglected here).

Assuming the gap voltage amplitude $\hat{V}_\sim$ to be kept constant during the acceleration [5], it follows from eq. (2) that the amplitude of the r.f. flux density varies inversely with the frequency:

$$f_i(\hat{B}_\sim)_i = f_f(\hat{B}_\sim)_f. \quad \ldots . \quad (7)$$

The cavities used in practice

Before discussing the properties of ferrites we shall turn briefly to the accelerating cavities actually used in the five synchrotrons. These are shown in longitudinal section in *fig. 7*, all drawn to the same scale. The basic pattern of fig. 5 is easily recognized in the first four cases. If more than one gap is required, foreshortened $\frac{1}{2}\lambda$ sections can be introduced between the foreshortened $\frac{1}{4}\lambda$ end-sections. The largest type of cavity used is that of Nimrod at Chilton; this is shown in *fig. 8*. In the PPA at Princeton the acceleration is obtained in two steps: a "drift tube" (explained in the caption of fig. 7) is used in the first part of the cycle, and a cavity in the second part; the reason is that the required frequency swing is too large to be obtained in one step by ferrite tuning. *Fig. 9* shows a PPA accelerating station under construction.

Fig. 7 shows how the bias field is applied to the ferrite. In the AGS, the ZGS, Nimrod and the PPA this is done by means of one or more turns (which may be formed by the cavity wall itself) around the ferrite. In these cases only a limited number of turns are used to prevent interference with the r.f. system; this leads to large polarizing currents (of the order of 1000 A). In the CPS the ferrite is polarized by an external iron-cored yoke magnet; the number of turns on this magnet is relatively large and only a few amperes are necessary for polarization (see *Table I*). An important consideration is that the d.c. and the r.f. circuits should be decoupled; the decoupling is obtained by arranging the winding or the geometry in such a way that at any instant the r.f. field and the polarizing field

Table I. Maximum ferrite polarizing field and current (per cavity).

	max. current in A	number of turns	$H_{p\,max}$ in A/mm	$H_{p\,max}$ (in Oe)	μ_{max}	μ_{min}
AGS	850	1	1.0	(12)	400	40
CPS	1	2200×2 [*]	0.8	(10)	250	25
ZGS	650	20	4.2	(53)	176	9.5
Nimrod	800	10	1.2	(15)	557	12
PPA cavity	14 000	1	10.8	(135)	150	3.5
PPA drift-tube inductor	6	1000	4.8	(60)	150	8

[*] In each cavity there are 2 parallel sections of 2200 turns with 1 A maximum current; each section is applied to one of the two ferrite cores in the cavity.

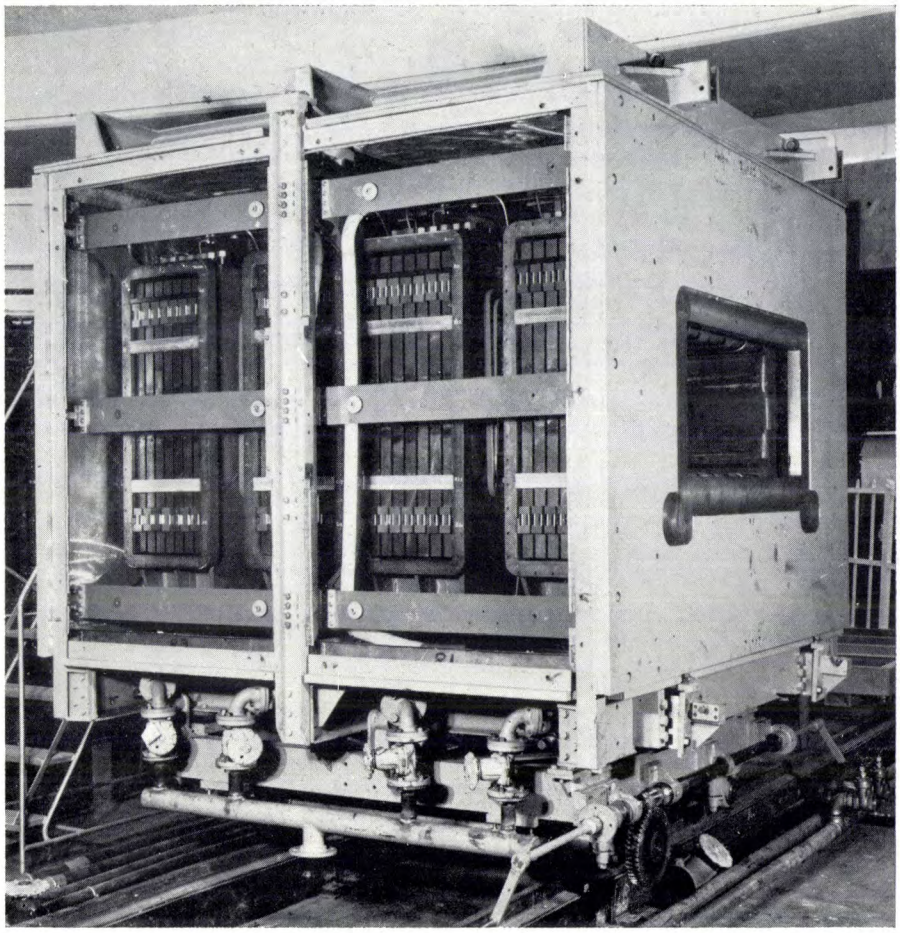


Fig. 8. The accelerating cavity of Nimrod, shown under construction.

Photo Rutherford High Energy Laboratory

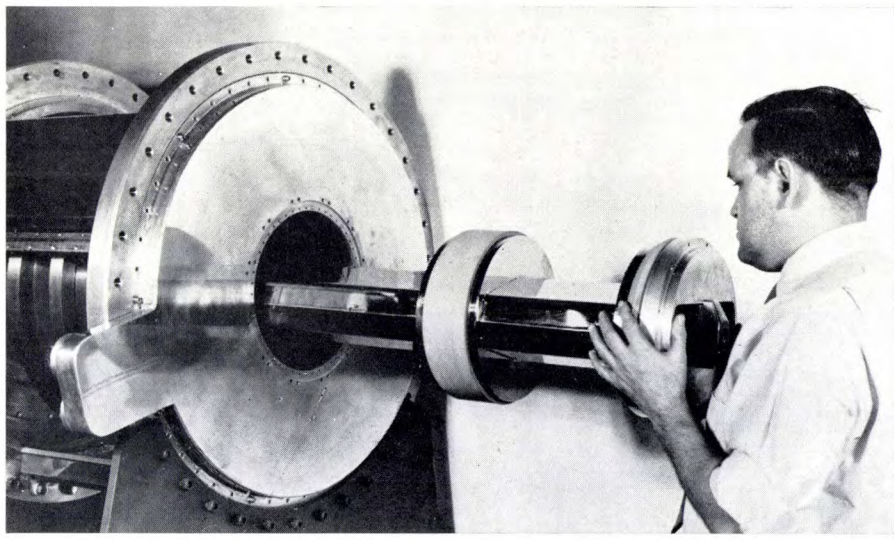


Fig. 9. A PPA accelerating station under construction. The drift-tube section is inserted in the cavity. The stub pointing to the left will carry the drift-tube inductor.

Photo Princeton-Penn Accelerator

are of equal sense in one region of the ferrite and of opposite sense in a symmetrical region. In *fig. 10* the polarizing winding of the ZGS cavity can be seen.

Ferrite varieties and their properties

The outstanding feature of ferrites, as magnetic materials, is their high resistivity, which is at least 10^6 (and in some ferrites as much as 10^{16}) times the resistivity of iron and the magnetic alloys. This makes the ferrites particularly suited for applications at high frequencies, the high resistivity usually precluding eddy currents. The usual methods of dealing with the eddy-current losses in metallic magnetic materials (lamination or powdering) are insufficient to render these materials useful in this application: the permissible thickness of the lamination becomes impracticably small; the

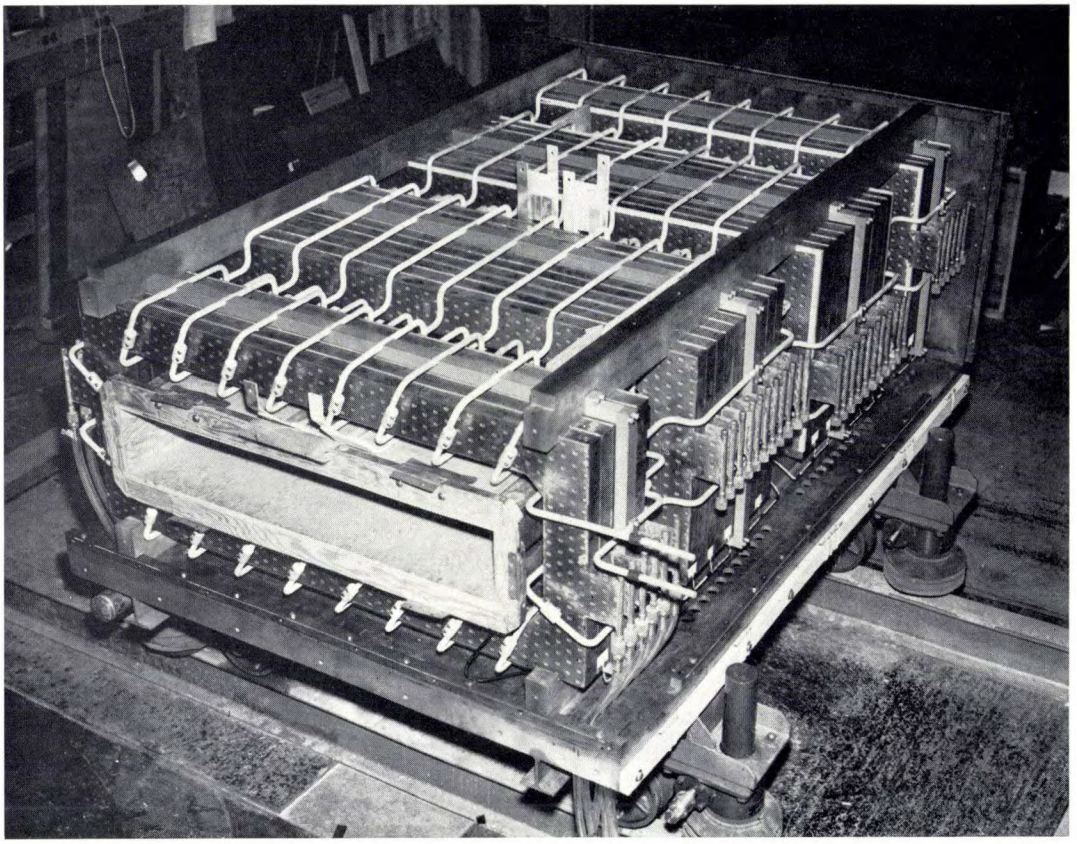


Photo Argonne National Laboratory

Fig. 10. The ferrite frames for the ZGS (Argonne) mounted around the (dummy) vacuum chamber. The twenty polarizing windings around each group of ferrite frames are clearly shown. Between the frames flat copper channels are mounted, guiding the water to cool the ferrite.

achievable permeability is too small in the case of powder cores [6].

The ferrite industry was established about 20 years ago following the now classical investigations of J. L. Snoek [7]. He showed that useful materials can be obtained by preparing mixed crystals of known magnetic ferrites and a non-magnetic ferrite such as zinc ferrite. Over the years two systems have emerged as the most useful. These are the manganese-zinc-ferrites (ferroxcube 3) and the nickel-zinc-ferrites (ferroxcube 4). It will be seen below that varieties of ferroxcube 3, although exhibiting some very good magnetic properties, cannot be used for the present application. Varieties of ferroxcube 4 have been chosen for the five accelerators.

Preparing a ferrite for a highly specialized application like the one under discussion involves a separate development of the material. The design of an accelerating station — only one of the many interrelated elements in the highly complex problem of designing a proton synchrotron — is based on the properties of ferrites, and the designers are initially guided by the specifications of marketed materials. However, the manufacturers of ferrites have carried out continuous research and development on their products to improve them. Whereas the specifications that are finally drawn up by the synchrotron builders are usually the outcome of an interplay of ideas on what is needed and on what is possible, it is not unusual for a specification to be included that constitutes a real challenge to the manufacturer. In this article it will be shown how Philips, the pioneer ferrite manufacturer, who are well equipped with means for research, development and manufacture of these materials, have succeeded in meeting such challenges.

In what follows we shall distinguish between two guiding principles when trying to meet a certain set of specifications. In the first place, the properties of the ferroxcube varieties depend essentially on their *nominal chemical composition*, and investigation is required to choose the variety best suited to the purpose. Secondly, even the best variety will have some bad characteristics if it is not prepared very carefully and it is a matter of *technology* to prepare a product that is sufficiently homogeneous, stoichiometric, mechanically strong, etc.

[6] See reference [4], in particular p. 795.

[7] J. L. Snoek, Non-metallic magnetic material for high frequencies, Philips tech. Rev. **8**, 353-360, 1946. See also J. L. Snoek, New developments in ferromagnetic materials, Elsevier, New York 1947.

Dependence of properties on chemical composition

Electrical properties

Whereas the manganese-zinc-ferrites (ferroxcube 3 varieties) have some very desirable magnetic properties: high relative permeabilities (1000 and over) and low hysteresis losses, they had to be rejected for the present purpose, mainly because of two undesirable electrical properties: a resistivity ϱ that is relatively low (around 1 Ωm) and a relative dielectric constant ε_r that is very high (several tens of thousands). The first gives rise to eddy-current losses, the latter, coupled with a high μ_r, to a resonance phenomenon called "dimensional resonance" [8] that also leads to unacceptable losses.

To ensure that such undesirable effects are avoided, specifications may be drawn up for ε_r and ϱ. Closely related to a relatively low resistance is the existence of relatively high dielectric losses, and to avoid these a minimum dielectric Q factor Q_ε may be specified. Typical electrical specifications are shown in *Table II*.

Table II. Ferrite specifications concerning ϱ and ε_r.

AGS	$\varepsilon_r < 20$ (at 1.5 MHz)
CPS	$\varepsilon_r < 12$ (at 2 MHz)
ZGS	$\varepsilon_r < 20$ (at 14 MHz)
Nimrod	$\varepsilon_r < 20$, $Q_\varepsilon > 5$, $\varrho > 0.2\times10^4$ Ωm (at 1.2-8.1 MHz)
PPA	$\varepsilon_r < 12$, $Q_\varepsilon > 25$ (at 30 MHz and 1 kV/cm)

If prepared with sufficient care, these specifications are easily met by ferroxcube 4, which has a high resistivity and a low dielectric constant. Varieties of this material are used in the accelerators and the further discussion will be restricted to ferroxcube 4.

A rough estimate of the importance of eddy-current losses can be obtained by comparing the theoretical skin depth $d = \sqrt{2\varrho/\omega\mu_r\mu_0}$ with the thickness D of the slabs. For eddy currents to be negligible we must have $d \gg D$. In our case D is around 1 or 2 cm (cf. fig. 20). For ferroxcube 3 with $\varrho = 1$ Ωm, $\mu_r = 10^3$ we obtain at $\omega = 2\pi f = 2\pi\times10^6$ s^{-1}: $d = 1.7$ cm. For ferroxcube 4, on the other hand, with $\varrho = 4\times10^4$ Ωm, $\mu_r = 100$ we find, at the same frequency, $d = 10$ m. Thus in ferroxcube 3 the eddy current losses cannot be neglected, but in ferroxcube 4 they present no problem.

"Dimensional resonance" is the occurrence of standing-wave effects that become apparent when the wavelength of electromagnetic waves *in* the material becomes so small (owing to very large values of ε_r and μ_r) that a half-wavelength fits into the smaller dimension of the core perpendicular to the magnetic flux. For $\varepsilon_r = 50\,000$, $\mu_r = 1000$, $f = 10^6$ Hz the wavelength is $\lambda = c/f\sqrt{\varepsilon_r\mu_r} \approx 4$ cm, so that in ferroxcube 3 the condition for dimensional resonance, $D \approx \lambda/2$, is easily approached. This effect includes a decrease in the apparent permeability and a decrease in the apparent Q. For ferroxcube 4, where typical values are $\varepsilon_r = 10$, $\mu_r = 100$, we have $\lambda = 10$ m at $f = 10^6$ Hz and there is no danger of dimensional resonance.

Magnetic properties

The magnetic properties that are of practical interest for the present discussion — permeability, variability of permeability by a bias field, and magnetic losses — are closely interrelated, and are of course related to other magnetic properties such as magnetic anisotropy, coercivity and saturation magnetization. For a qualitative understanding of these interrelations we shall occasionally have to go into the microscopic phenomena underlying the magnetization processes involved.

a) Permeability

The permeability μ_r used in the preceding sections is in fact the relative *incremental permeability*; it is the permeability for oscillating fields whose amplitude is small compared to the amplitude of the polarizing field. If the ferrite is polarized by a (quasistatic) bias field that cycles between a large positive and a large negative value, it describes a hysteresis loop in a *B-H* diagram (*fig. 11*). If in a certain state of polariza-

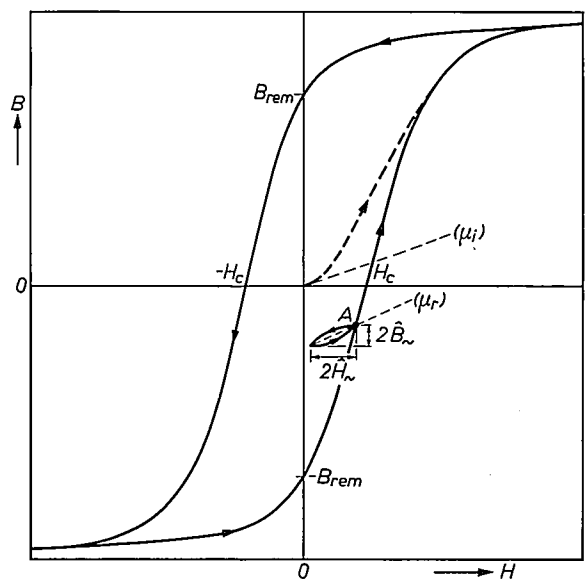


Fig. 11. The hysteresis loop traced by a ferromagnetic material when subjected to a field that varies between large positive and negative values (schematic). If in some state A an r.f. field of relatively small amplitude $\hat{H}_\sim$ is applied, the substance traces a loop that is narrow when the losses are small. The *incremental permeability* μ_r is the slope of this loop, $\mu_r = \hat{B}_\sim/\mu_0\hat{H}_\sim$. In the demagnetized state, for vanishing $H_\sim$, the *initial permeability* μ_i is obtained. When the material is magnetized starting from the demagnetized state, it follows the "virginal curve" (dashed), the incremental permeability (not directly related to the slope of the curve) decreasing steadily. H_c is the coercivity, B_{rem} the flux density at remanence.

tion (e.g. A in fig. 11), a small (parallel) oscillating field is superimposed on the bias field, the ferrite describes a minor loop, and the slope of this loop — which evidently depends on the polarization — is the relative incremental permeability $\mu_r = \hat{B}_\sim/\mu_0\hat{H}_\sim$. A special case is the *initial permeability* μ_i, the μ_r in the demagnetized state, for vanishing $\hat{H}_\sim$ [*]. Fig. 11 may also serve to recall the notions of *coercivity* H_c and the flux density at *remanence* B_{rem}.

Microscopically, the material is divided into many magnetic domains; each domain is magnetically saturated. Within each domain the magnetization has certain preferred directions, due to magnetic anisotropy of the material. In the demagnetized state the domain magnetizations are oriented randomly and the average magnetization is zero. Magnetization of the body as a whole proceeds by two processes: a) by rotation of the magnetization within each domain from the preferred direction; b) by motion of the walls between the domains; during the latter process a domain that is magnetized advantageously with respect to the field grows at the expense of adjacent domains. In very strong magnetic fields all the elementary magnetic moments are aligned and the material is saturated; the *saturation magnetic polarization* J_s is determined by magnitude and concentration of the elementary magnetic moments. Hysteresis occurs because some processes, e.g. the domain-wall motion, are not strictly reversible.

The magnetic anisotropy may be due to the anisotropy of the crystal structure (giving rise to the "magneto-crystalline anisotropy") as well as to internal stresses in the material. A local change in magnetization is "advantageous" if the local change in dimensions that accompanies it (because of the existence of *magnetostriction*), is such as to yield to the local stress. The total anisotropy is measured by the *anisotropy constant* K, or by the *anisotropy field* $H_A = 2K/J_s$. A local magnetic moment is bound to its preferred direction as if by a magnetic field H_A along this direction.

The varieties of ferroxcube 4 can be represented by the chemical formula $Ni_{(1-x)}Zn_xFe_2O_4$. When the synchrotrons were designed five varieties were readily available. In these varieties, indicated by 4A to 4E, x in the formula ranges from 0.64 to 0. (Many more commercial varieties are available today [9].) The change in properties that corresponds to a progress from the variety A via B, C and D towards E can qualitatively be described by saying that one progresses from the "softer" to the "harder" material. In the softer ferrite (A) the internal forces are relatively small, implying that the anisotropy is small and the domain walls are loosely bound. Thus the magnetic moments easily yield to the external field, giving a high permeability; see *Table III*. Also the coercivity

Table III. Typical values of parameters for the ferroxcube 4 series. These values are mainly from reference [10]. For the coercivity and the high-frequency dielectric constant see also reference [11].

Ferroxcube		4A	4B	4C	4D	4E
Saturation magnetic polarization J_s	tesla [*]	0.3665	0.4170	0.4030	0.3550	0.2460
Initial (relative) permeability μ_i		650	230	90	45	17
tan δ at 0.1 MHz (for $\hat{B}_\sim$ small)		0.013	0.015	0.010	0.006	0.006
Frequency for which tan $\delta = 0.1$ (for $\hat{B}_\sim$ small)	MHz	1.5	5.5	16	29	60
Resistivity ϱ	Ωm	$>4\times10^4$	$>4\times10^4$	$>4\times10^4$	$>4\times10^4$	$>4\times10^4$
Coercivity H_c	A/m	30	120	300	650	1100
	(Oe)	(0.4)	(1.5)	(4)	(8)	(14)
High frequency (relative) dielectric constant ε_r		15	15	15	15	15
Curie point	°C	140	265	385	465	585

[*] 1 tesla = 10^4 gauss.

[*] The μ's in this article are relative permeabilities. This is indicated by the subscript r, but the r is omitted when primes or further subscripts are added.

[8] F. G. Brockman, P. H. Dowling and W. G. Steneck, Phys. Rev. 77, 85-93, 1950. See also reference [1].

[9] See for example Philips Data Handbook, Components and Materials Part 4, March 1969.

[10] J. J. Went and E. W. Gorter, The magnetic and electrical properties of ferroxcube materials, Philips tech. Rev. 13, 181-193, 1951/52.

[11] "Ferroxcube", issued by the Philips Industrial Components and Materials Division, pages 18, 20, 21.

is small and the Curie point is low. Hysteresis loops of the softest (A) and the hardest (E) variety are shown in *fig. 12*. Physically, the permeability is also determined by magnitude and concentration of the elementary magnetic moments, i.e. by the saturation polarization J_s. The dependence of J_s upon composition, however, is not very pronounced, and the variation of μ_r with composition is largely a matter of the material being softer or harder.

For the present purpose, the permeability must be sufficiently high, in other words, *the variety must be sufficiently soft.*

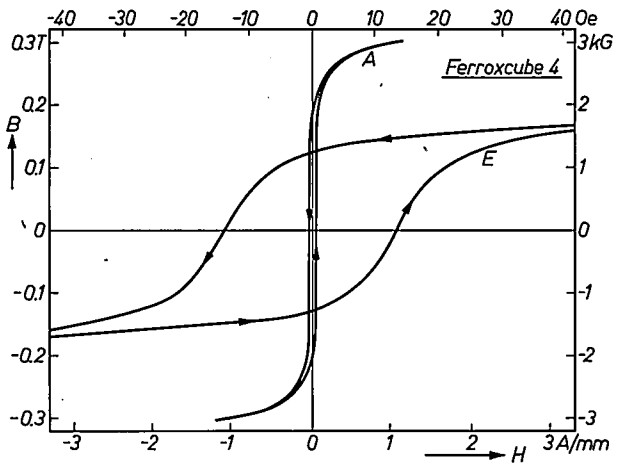


Fig. 12. The hysteresis loops of ferroxcube 4A and 4E.

b) Variation in permeability by polarization

The harder the ferrite, the larger the bias field that is needed to drive it into saturation (see fig. 12), and therefore the larger the bias field needed to obtain a given reduction factor in incremental permeability. This is found to be true experimentally for the five ferroxcube 4 varieties, as is shown in *fig. 13*. This is therefore another reason why the ferrite must be *sufficiently soft.*

The cycle of the bias field may be chosen in several ways. For convenience it may be cycled between zero and a positive value only, but much larger variations in μ can be obtained if the field is also reversed in the cycle. To illustrate this, the results obtained with various cycles for a sample of ferroxcube 4C are shown in *fig. 14*. In fig. 14*a* the field H is cycled between large positive and negative values, and a peak value of μ_r is obtained that is somewhat smaller than the initial permeability μ_i. In fig. 14*b*, after magnetization starting from the demagnetized state, H is cycled between zero and positive values only, and the peak value of μ_r in the cycle is the *permeability at remanence*, μ_{rem}, which is considerably smaller than the peak value in fig. 14*a*. Finally, in fig. 14*c*, it is shown that a peak permeability very close to μ_i can be obtained if H is cycled between a large positive and a small negative value (a little larger than the coercivity H_c). In all of the machines except the PPA the field is cycled between zero and a positive value only (as in fig. 14*b*, but for different grades of ferroxcube 4), and the peak permeability is μ_{rem}. In fig. 13 μ_{rem} is also indicated for the ferroxcube 4 varieties. In the PPA, however, a cycle of the type of fig. 14*c* has been used, so that the largest value of μ_r that can be obtained is exploited. (The choice of this cycle was also prompted by other considerations.) The hysteresis loop that is actually traced at PPA is shown in *fig. 15*.

In the above it has been assumed that the bias and r.f. magnetic field are parallel. Other relative orientations might also be chosen, but the bias field is then less effective in producing a μ_r variation. In the CPS the fields are not parallel in *all* parts of the ferrite (see fig. 7*b*), in the other machines they are.

c) Magnetic losses

Formally, the losses in a magnetic material under r.f. excitation can be described by the *Q factor* or by the *loss angle* δ of the material. A *complex permeability* $\mu_r = \mu' - j\mu''$ is often introduced as well, $\mu'' \neq 0$ then implies a phase lag of $B_\sim$ with respect to $H_\sim$, and μ'' is a measure of the losses. By definition Q, δ, μ'

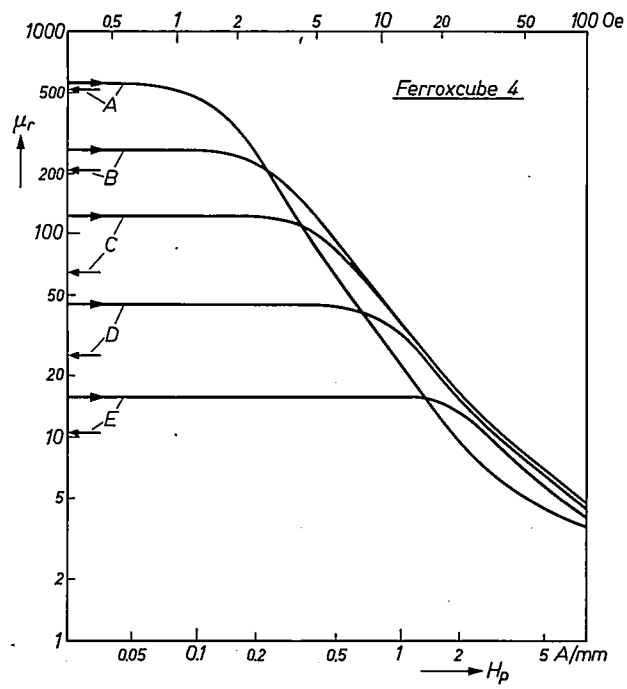


Fig. 13. The incremental permeability μ_r as a function of a polarizing d.c. field H_p for five grades of ferroxcube 4. Each measurement started from the demagnetized state. The remanence values μ_{rem} of μ_r are indicated by the arrows by the vertical axis pointing to the left. In a soft ferrite (μ_i large) a large reduction of μ_r can be obtained with a relatively small polarizing field.

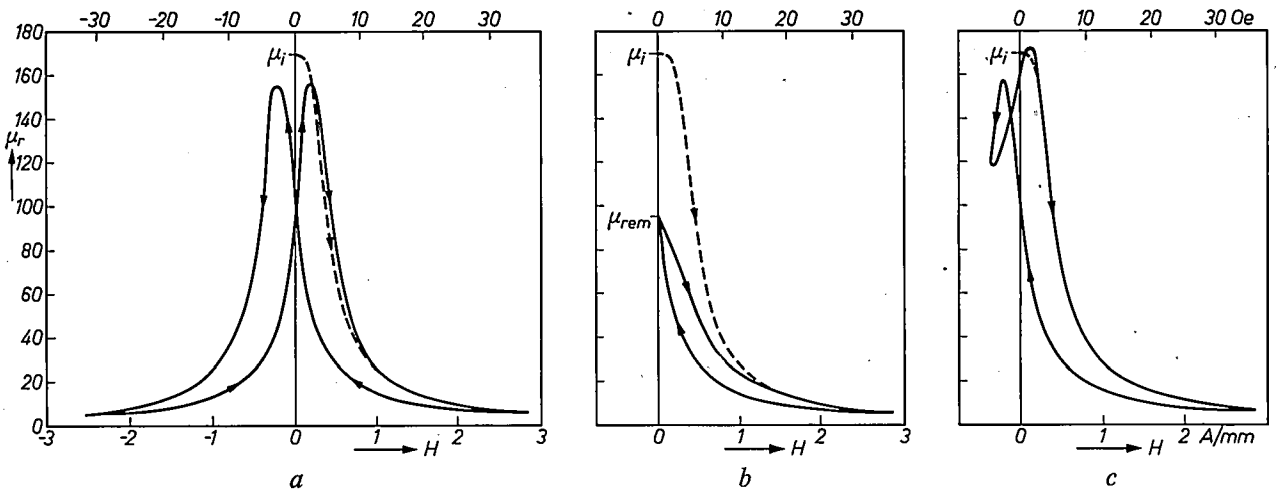


Fig. 14. The variation in μ_r obtained with different cycles. *a*) H is cycled between large positive and negative values. The peak value of μ_r is somewhat below μ_i. *b*) If H is cycled between zero and positive values only, the peak value of μ_r in the cycle is μ_{rem}, considerably smaller than μ_i. *c*) If H is cycled between a large positive value and a small negative value (a little larger than H_c), a peak value close to μ_i is obtained. The curves are the results of measurements on samples of ferroxcube 4C.

and μ'' are interrelated by:

$$1/Q = \tan\delta = \mu''/\mu'. \quad . \quad . \quad . \quad . \quad (8)$$

The power loss per unit volume P is given by [2]:

$$P = \frac{\omega \hat{B}_\sim \hat{H}_\sim}{2Q} = \tfrac{1}{2}\omega\mu''\mu_0\hat{H}_\sim^2 = \frac{\pi f \hat{B}_\sim^2}{\mu_r\mu_0 Q} \quad . \quad (9)$$

(where it is assumed that $\mu''/\mu' = 1/Q \ll 1$).

It should be noted from the outset that in the materials of interest here Q, δ and μ'' (and, to a lesser degree, μ') are affected by the amplitudes of the r.f. magnetic field and flux density: they are not "material constants" in the proper sense of the word.

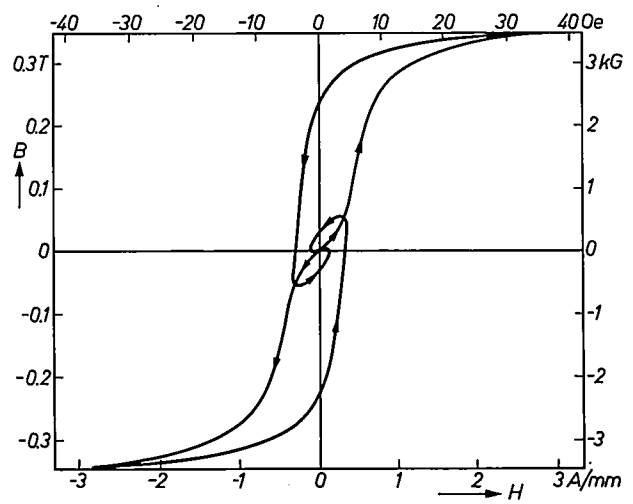


Fig. 15. The *B-H* curve that is traced by ferroxcube 4C in the PPA. The upper and the lower part each correspond to a complete period of acceleration, so that the period of the bias field is twice as long as the main synchrotron period. A high peak value of μ_r is (among other things) obtained by the introduction of the small loops near $B = 0, H = 0$ in the cycle.

To evaluate the losses (and also the permeability) of a ferromagnetic substance, a copper-wire winding can be applied to a toroidal core of the material, and the circuit can be represented by an equivalent circuit of an inductance L and a resistance R in series. Neglecting ohmic losses in the wire, R then represents the losses in the material. The Q factor is then related to R and L by $Q = \omega L/R$.

The *physical* causes of the r.f. losses in ferromagnetic materials constitute a problem that for many substances is still an object of research. Traditionally one distinguishes *hysteresis loss* (related for instance to irreversible domain-wall motions), *eddy-current loss* (related to induced currents that occur when the material is conducting) and *residual loss*. This distinction can be established experimentally because of the differences in dependence upon f and $\hat{B}_\sim$. If $\hat{B}_\sim$ is sufficiently small, one then writes [12]:

$$2\pi/\mu_r Q = a\hat{B}_\sim + ef + c, \quad . \quad . \quad . \quad (10)$$

the three terms representing the three kinds of losses. This kind of analysis has been successful in magnetic metals and also in some ferrites at the lower frequencies, but it loses its point when the "constants" a, e and c turn out to depend upon f and $\hat{B}_\sim$, as is the case with ferrites at high frequencies. In this case a more fruitful line of attack is to study the losses as a function of the frequency at low amplitudes of the r.f. magnetic field and flux density, leading to well-defined Q values. As mentioned, the Q factor will usually depend upon the amplitude at larger amplitudes. Most of the experimen-

[12] V. E. Legg, Magnetic measurements at low flux densities using the alternating current bridge, Bell Syst. tech. J. **15**, 39-62, 1936.

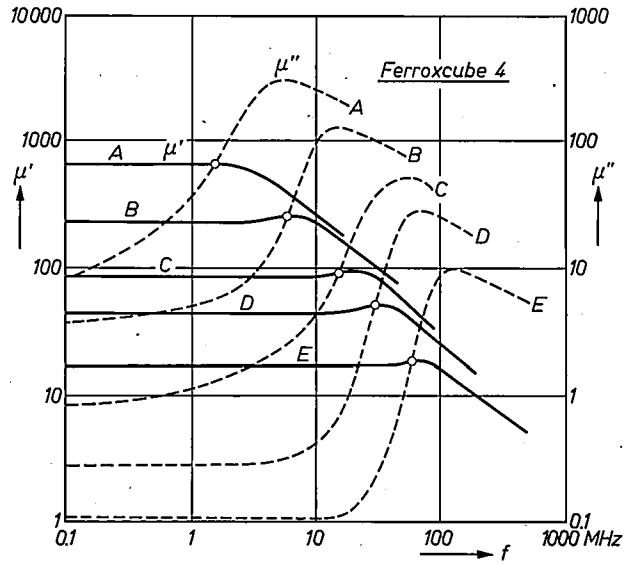


Fig. 16. Real part μ' and imaginary part μ'' of the permeability as a function of frequency for ferroxcube 4A, B, C, D and E [10]. The peak in μ'' is due to ferromagnetic resonance absorption. The approach of resonance is marked by the frequency at which $\tan\delta = 0.1$ ($Q = \mu'/\mu'' = 10$). As the scales for μ' and μ'' differ by a factor of 10, this frequency is found at the intersection of the μ' and the μ'' curves (see also Table III).

tal results discussed below refer to the low-amplitude case.

The losses are very sensitive to the method of preparation, and we shall return to this point when discussing the technology of ferrites. Here, however, with respect to the influence of the nominal chemical composition, it may be noted that in general there are no large *systematic* differences in Q factor for the varieties A, B, . . . E, *except in one respect*: the loss factor (for small $\hat{B}_\sim$) in all ferrites shows a large increase at some frequency and this increase sets in at higher frequencies the harder the ferrite. In terms of μ'' this is illustrated in *fig. 16*. It has been attributed by Snoek [13] to an intrinsic *ferromagnetic resonance*, the essence of which is as follows. If it were not for strong damping, a domain magnetization would perform a free precessional motion ("Larmor precession") around its preferred direction with an angular frequency ω_r determined by the anisotropy field H_A:

$$\omega_r = \gamma H_A. \quad \ldots\ldots \quad (11)$$

The constant γ is the gyromagnetic ratio. If the angular frequency ω of the applied r.f. field approaches ω_r, the precessional motion will be excited by the r.f. field and the accompanying absorption has a maximum for $\omega = \omega_r$. If the ferrite is hard, H_A will be high, and the resonance absorption will occur at a high frequency. In order to avoid this cause of losses the frequency of application should be chosen well below the resonance frequency. At a given frequency this means that, from this point of view, *the harder material is to be preferred.*

The Q factor generally *decreases* for increasing $\hat{B}_\sim$. This fact is very important when choosing a ferrite. Many materials with extremely high Q at low r.f. flux densities become very lossy as the r.f. flux density increases. Therefore, in the specifications for the ferrite, it is often the value of Q (or $\mu_r Q$ or $\mu_r Q\omega$) *at a given value of* $\hat{B}_\sim$ that is specified, and this may very well be the most difficult specification to satisfy.

If the loss of quality due to the fact that $\hat{B}_\sim$ is large is prohibitive, the required $\hat{B}_\sim$ may in principle be reduced by increasing the cross-sectional area A of the cavity (implying the use of more ferrite), because it is in fact a given amplitude $\hat{V}_\sim$ of the gap voltage that is required (cf. eq. 2).

d) The value of $\mu_r Q\omega$

As we have seen before it is the value of $\mu_r Q\omega$ that is required to be large. This value is shown in *fig. 17*

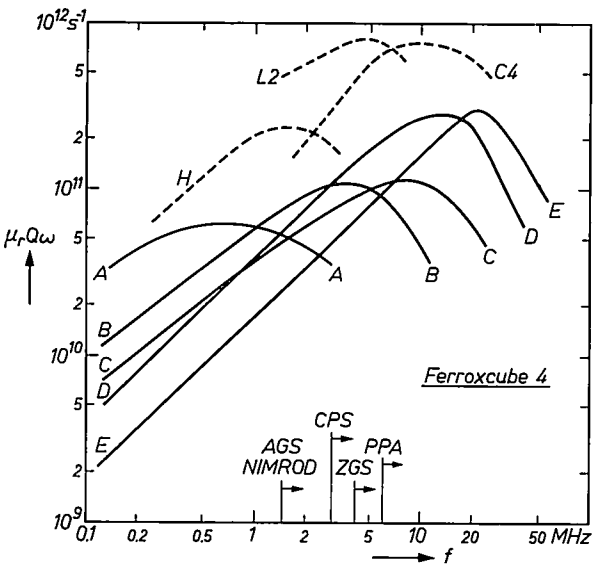


Fig. 17. $\mu_r Q\omega$ as a function of the frequency f for the five grades of ferroxcube 4A, B, C, D and E (unpolarized, and for vanishing $\hat{B}_\sim$), for ferroxcube 4H (developed for the AGS and the CPS), and for some more recently developed materials, 4L2 and 4C4. The lowest frequency of the accelerating voltage in each of the machines (cf. Table V) is also indicated. (The lowest frequency indicated for the PPA refers to the cavity proper and not to the drift tube.)

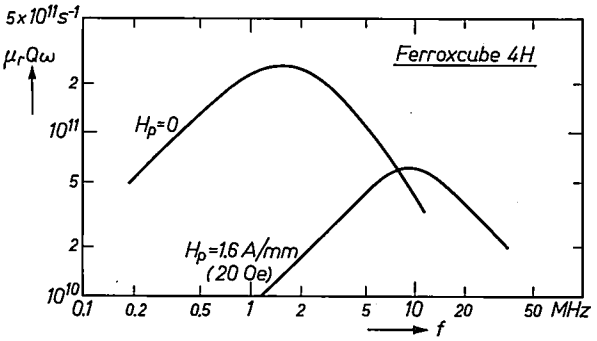


Fig. 18. $\mu_r Q\omega$ as a function of the frequency f for ferroxcube 4H (for vanishing $\hat{B}_\sim$), in the unpolarized state and in a polarizing field of 1.6 A/mm (20 Oe).

as a function of frequency for ordinary samples of ferroxcube 4A, B, C, D and E, in the unpolarized state, for vanishing values of $\hat{B}_\sim$. The rise at the lower frequencies is due to the factor ω, since μ_r and Q are roughly constant; the leveling off and decrease at the higher frequencies is due to the approach of ferromagnetic resonance. For comparison, $\mu_r Q\omega$ curves of ferroxcube 4H, a variety that will be discussed below, and of some materials that have been developed more recently (4L2 and 4C4) are also shown in fig. 17.

When a ferrite sample is polarized, the $\mu_r Q\omega$ curve shifts downwards and to the right. This is shown in *fig. 18* for ferroxcube 4H. A downward shift would be expected because of the decrease in μ. The shift to the right may be explained by the increase of the *total* effective local field H_{eff}, implying an increase of the ferromagnetic resonance frequency $\omega_r = \gamma H_{eff}$.

The curves of figs. 17 and 18 were obtained using low values of $\hat{B}_\sim$. During operation in the accelerator, however, the values of $\hat{B}_\sim$ are not very low, and $\mu_r Q\omega$ will be smaller than figs. 17 and 18 indicate (see the preceding section). This effect will be larger at injection (unpolarized ferrite, upper lefthand curve of fig. 18) than at the end of the cycle (polarized ferrite, lower right-hand curve of fig. 18) because, by eq. (7), $\hat{B}_\sim$ is larger at injection than at the end of the cycle. Thus the difference in the $\mu_r Q\omega$ peak values of the two curves in fig. 18 will be compensated to some extent. As a matter of fact, it has been observed experimentally that, under the conditions (6) and (7), the quantity $\mu_r Q\omega$ remains essentially constant [14].

The choice of a ferroxcube variety for a particular synchrotron

We shall now show from the above that the ferrite varieties that were actually chosen are well adapted to the requirements for the synchrotrons. The composition of these materials, together with those of the commercial varieties 4A-4E, are shown in *Table IV*. In *Table V* the frequency range and the frequency swing of each synchrotron are shown. From the latter, using eq. (6), a specification for a minimum permeability swing is derived. The specified test value of μ_{max}/μ_{min} is often larger than the $(f_t/f_i)^2$ value of the machine, to allow

Table IV. Approximate composition of commercial and synchrotron varieties of ferroxcube 4.

Ferroxcube	4K1	4A	4H	4L1	4B	4C	4D	4E
Main components (mol %)								
NiO	15	18	21.5	23.9	25	32	40	50
ZnO	32.5	32	28.5	25.9	25	18	10	0
Fe_2O_3	49	50	50	50	50	50	50	50
Additions (mol %)								
MnO	1							
Mn_2O_3				0.24				
CuO	2.5							
Used in	Nimrod		AGS CPS	ZGS		PPA		

Table V. The synchrotron frequency range [3] and the required variation in μ.

	AGS	CPS	ZGS	Nimrod	PPA	
					drift tube	cavity
Frequency at injection f_i in MHz	1.4	2.9	4.4	1.4	2.7	6
Final frequency f_t in MHz	4.45	9.5	14	8.0	6.5	30
Frequency ratio f_t/f_i	3.2	3.2	3.2	5.6		5
$(f_t/f_i)^2$	10.2	10.2	10.2	31		25
Specified μ_{max}/μ_{min}	10	25	12	45		
Specified upper limit for $H_{p\,max}$ in A/mm	1.2	2.0	2.8	1.5		
Specified upper limit for $H_{p\,max}$ (in Oe)	(15)	(25)	(35)	(19)		

[13] J. L. Snoek, Dispersion and absorption in magnetic ferrites at frequencies above one Mc/s, Physica **14**, 207-217, 1948. See also: H. G. Beljers and J. L. Snoek, Gyromagnetic phenomena occurring with ferrites, Philips tech. Rev. **11**, 313-322, 1949/50.

[14] See for instance: G. Rakowsky, RF accelerating cavities for AGS conversion, IEEE Trans. **NS-14**, No. 3, 315-319, 1967.

a margin for incomplete filling of the inductor with ferrite, for differences between test and actual operating conditions, and to leave some scope for altering the operating conditions, etc. The largest μ_r variation required is at Nimrod, the machine with the largest frequency swing achieved by ferrite tuning in one step.

Thousands of amperes may be required to tune the ferrite (see Table I), and a limit is usually specified for the field that will be necessary to obtain the tuning (6th row of Table V). In *fig. 19* these specifications have been combined with the μ_r-H_p curves of fig. 13, taking into account that the reduction in μ_r is usually to be obtained starting from μ_{rem}.

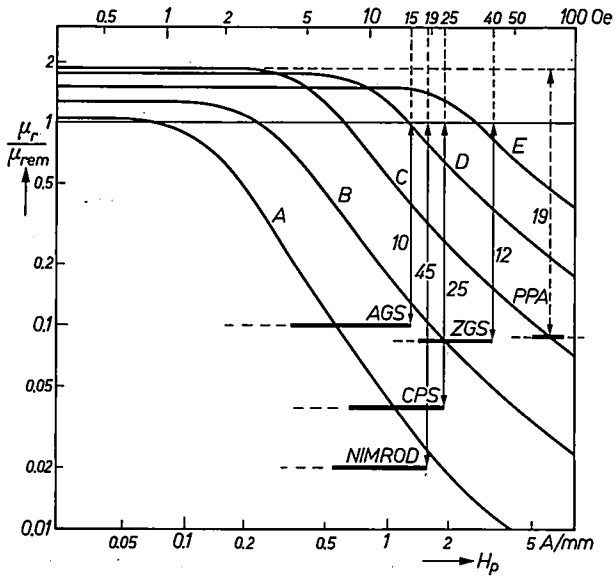


Fig. 19. The variation of μ_r/μ_{rem} with polarizing field H_p for ferroxcube 4A, B, C, D and E according to fig. 13, combined with the specified μ_r variation (indicated by the double arrow) and the upper limit for the polarizing field (end of heavy line) for each synchrotron according to Table V. The μ_r variation should be obtained starting from μ_{rem} (i.e. μ_r/μ_{rem} starting from 1), except for the PPA (see text). A ferrite will comply with the specifications if its μ_r-H_p curve intersects the heavy line.

The maximum of $\mu_r Q\omega$ of the unpolarized ferrite should preferably occur at the lowest operating frequency of the machine. Thus, comparing Table V with fig. 17, for AGS and CPS a grade is indicated between A and B. In fig. 19 it is seen that in such a grade the required μ variation could be obtained with a polarizing field not exceeding the specified upper limit. For AGS and CPS a ferrite between A and B, ferroxcube 4H, has been developed (see Table IV). Fig. 18 shows that it is very well adapted especially to CPS, as in the polarized state $\mu_r Q\omega$ has again its maximum near the operating frequency (10 MHz).

For Nimrod the lowest frequency also pointed towards a ferrite between A and B. The μ variation of none, however, was large enough (fig. 19): a grade "before A" was required. The ferrite that has been developed for Nimrod (4K1) can indeed be considered as such (Table IV). The reason for the addition of other components will be indicated below.

For ZGS the lowest frequency indicated ferroxcube 4B (fig. 17), and this gave a sufficient μ variation too (fig. 19). The ferrite for ZGS, 4L1, is indeed quite near B (Table IV).

At PPA the cavity must cover the second part (6-30 MHz) of the frequency range of the machine. The lowest frequency of this part, 6 MHz, suggests ferroxcube 4C for the cavity ferrite. As a maximum value for H_p is not prescribed, this grade could be and is indeed used. However, a special, rather complicated, cycle of polarization is used, as mentioned earlier (fig. 15); this avoids the necessity of too large H_p values for the required μ variation.

Technology

The relevant properties of the material, permeability and Q factor, are determined not only by the chemical composition but to a high degree by the gross structure. If, for instance, the body is porous, or has non-magnetic inclusions, the permeability may be affected in two ways. a) Internal demagnetization will occur at the pores so that the field in the material proper is smaller than the externally applied field. b) Domain walls tend to be fixed at inclusions so that the contribution to the permeability from domain-wall motion is diminished. If domain-wall motion were eliminated completely, the permeability would be due to rotation only.

The gross structure is largely a matter of preparing the material. The ferrites are made by ceramic techniques and it is in the control of the sintered structure that experience is a major factor. Generally speaking, homogeneity, both in chemical composition and in microstructure, is extremely important. The quality also appears to be very sensitive to the degree of stoichiometry [15]. The more care paid to factors of this kind, the better the material. This is illustrated by comparing the $\mu_r Q\omega$ values of ferroxcube 4H, which was especially developed for the AGS and the CPS, with those of the ordinary ferroxcube 4 varieties (fig. 17). The large dimensions (fig. 20) of these ferrite parts presented the ferrite factory with unusual problems. The PPA ring — the basic element of the PPA core — is the largest single-piece ferrite made anywhere. The pressing

[15] F. G. Brockman and K. E. Matteson, New technique for the precise control of the stoichiometry of nickel zinc ferrite and the resulting high quality product, presented at the 71st Annual Meeting of the American Ceramic Society; abstract given in Amer. Ceramic Soc. Bull. **48**, 430, 1969 (No. 4).

[16] L. G. Van Uitert, Proc. IRE **44**, 1294-1303, 1956.

of such a ring prior to firing requires a hydraulic press rated at 300 tons. Firing is especially difficult for such large parts; the furnace must have a very uniform temperature over the volume of the furnace and the rate of heating must be such that the parts shrink uniformly to prevent cracking. It may be recalled that firing may involve a decrease in the linear dimensions of as much as 25%. Firing cycles of as long as five days were required for some parts.

Several details of the technology are aimed at reducing the losses that are due to the presence of ferrous ions. If a considerable concentration of these are present along with the ferric ions, the electrons appear to jump rather easily between the two kinds of iron ions in the lattice. This mobility of charges gives rise to extra magnetic losses. For large concentrations it is equivalent to a low value of ϱ as well as a high value of ε_r. This explains why ferroxcube 4 (high ϱ, low ε_r) is more suitable than ferroxcube 3: in ferroxcube 3 ferrous ions do belong to the composition nominally, in ferroxcube 4 they do not.

The sintering process is carried out at high temperatures (roughly 1200 °C). A high sintering temperature will lead to a dense product and to a high value of μ_r. At these temperatures there is a tendency to lose oxygen, leading to the formation of ferrous ions and thus to a decrease of the Q factor. To counteract the loss of oxygen the firing is carried out in an oxygen atmosphere but to stop the reaction completely an impracticably high oxygen pressure would be required. A lower pressure is therefore applied and one must rely on reoxidation during the cooling-down period for compensation of the oxygen loss. If, however, the product has become too dense because the sintering temperature was chosen too high, the oxygen may not be able to penetrate sufficiently, resulting in incomplete reoxidation and degradation of the Q factor.

There are other means to counteract the formation of ferrous ions. In the first place the firing temperature leading to a sufficiently dense product may be lowered by the addition of small amounts of a "densifying agent"; copper oxide (used in 4K1, see Table IV) appears to act as such. Secondly, a small amount of manganese oxide may be added [16]. Iron is more readily oxidized than manganese so that Fe^{2+} ions in the presence of Mn^{3+} will be reoxidized by the reaction: $Fe^{2+} + Mn^{3+} \rightarrow Mn^{2+} + Fe^{3+}$. This method was applied when 4K1 and 4L1 were prepared (Table IV).

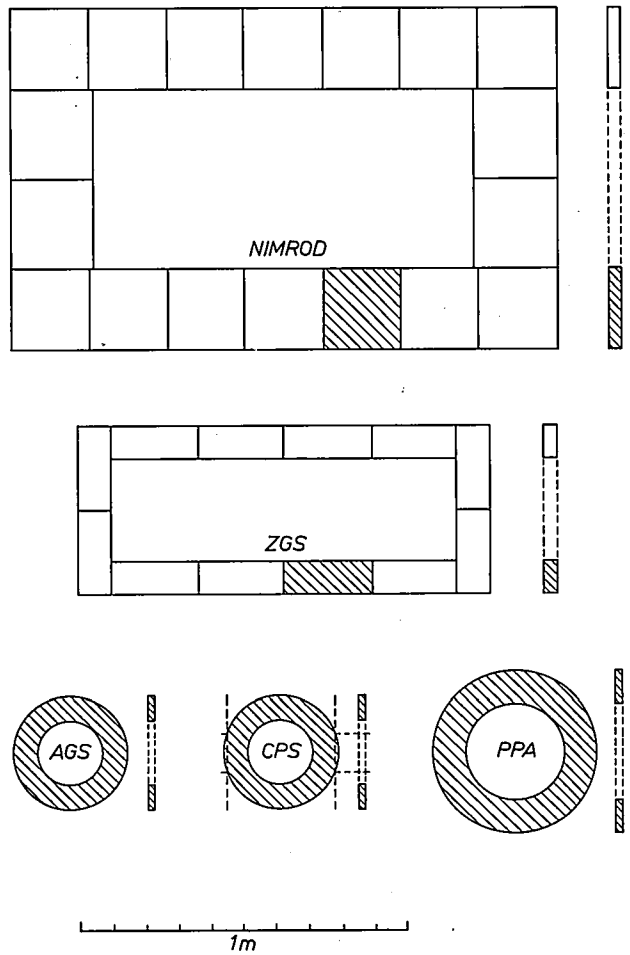


Fig. 20. The ferrite frames and rings that have been manufactured for the five synchrotrons, to scale. For the ZGS and for Nimrod blocks are manufactured, which were assembled into frames by cementing (ZGS) or clamping (Nimrod). At the CPS flat faces were ground on the material after the rings were cemented together into cores.

The final cores

In *fig. 20* the blocks, frames and rings that were manufactured for the five synchrotrons are drawn to scale. The total amount of ferrite for the cavities of each synchrotron, and the way it is distributed in and among the cavities is shown in *Table VI.*

For the CPS the rings were cemented together to

Table VI. Distribution of the ferrite in the synchrotrons.

	number of cavities	per cavity	ferrite mass per cavity (kg)	total ferrite mass (kg)
AGS	12	4 cores = 80 rings	500	6000
CPS	15	2 cores = 90 rings	560	8400
ZGS	1	30 frames	1760	1760
Nimrod	1	28 frames	6000	6000
PPA	4	44 rings [*]	750	3000

[*] 40 rings in the cavity proper, 4 in the drift-tube inductor.

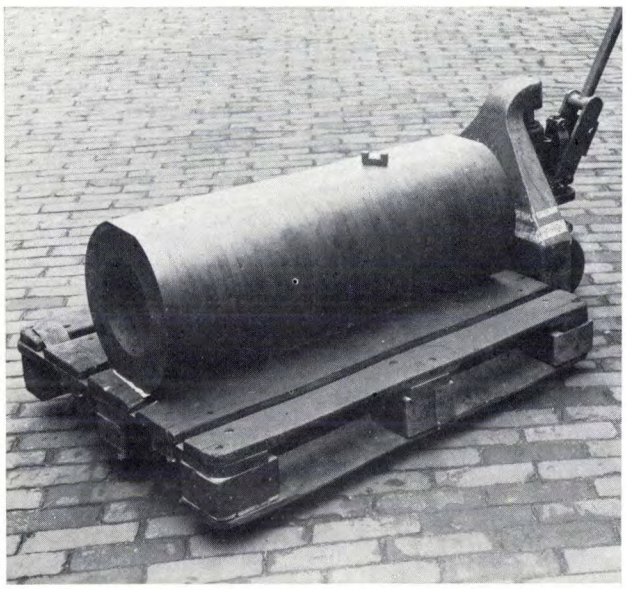


Fig. 21. A ferrite core for the CPS.

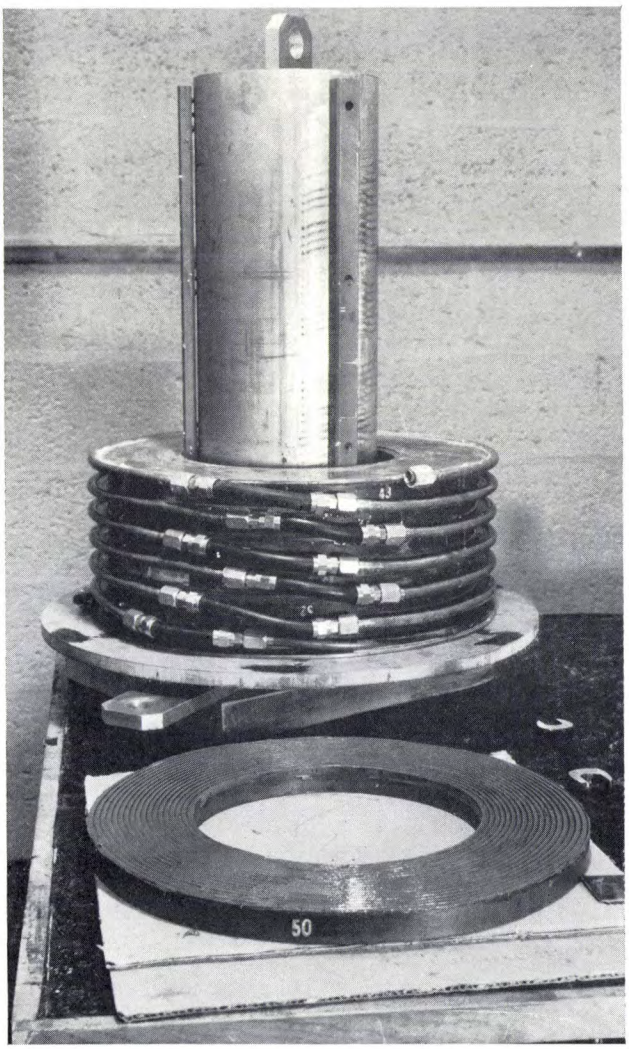


Photo Brookhaven National Laboratory

Fig. 22. A ferrite core of the AGS under construction. The ferrite rings are alternated with water-cooled copper plates.

form a complete core as shown in *fig. 21*. After cementing, flat faces that serve to fit the core into the bias-field external magnet (see fig. 7*b*) were ground on the core. For the AGS copper cooling plates were mounted between the rings (*fig. 22*). In the PPA the rings were mounted separately in the cavities, as were the frames for the ZGS and Nimrod (see figs. 8 and 10). The frames for the ZGS were assembled from large blocks cemented together. (The frames for the Cosmotron were also made in this fashion [1].) For Nimrod this procedure was also applied (*fig. 23*), but as cementing led to some disappointments in the earlier stages, blocks have also been clamped into position in later procedures rather than cemented.

In manufacturing the blocks, rings, frames and complete cores, strict dimensional tolerances had to be kept. This was accomplished by surface grinding using diamond tools. One important purpose of dimensional accuracy is to minimize gaps in the bias and r.f. magnetic circuits, as any gap in the magnetic circuit reduces the apparent permeability of the assembly.

The r.f. input power of a synchrotron is many kilowatts (see Table XII in the following article [3]), and a considerable part of this power is dissipated in the ferrite; therefore the cores must be cooled. As the ferrites are poor heat conductors, all parts must be brought into contact with the cooling water (or oil) as intimately as is feasible. In Nimrod and in the PPA the ferrite frames or rings are mounted with spaces in between (see fig. 8), and every frame or ring is immersed

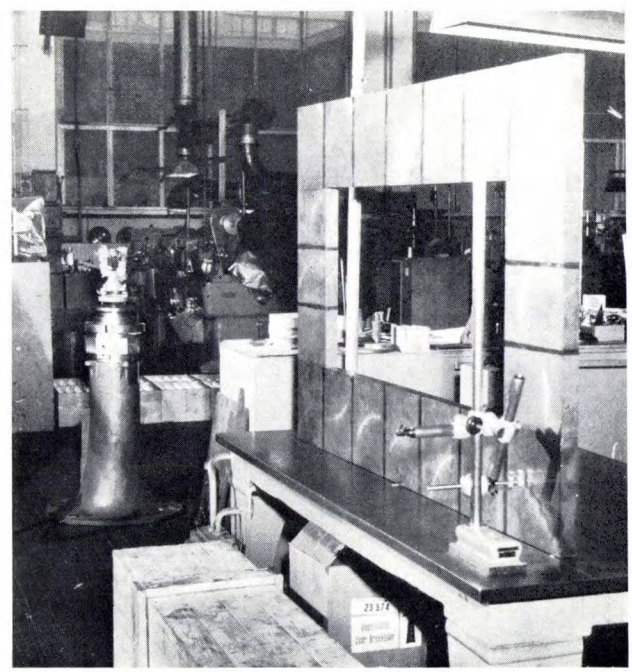


Fig. 23. A Nimrod frame being tested for flatness.

in cooling oil. In the AGS the ferrite rings are cooled by the copper plates mentioned earlier (fig. 22). In the ZGS very flat copper channels, mounted between the frames, carry the cooling water (fig. 10). In the CPS cooling tubes are mounted on a copper cover (which is the cavity wall) on the outside of the core (*fig. 24*). In some cases the dimensional accuracy mentioned above is also of importance for the thermal contact between the core and the cooling structure.

Correlated to the development of heat and the cooling are a few requirements to be met by the ferrite. To prevent "hot spots" in the center the slabs must be sufficiently thin and the thermal conductivity sufficiently large. A typical specification for the latter is > 3 W/m °C. The temperature differences that can build up in the material (e.g. 60 °C between the inside and the outside of the CPS core) will have an effect on μ and Q; therefore the temperature coefficient of these quantities is sometimes specified (e.g. $< 0.5\%$ per °C). Also, the material must be able to stand thermal (and other) stresses; a required tensile strength of 2000 N/cm^2 is typical. To meet such specifications is largely a matter of technology rather than of chemical composition.

Magnetic testing of the large pieces of ferrite at the specified frequency and flux density requires a high-power r.f. test installation. During the earlier stages of selection and development of the materials, a pulse-test method was used at the Philips laboratories in Eindhoven, in order to avoid the need for a high-power r.f. test installation at that stage. In this method the test piece is subjected to short pulses of r.f. power [17]. The instantaneous power can then have the required level whereas the average power remains low. The test piece is made the core of an inductor of an LC circuit, tuned to the required frequency. The free oscillation of this circuit excited by the pulse is displayed on an oscilloscope. From the frequency and the value of the capacitance C in the circuit the permeability is found, and the Q factor is derived from the damping of the oscillations. Later however, a special high-power test equipment was built at the factory. With this equipment the rings, the blocks, and the assembled frames could be tested in the polarized state, at the specified frequency and flux density, before they were delivered.

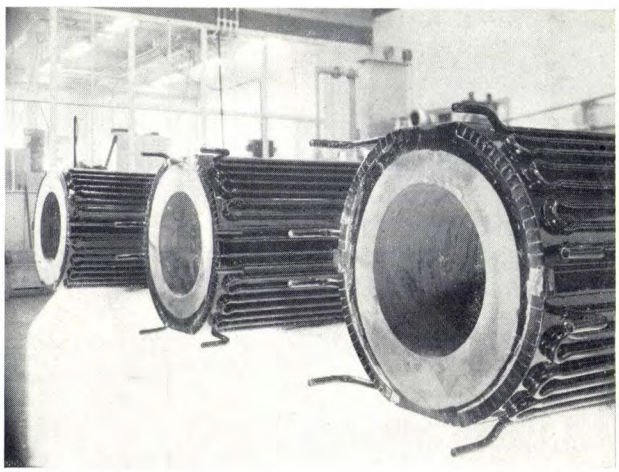


Fig. 24. CPS ferrite cores with water-cooling tubes. The tubes are mounted on a copper cover, which is also the cavity wall.

Summary. Philips have supplied the cores of ferrite material for the accelerating stations of five proton synchrotrons, which are among the largest accelerators in the world. These machines, the AGS at Brookhaven, the CPS at Geneva, the ZGS at Argonne, Nimrod at Chilton and the PPA at Princeton are more fully discussed elsewhere in this issue. The pieces of ferrite are quite large, e.g. rings of 50 cm diameter or blocks (to be assembled into frames) 25 cm long, and considerable quantities are needed: from 1760 kg at ZGS to 8400 kg at CPS. An accelerating station can be considered as a single-turn inductor with the accelerating voltage across its ends. The ferroxcube serves not only as the core material but also as the means of tuning the inductor to the rapidly changing orbital frequency of the protons. The tuning is effected by controlling the incremental permeability μ_r of the ferroxcube by a magnetic biasing field. For this kind of application it is desirable to have a wide range of variation for μ_r and a high value of $\mu_r Q\omega$. The most suitable ferrites then readily available were the nickel-zinc ferrites (a continuous series of compositions, five of which were marketed: the ferroxcubes 4A, 4B, 4C, 4D and 4E). The softest variety (A) has the advantage of a high permeability and a good range of variation, whereas the hardest variety (E) would be more suitable if it is important to avoid ferromagnetic resonance. The optimum composition is different for each synchrotron: between A and B for the AGS and the CPS, softer than A for Nimrod, close to B for the ZGS, and C for the PPA. The material is a ceramic product; the mixing, pressing and firing must be performed very carefully to achieve sufficient homogeneity in gross structure and composition in such large components. To avoid extra losses the formation of ferrous ions is counteracted by firing in an oxygen atmosphere and in some cases by adding chemical densifying or oxidizing agents.

[17] H. van der Heide, Ferroxcube in proton accelerators, Philips Matronics No. 14, 256-264, 1958.

Five major proton synchrotrons

R. Gouiran

Introduction

The preceding article [1], concerned with the ferrites used in large quantities in five of the world's largest particle accelerators, touched only briefly on the accelerators themselves. However, to do justice to the remarkable physical and engineering aspects of these five machines, and to profit from the unique opportunity for a combined description, it seemed desirable to present a more detailed account of the machines here. This presentation, with its opportunity for comparisons, may help to give some idea of the philosophy behind the choice of certain design parameters.

The machines are listed here again for reference:

1) The Alternating Gradient Synchrotron at Brookhaven near New York.
2) The CERN Proton Synchrotron at Geneva.
3) The Zero Gradient Synchrotron at Argonne near Chicago.
4) The "Nimrod" at Chilton near Harwell in England.
5) The Princeton-Pennsylvania Accelerator at Princeton (New Jersey).

Dr. R. Gouiran is with the Proton Synchrotron Machine Division, CERN, Geneva.

The abbreviations AGS, CPS, ZGS, Nimrod and PPA will again be used in this article.

These five accelerators are proton synchrotrons and thus of the same basic design as outlined in the preceding article; the principal elements being an annular vacuum chamber in a circular array of guiding magnets and, between the magnets, provision for accelerating the protons. In appearance and in detail of the design, however, there is a remarkable variety. The difference in size of the ring, for instance, is illustrated in *figs. 1*, *2* and *3*, showing the smallest machine, the PPA (fig. 2), the ring of concrete being made to house the largest machine, the AGS (fig. 3), and the "middle-sized" ZGS (fig. 1). Another contrast is that between the magnet units of the CPS and those of Nimrod; in the CPS there is an alignment of many relatively small magnet units (*fig. 4*), designed for a narrow vacuum chamber, whereas Nimrod has a massive magnet enclosing a huge vacuum chamber (*fig. 5*). In

[1] F. G. Brockman, H. van der Heide and M. W. Louwerse, Ferroxcube for proton synchrotrons, the preceding article in this issue.

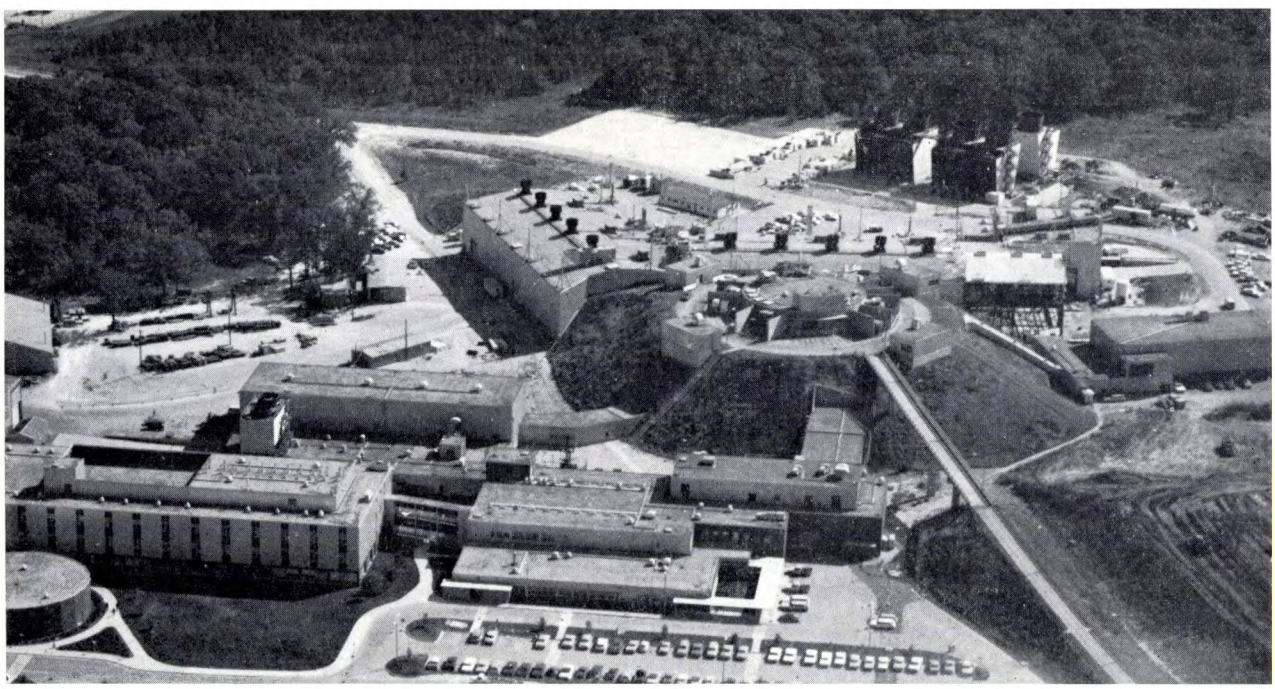
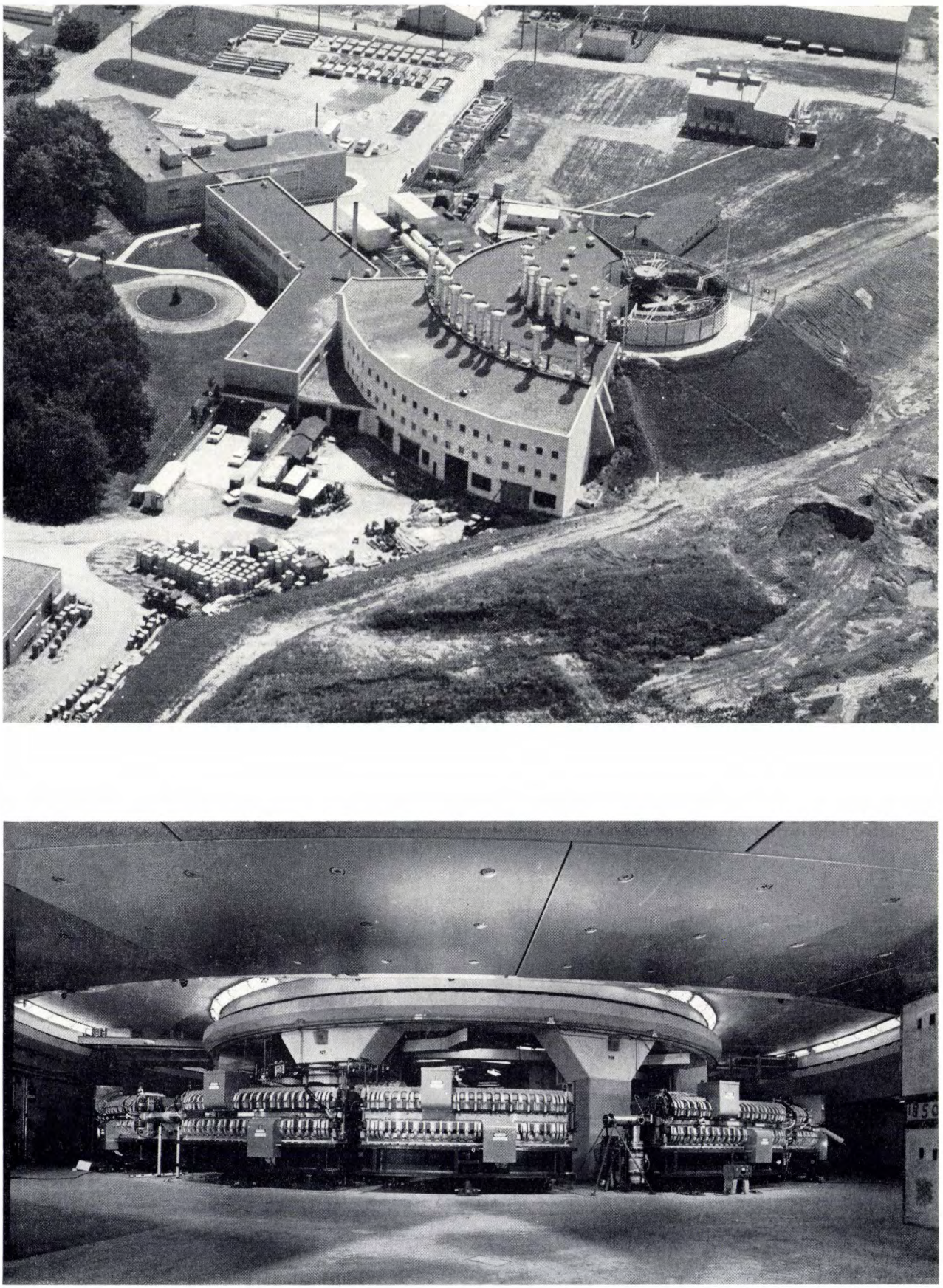


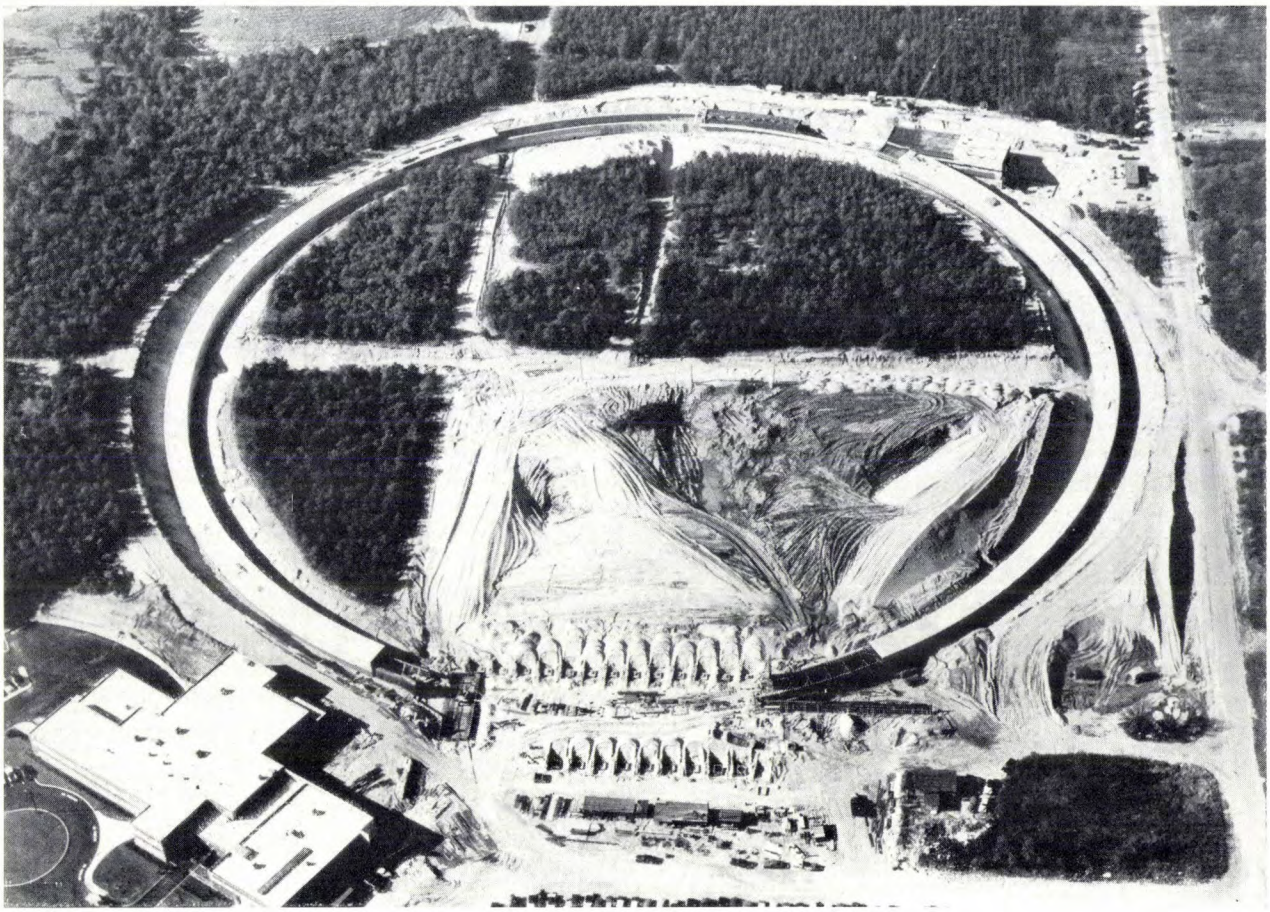
Photo Argonne National Laboratory

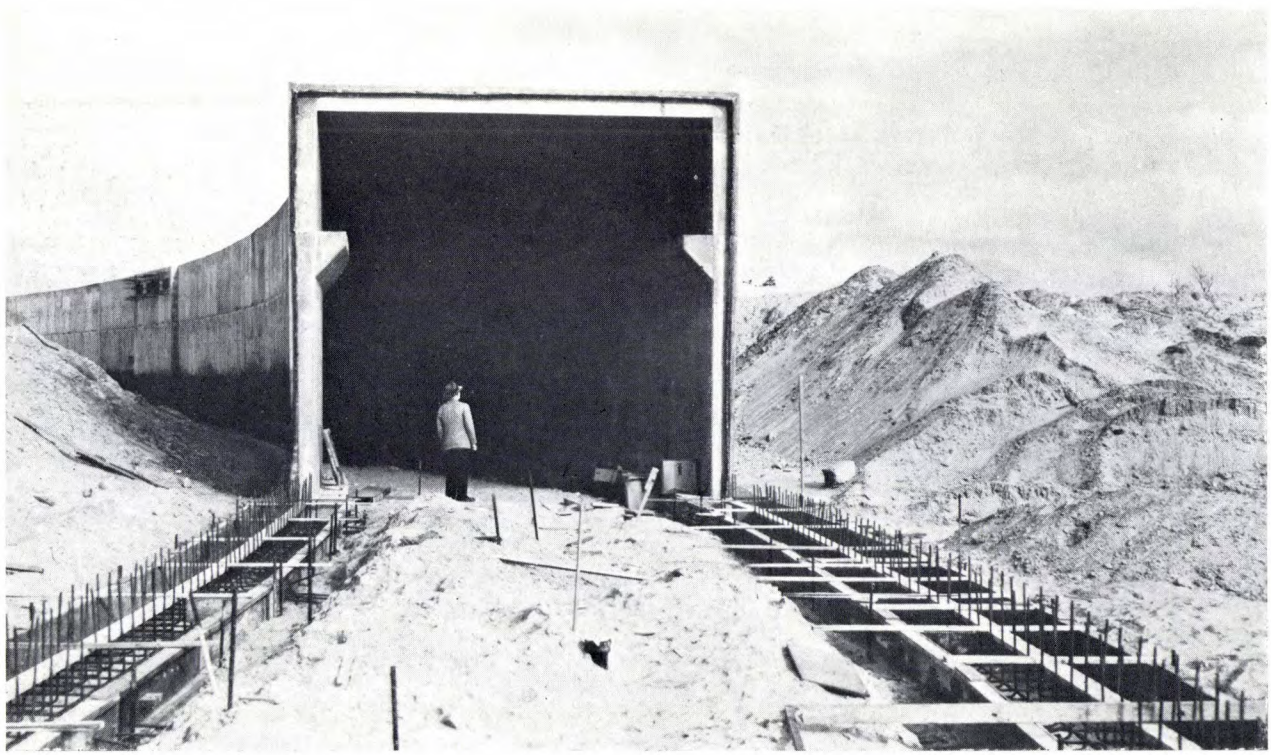
Fig. 1. General view of the building complex for the Zero Gradient Synchrotron at Argonne near Chicago. The accelerator proper is surrounded by a mound of earth.



Photos Princeton-Penn Accelerator

Fig. 2. The Princeton-Pennsylvania Accelerator at Princeton, New Jersey. The upper photograph gives a general view of the building complex. A bank of capacitors forming a resonant system with the synchrotron magnets can be seen on top of the building housing the accelerator. The lower photograph shows the ring of magnets under construction.





Photos Brookhaven National Laboratory

Fig. 3. Two photographs of the ring of concrete to house the Alternating Gradient Synchrotron at Brookhaven near New York.

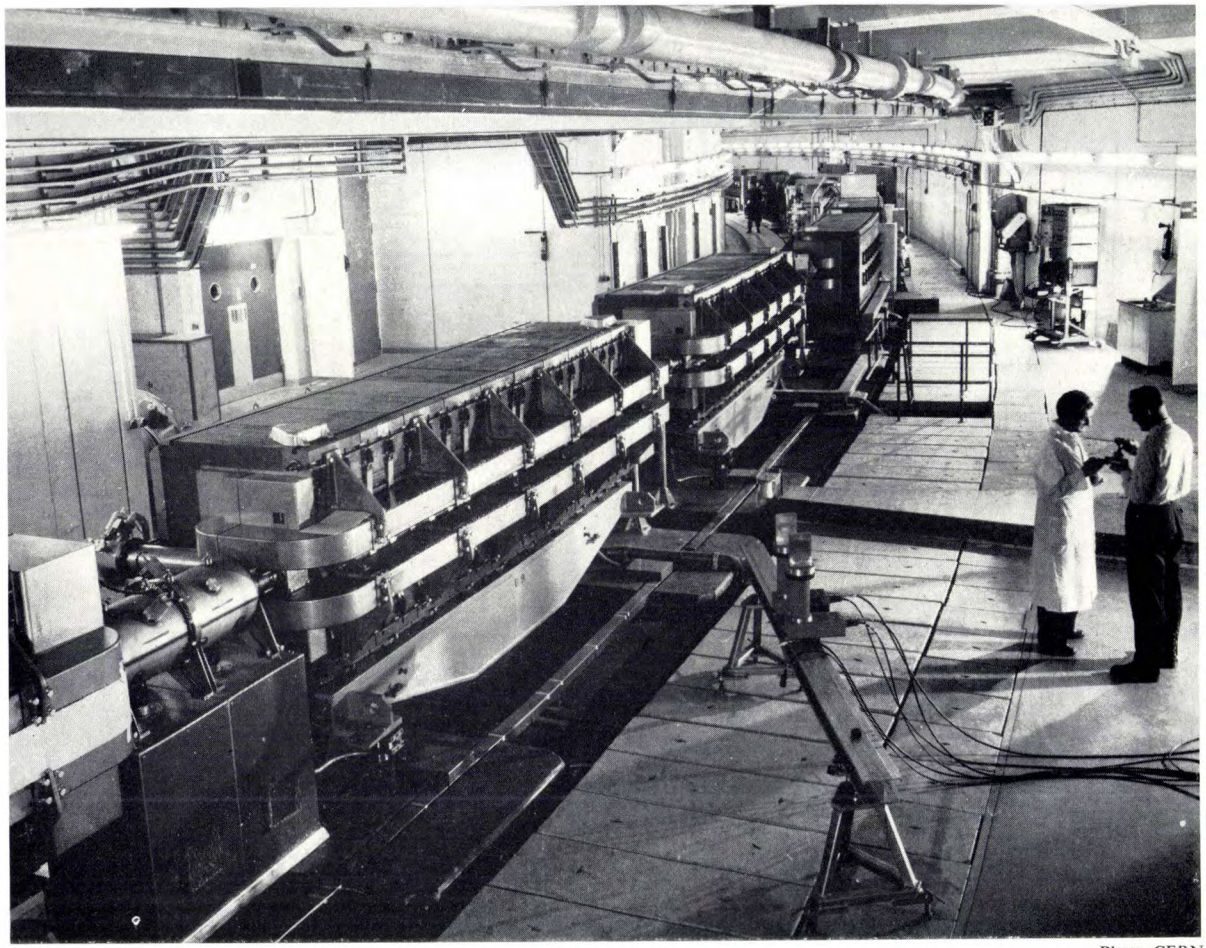


Photo CERN

Fig. 4. A number of magnet units of the CERN Proton Synchrotron at Geneva in 1959 just after commissionering.

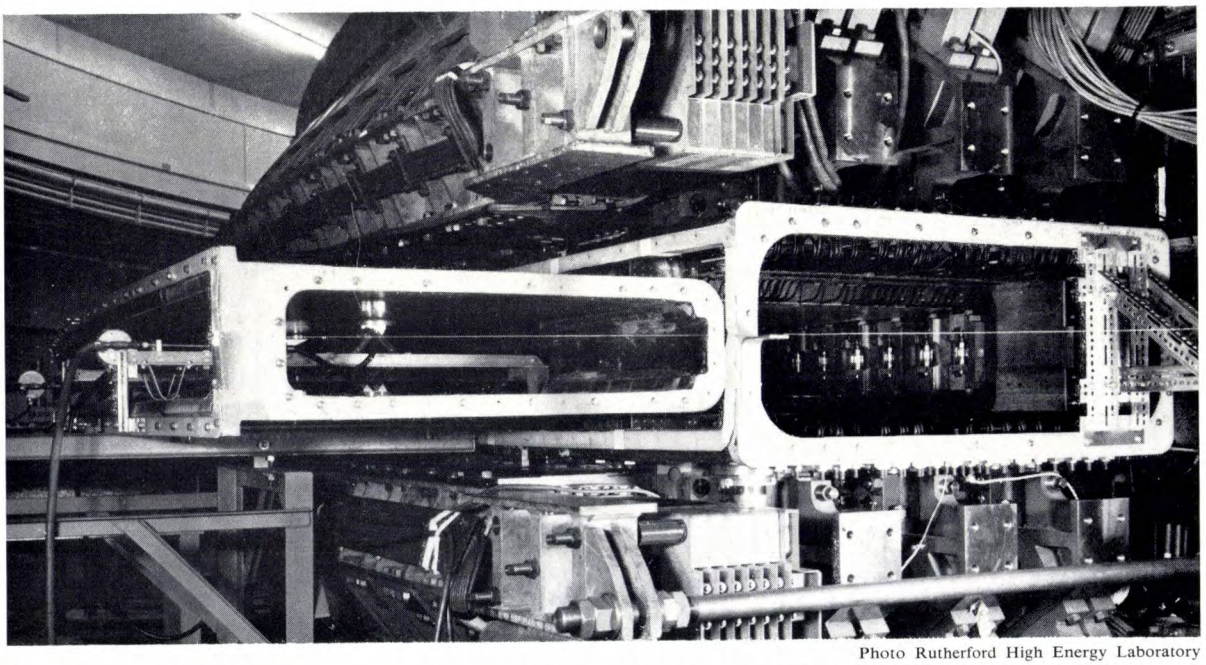


Photo Rutherford High Energy Laboratory

Fig. 5. Section of the double-walled vacuum chamber in the gap of the magnet of Nimrod at Chilton near Harwell, England. The chamber is shown under construction: the inner chamber on the left still has to be moved into the outer chamber.

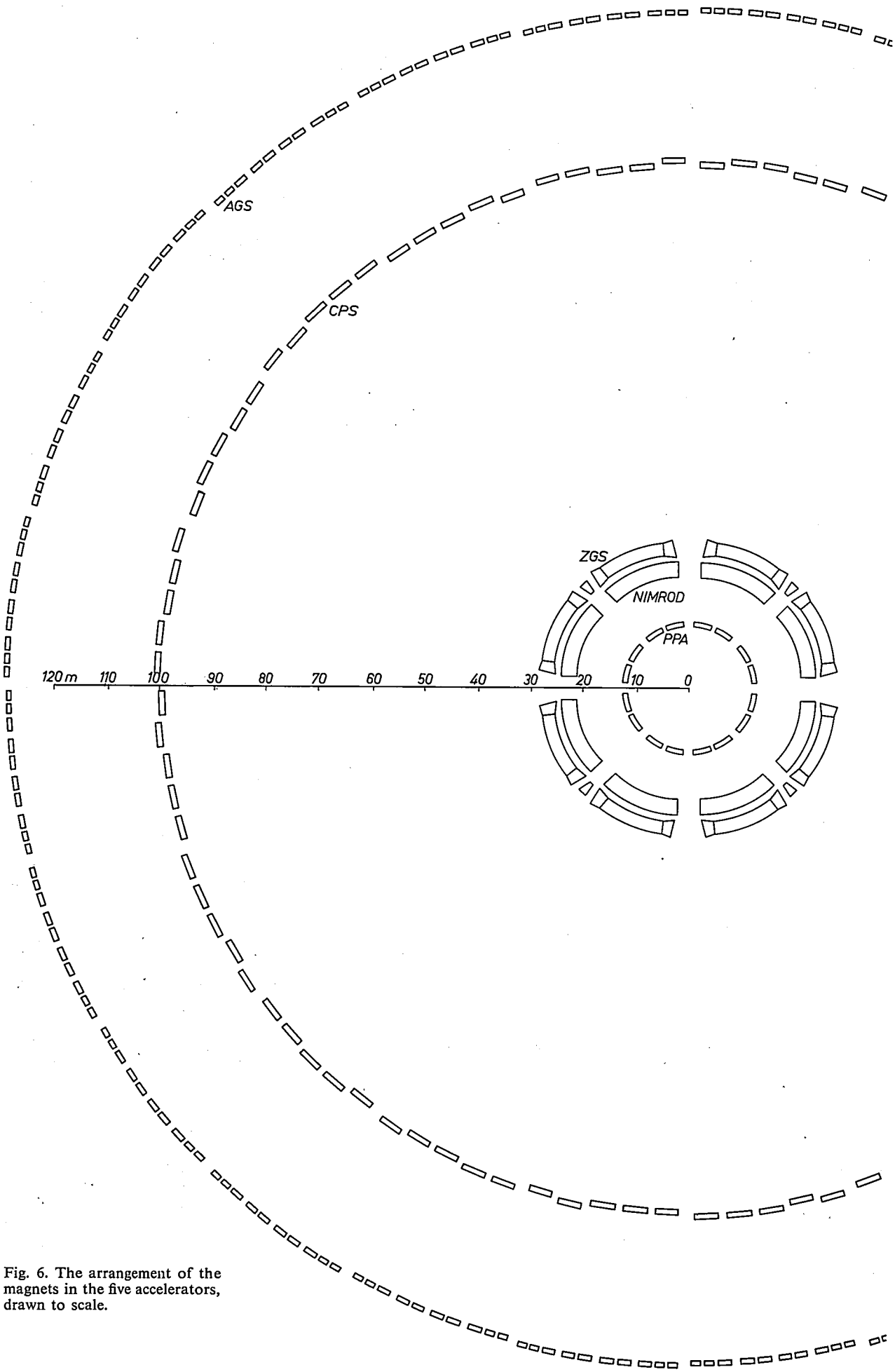


Fig. 6. The arrangement of the magnets in the five accelerators, drawn to scale.

fig. 6 the arrangement of the magnets is drawn on the same scale for all of the machines, and in *fig. 7* the vacuum-chamber cross-sections are all shown on the same scale. Finally, the values of a few characteristic quantities are given in *Table I.*

Table I. General data.

		AGS	CPS	ZGS	Nimrod	PPA
Final particle energy	GeV	33	28	12.7	7	3
Typical cycle period	s	2.5	2.5	2.3	2.6	0.054
Mean radius of the ring	m	128.5	100	27.4	23.6	12.51
Mass of magnet steel	t	4000	3000	4700	7000	400
Date of first acceleration		July 1960	Nov. 1959	Sept. 1963	Aug. 1963	April 1963

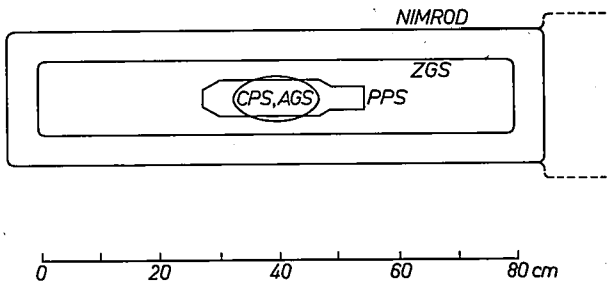


Fig. 7. Vacuum chamber cross-sections drawn to scale.

These considerable differences in design can ultimately be related to considerable differences in the *final energy* and the *intensity* of the beam of particles for which each machine is designed. The maximum field of a conventional bending electromagnet is mainly determined by saturation of the magnet steel, and usually corresponds to less than 2.0 teslas (20 kilogauss), which, together with the final energy, defines the length of the ring (excluding the straight sections). The intensity is mainly determined by the "brightness" of the source and by the "acceptance" of the accelerator — concepts to be dealt with later; as the acceptance depends upon the aperture of the vacuum chamber, a requirement concerning the intensity defines the cross-section of the magnet yoke. From the accelerator diameter and the magnet cross-section the overall cost — much of which is determined by the amount of magnet steel required — can be estimated.

Before reviewing the five synchrotrons in detail (p. 351 *et seq.*), we shall try to set the background by giving a fairly extensive discussion of the general principles of a proton synchrotron. Only the production of the beam of protons will be discussed. One should keep in mind, however, that the only true value of an accelerator comes from the experiments that can be done with it. For these experiments the beam of particles at the end of the acceleration cycle is put to use by means of various techniques (insertion of internal targets, beam ejection). We shall refrain from describing these, but they are nevertheless essential.

A few relativistic formulas

As after acceleration (and sometimes even before) the particles have velocities very close to the velocity of light, relativistic mechanics is needed for a proper description of the particle behaviour. The relativistic formulae that we shall use most frequently are now given.

If m is the particle rest mass, v its velocity and c the velocity of light (3×10^8 m/s), the particle has a momentum p given by:

$$p = \frac{mv}{\sqrt{1-v^2/c^2}} = \beta\gamma mc, \quad \ldots\ldots \quad (1)$$

where $\beta = v/c$ and $\gamma = 1/\sqrt{1-\beta^2}$.

The total energy E is:

$$E = \frac{mc^2}{\sqrt{1-v^2/c^2}} = \gamma mc^2. \quad \ldots\ldots \quad (2)$$

The kinetic energy T is the difference between the total energy and the rest energy mc^2:

$$T = E - mc^2. \quad \ldots\ldots \quad (3)$$

Eliminating the velocity v from (1) and (2), the relation between the energy and the momentum is found to be:

$$E^2 = m^2c^4 + p^2c^2. \quad \ldots\ldots \quad (4)$$

I. General description of a proton synchrotron

Motion of a particle in a magnetic field

The most important relation determining the motion of the proton (of charge e) in the synchrotron is:

$$p = \varrho eB \quad \ldots\ldots \quad (5)$$

between the particle momentum p, the magnetic field of flux density B and the radius of curvature ϱ of the particle orbit. *Fig. 8* shows a simple derivation of this relation, which has already been used in the preceding article.

The final energy of the particle, as (4) and (5) show, is determined by ϱ and by the flux density B_f of the final field, which in turn is determined by saturation of the magnet steel. In practice B_f is usually around 1.4 teslas (14 kilogauss) [2]. It then follows that the larger the final energy required, the larger the radius of curvature should be. For example, if we want an accelerator for protons ($mc^2 = 0.938$ GeV) to give a final total energy of 30 GeV (which then is mainly kinetic, $E \gg mc^2$), we have $pc \approx E = 30$ GeV, and with the B_f value mentioned above the radius is given by $\varrho = p/eB \approx 3 \times 10^{10}/3 \times 10^8 \times 1.4 = 70$ m.

The *mean radius* R_0 of the accelerator is larger than the radius of curvature ϱ, owing to the presence of straight field-free sections. R_0 is defined as $1/2\pi$ times the total circumference, measured along the nominal orbit of the particles.

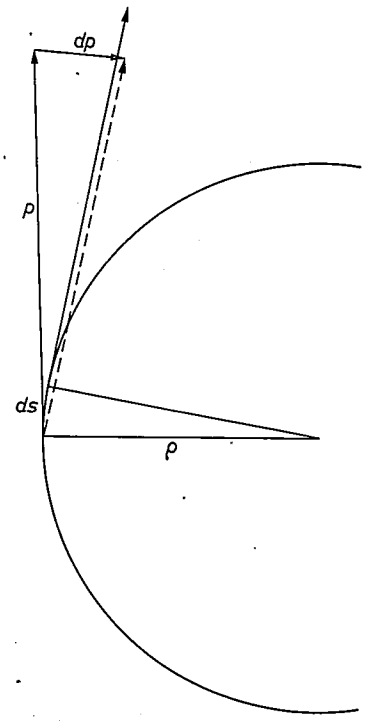


Fig. 8. A particle of charge e, momentum p, in a homogeneous magnetic field of flux density B describes a circular orbit. The change in momentum dp in a time dt is $dp = evBdt$ (evB is the Lorentz force). Since $vdt = ds$ and $dp/p = ds/\varrho$ (the momentum remains perpendicular to the radius vector), this immediately leads to eq. (5).

Acceleration; magnetic field cycle and frequency swing

The particles are accelerated by accelerating stations around the ring, which provide alternating electric fields. As the velocity and the momentum of the particles increase, both the magnetic field and the frequency of the accelerating electric field have to vary in a suitable way: the magnetic field must vary in accordance with eq. (5) to keep the particles in orbit, and the frequency of the electric field must vary in such a way that the field is always in the proper phase to accelerate the particles when they pass the accelerating station.

In practice, as the system providing the magnetic field is inherently a slow-acting one, and thus difficult to regulate, a programme is selected for the magnetic field, and the frequency of the electric field is adjusted to suit this. *Fig. 9* shows a typical magnetic field cycle, with a linear rise during acceleration and a flat top for experimental use. (In the PPA the variation of the

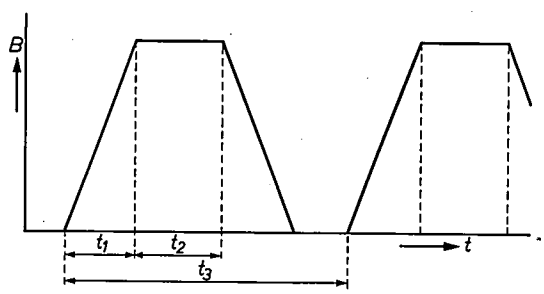


Fig. 9. Typical programme for the magnetic field; t_1 rise time during which the acceleration takes place; t_2 flat top for experimental use; t_3 repetition time.

magnetic field with time is represented by a sine-wave superimposed on a constant bias.) The huge energy in the "top" magnetic field is stored in flywheels (or in the PPA in capacitors) during the low-field period. The particles must now be accelerated so that their velocity matches the magnetic field given by the programme. This requires, first, that the frequency of the electric field should be an exact harmonic of the orbital frequency f of the particle,

$$f = v/2\pi R, \quad \ldots\ldots \quad (6)$$

where $2\pi R$ is the length of the particle orbit; v is dictated by the magnetic field via eq. (1) and (5). Secondly, for the particles to obtain the proper acceleration, they must find a specific electric field value in the accelerating station when they go through it. This value is not obtained by precisely controlling the amplitude of the r.f. electric field, but, as will be seen later, by an adjustment of the relative phase of the r.f. field and the revolution of the particles.

As in all the accelerators under consideration the velocity of the particle is well into the relativistic range after acceleration, i.e. the final velocity is approximately equal to c, the frequency ratio f_f/f_i to be covered by the accelerating field is approximately c/v_i (where f_f is the frequency after acceleration, and f_i and v_i are the frequency and the velocity at injection).

At this point it should be noted that in practice there is a lower limit to the flux density of the magnetic field

at injection, B_i: in the first place it cannot normally be lower than the remanence flux density of the magnet, and secondly, to be well defined, it should be large compared with the stray flux densities around the ring. For acceptable values of B_i (about 10 milliteslas = 100 gauss at least) and for the larger machines, eq. (5) requires the particle to be injected near the relativistic range (for $B = 10$ milliteslas, $\varrho = 70$ m we obtain $\beta = v/c \approx 0.2$). Other considerations may require the injection field and energy to be still higher, as we shall see later. A convenient corollary of these requirements for the injection energy in the larger machines is that the frequency range to be covered is small, as v_i is not very far from c. In the smaller machines, on the other hand, with injection in the pre-relativistic range, the required frequency swing is quite large.

From (1), (5) and (6) it is readily seen that the variation of the frequency with the magnetic field is given by:

$$f/f_c = \frac{B}{\sqrt{B^2 + B_\varrho{}^2}}, \quad \ldots\ldots \quad (7)$$

where $f_c = c/2\pi R$ and $B_\varrho = mc/e\varrho$. The quantity f_c is the asymptotic value of the frequency when the velocity approaches c ($f_f \approx f_c$).

The frequency of the accelerating field f_a is chosen to be some integral multiple of the orbital frequency f:

$$f_a = hf. \quad \ldots\ldots \quad (8)$$

The choice of the "harmonic number" h will be discussed later.

The actual devices which accelerate the protons are resonant cavities (in the PPA combined with a drift tube) in which the accelerating voltage appears across a gap in the metal of the vacuum-chamber wall. These cavities have been treated in some detail in the preceding article (see figs. 5 and 7 in that article), so that we can leave the subject here.

Usually a number of cavities (M) are distributed evenly around the ring. For a given energy gain of the particle per turn, the required voltage per cavity V is proportional to M^{-1}. For a cavity of a given design the power dissipated is proportional to V^2 and thus to M^{-2}. The total power required is therefore proportional to M^{-1}, and the total power can be reduced by using more cavities.

Accelerator focusing characteristics and particle oscillations

If (7) were ideally satisfied, then the relations (5) and (6) would define an ideal particle moving in a nominal orbit with nominal momentum. However, the principles described above would be of no use at all in practice if it were not for the fact that particles deviating slightly from the ideal behaviour can also be accelerated. This is due to the "focusing" characteristics of the accelerator.

Two kinds of deviation from the ideal behaviour and two kinds of focusing can be distinguished. First, particles may have lateral deviations from the nominal orbit, and it is by "transverse" focusing that they are kept in the track. Such particles will perform "betatron oscillations" around an equilibrium orbit in the ring. Secondly, particles may have a longitudinal deviation (a deviation in azimuth) from the ideal particle passing the gap at the correct phase of the accelerating voltage. Such particles are kept in the acceleration cycle by "longitudinal" or "phase" focusing, and they will perform "synchrotron oscillations" around the ideal particle. A discussion of these focusing characteristics is simplified by the fact that the increase in particle energy per turn is small compared with the particle energy itself. Therefore, transverse focusing and betatron oscillations can be studied separately in the "adiabatic approximation" in which the acceleration is neglected. We shall first confine our attention to this case.

Transverse focusing and betatron oscillations

In relation to transverse focusing two kinds of machine can be distinguished. In the first, the *combined-function machines*, transverse focusing is obtained by the same magnetic field that establishes the curvature of the particle orbit around the ring. The focusing is a result of a carefully designed transverse gradient in the bending field. In a *separated-function machine* focusing and bending are separated, the bending being established by flat-pole magnets and the focusing by (for example) quadrupole magnets inserted between the bending magnets or (as in the ZGS) by specially shaped edges on the bending magnets. Of the five machines under discussion only the ZGS is of the separated-function type; the others are combined-function machines. Apart from the main bending and focusing elements a number of correcting magnetic elements, producing multipole (dipole, quadrupole, sextupole or octupole) magnetic fields, are generally installed in the straight sections for a fine adjustment of the bending and focusing characteristics of the ring.

Focusing by a transverse field gradient

As indicated, focusing can be obtained by introducing a transverse field gradient. This is achieved by using flared pole pieces. Consider for example a magnet of cross-section like that of *fig. 10a*, with the pole pieces flared outwards. The field lines are curved,

[2] This saturation limitation may be removed in future accelerators by using magnet coils at very low temperatures.

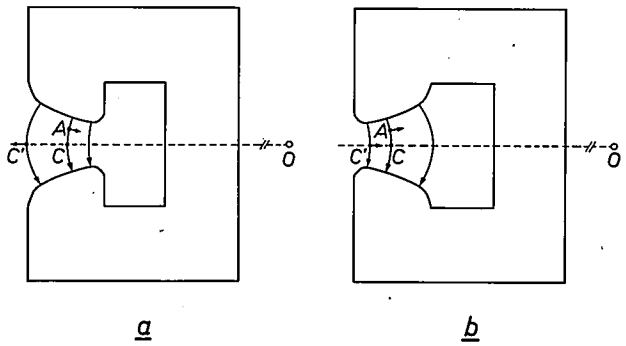


Fig. 10. Profiles of focusing magnet sectors (not to scale). The centre O of the synchrotron is on the right. C is the nominal orbit. *a*) A D sector, in which a particle at A, with vertical deviation, is pushed back towards C, but a particle at C', with horizontal deviation, will move away from C (if n is large). This sector is vertically focusing but horizontally defocusing. *b*) An F sector, in which the situation is reversed.

bulging outwards, and the field decreases as r, the distance from the centre of the accelerator, increases. The *field index n* is defined by:

$$\frac{\partial B}{\partial r} = -n\frac{B_0}{\varrho_0}, \quad \ldots\ldots (9)$$

where B_0 and ϱ_0 are the flux density and the radius of curvature at the nominal equilibrium orbit C. In fig. 10*a* n is positive. As the Lorentz force is always perpendicular to the field, a particle in A, moving with nominal velocity, experiences a force with a component pushing it back to the median plane: the magnet gives *vertical focusing*. However, if n is large, the field at C' will be much smaller than at C, the curvature of the orbit of a particle moving with nominal velocity will be too slight and it will move away from C: the magnet is *horizontally defocusing*. The sector is called a D sector. The sector shown in fig. 10*b* is called an F sector; the field index is negative; this magnet focuses horizontally and defocuses vertically.

When the flux density is resolved into a vertical component B_z and a horizontal component B_r, then:

$$B_z = B_0 + g\Delta r, \quad \ldots\ldots (10a)$$

$$B_r = gz, \quad \ldots\ldots (10b)$$

where Δr is the horizontal and z the vertical distance from the nominal orbit, and g is proportional to n and to the particle momentum p. Comparing (10a) with (9) and using (5), we have $g = -np/e\varrho_0^2$. The coefficient in (10b) must also be g to satisfy the equation curl $B = 0$. The field is a superposition of a homogeneous guiding field B_0 and a quadrupolar focusing field with components $g\Delta r$ and gz. In a separated-function machine the only field *inside* the (flat-pole) magnets is a homogeneous guiding field $B_z = B_0$, and *between* the magnets the field only has quadrupolar components $B_z = g\Delta r$, $B_r = gz$, which are produced by quadrupole lenses, or by shaping the magnet edges.

The rule that a focusing effect in the horizontal direction is accompanied by a defocusing effect in the vertical direction, and *vice versa*, is a direct consequence of the relation curl $B = 0$.

Weak or constant-gradient focusing

From the above one might be inclined to conclude that focusing in both directions, horizontally *and* vertically, at the same time is impossible. A closer look, however, shows that a D sector ($n > 0$) focuses vertically *and* horizontally if n is restricted to sufficiently small values. Consider a particle at C' (*fig. 11*) with nominal velocity. We should now distinguish between the distance r from the centre of the nominal orbit

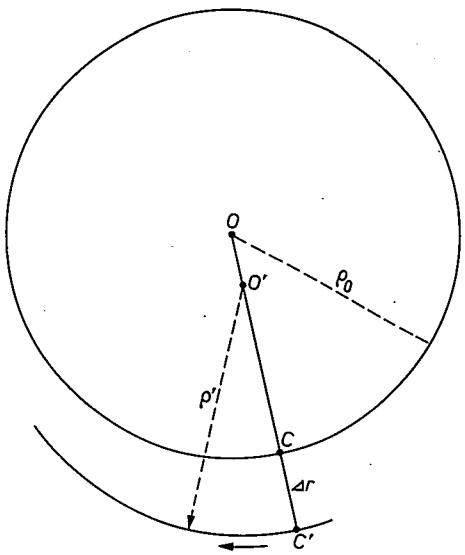


Fig. 11. Horizontal weak focusing. A particle at C' of nominal velocity, but which is displaced by a distance Δr from the nominal orbit C, has an orbit with an instantaneous radius of curvature ϱ', determined by the field at C'. Horizontal focusing is obtained if OO' is of the same sign as $CC' = \Delta r$. This leads to the condition $n < 1$ (see text).

($r = OC' = \varrho_0 + \Delta r$) and the instantaneous radius of curvature $\varrho' = O'C'$, determined by the local flux density $B_0 + \Delta B$ at C'. The deviation ΔB is given by (9):

$$\Delta B = (\partial B/\partial r)\,\Delta r = -n(B_0/\varrho_0)\Delta r.$$

The instantaneous radius of curvature is:

$$\varrho' = \varrho_0 + \Delta\varrho$$

where, as (5) shows,

$$\Delta\varrho/\varrho_0 = -\Delta B/B_0.$$

Combining the equations we have:

$$\varrho' = \varrho_0 + n\Delta r.$$

The particle at C' will move back towards the equilibrium orbit C if O' is on the same side of O as C' is of C, in other words, if OO' and $CC' = \Delta r$ have the same sign. Now $OO' = OC + CC' - O'C' = \varrho_0 + \Delta r - (\varrho_0 + n\Delta r) = \Delta r(1 - n)$. Therefore, a D sector is horizontally focusing — although only weakly — if $n < 1$.

The type of accelerator that uses this principle is the

constant-gradient (CG) or *weak-focusing* accelerator. In this type of machine magnets of the D type only are used, with the same value of n all around the ring. The stability condition in a CG synchrotron is then:

$$0 < n < 1. \quad \ldots\ldots \quad (11)$$

Betatron oscillations

Owing to the magnetic restoring forces which draw the particles back towards the equilibrium orbit, the particles not in this particular orbit will oscillate about it, rather like a harmonic oscillator. These oscillations are called "betatron oscillations"; they were first discovered in the betatron [3].

In a combined-function machine the bending and the focusing fields, produced by the same magnets, increase at the same rate during acceleration and remain proportional to the particle momentum p. Therefore, the local radius of curvature, determining the orbit geometry, is constant everywhere in the magnet gap. As a result, the parameter ν, defined as *the number of betatron periods per turn* (the number of betatron wavelengths around the ring) is constant during acceleration [4]. The betatron oscillation frequency is νf, f being the orbital frequency (eq. 6). In particular, for a CG machine without straight field-free sections it can be shown that:

$$\left.\begin{aligned} \nu_H &= \sqrt{1-n}, \\ \nu_V &= \sqrt{n}, \end{aligned}\right\} \quad \ldots\ldots \quad (12)$$

where ν_H and ν_V are the number of betatron periods per turn in the horizontal plane and the vertical plane respectively.

The relations (12) are consistent with (11): an imaginary frequency implies exponential instead of oscillatory behaviour, i.e. instability. Under the stability condition (11) ν_H and ν_V are real. We note that both are smaller than 1, a fact that is typical for a weak-focusing machine.

The ν's can be adjusted by adding quadrupole lenses in straight sections (see p. 337). If this is done, ν does not automatically remain constant during acceleration (unless these additional lenses are electrically in series with the bending magnets and have the same saturation curves). This is particularly so for a separated-function machine.

As the particle is accelerated, the betatron frequencies $\nu_H f$ and $\nu_V f$ increase, in proportion to the orbital frequency f. This is accompanied by a decrease in the betatron amplitudes, so that *beam shrinkage* occurs: a wide beam becomes a narrow beam during the acceleration. This damping process corresponds to the decrease in amplitude of a conventional harmonic oscillator in which the mass and the restoring force are steadily increased. The betatron amplitudes in each plane of motion are proportional to $B^{-1/2}$ ($\propto p^{-1/2} \propto (\beta\gamma)^{-1/2}$), which can be shown by applying Liouville's theorem.

Resonance

In practice not all values of n which obey condition (11) are satisfactory; certain values — corresponding for example to ν_H or $\nu_V = \frac{1}{2}$ — have to be excluded to prevent resonance of the betatron oscillation with the revolution of the particle. If such a resonance occurs, any slight irregularity in the machine may cause a disturbance which will build up each time the particle passes through; the particle will then rapidly be lost. If resonance does not occur, such disturbances will average out. This consideration applies to all types of synchrotrons (including those in which the number of betatron periods ν per turn is larger than 1, see below) and in principle, for any accelerator, one should avoid the values of ν given by:

$$\left.\begin{aligned} l\nu_H + m\nu_V &= k, \\ \text{where} \quad k,l,m &= 0, \pm 1, \pm 2, \pm 3, \ldots \end{aligned}\right\} \quad \ldots \quad (13)$$

It is not of course possible to avoid the relation (13) for every set of integers k,l,m. In practice the "lower" resonances are the only ones that cause trouble.

Momentum compaction

Let us at this point introduce a concept which will be useful in the study of the injection procedure: *the momentum-compaction factor* α. For the nominal equilibrium orbit C the nominal values B_0, ϱ_0 and p_0 for the flux density, the radius and the particle momentum are related by (5). A particle with a momentum $p = p_0 + \Delta p$ can be accommodated in another equilibrium orbit, of radius $\varrho = \varrho_0 + \Delta\varrho$, if relation (5) between p, ϱ and B at this new orbit is satisfied. The cross-section of the magnet defines a relation between B and ϱ, or between ΔB and $\Delta\varrho$ (given by eq. 9). Therefore, for equilibrium orbits, Δp and $\Delta\varrho$ are directly related. For a given machine, Δp and ΔR are then also directly related, and the momentum-compaction factor α is defined by:

$$\alpha = \frac{\Delta R/R_0}{\Delta p/p_0}. \quad \ldots\ldots \quad (14)$$

A small value of α ("strong momentum compaction") is favourable since particles of widely varying momentum can then be accommodated in a narrow range of radii (in a narrow vacuum chamber).

From eq. (5) we have $\Delta p/p_0 = \Delta B/B_0 + \Delta\varrho/\varrho_0$,

[3] Betatrons have been discussed earlier in this journal: A. Bierman and H. A. Oele, Philips tech. Rev. **11**, 65, 1949/50.

[4] This number is also often denoted in the literature by "Q".

and from eq. (9), if n is constant along the ring, $\Delta B/B_0 = -n\Delta\varrho/\varrho_0$, so that for a CG accelerator $\Delta p/p_0 = (1-n)\Delta\varrho/\varrho_0$, and, if there are no straight sections ($\varrho_0 = R_0$):

$$\alpha = 1/(1-n) = 1/\nu_H^2. \quad . . \quad (15)$$

It may be shown that for any synchrotron with straight sections that cover no more than 20% of the circumference:

$$\alpha \approx 1/\nu_H^2. \quad . . . . . \quad (16)$$

Thus, in general, if the restoring forces in the horizontal plane are large, implying *fast* betatron oscillations (ν_H large), the momentum compaction is strong (α small).

Strong or alternating-gradient focusing

The alternating-gradient (AG) or strong focusing accelerators, instead of having only D sectors, have D *and* F sectors alternating around the ring. Generally these are of equal length, the field index n is alternatively positive and negative, its value is the same in the D and F sectors and it is very large (roughly 300).

In the CPS at Geneva the D and F sectors are combined into DF or FD units (a few of which can be seen in fig. 4). In the AGS at Brookhaven the D and F sectors are separate (for the AGS magnets see fig. 2 of the preceding article). Straight field-free sections (O) may be present between units. The whole ring then consists of a large number of identical "periods", each period consisting of an arrangement such as FDODFO (CPS) or FODODOFO (AGS).

It has been shown that an arrangement of this kind can have focusing characteristics, and that these can be much more effective than in CG machines [5]. If the orbit of a proton moving in a sequence of D and F sections (of equal length and equal, large, values of n) is calculated, it is of course found that in the horizontal plane the orbit is locally sinusoidal (converging) in an F sector, and hyperbolical (diverging) in a D sector; in the vertical plane the situation is reversed. The calculation shows, however, that *in both planes, up to large values of n, the net effect, after a series of alternate F and D sectors, is that of a focusing force, and that the orbit is stable.* If n exceeds a certain critical value (depending on the length of the sectors), the orbit becomes unstable [6]. A stable orbit is shown in *fig. 12a* and an unstable one in fig. 12*b*. If straight sections are introduced similar results are obtained, and a stable orbit for such a case is shown in fig. 12*c*.

The focusing characteristics of a system of D and F sectors have an analogy in optics. In the first place we note that a convergent and a divergent lens of equal strength $\pm 1/F$ separated by a distance L form a convergent optical system. More relevant to the present discussion, however, is the fact that a beam of light that has passed through a long sequence of well aligned lenses, alternately of strength $+1/F$ and $-1/F$, will remain restricted in width if $L < 2F$ (L = distance between lenses).

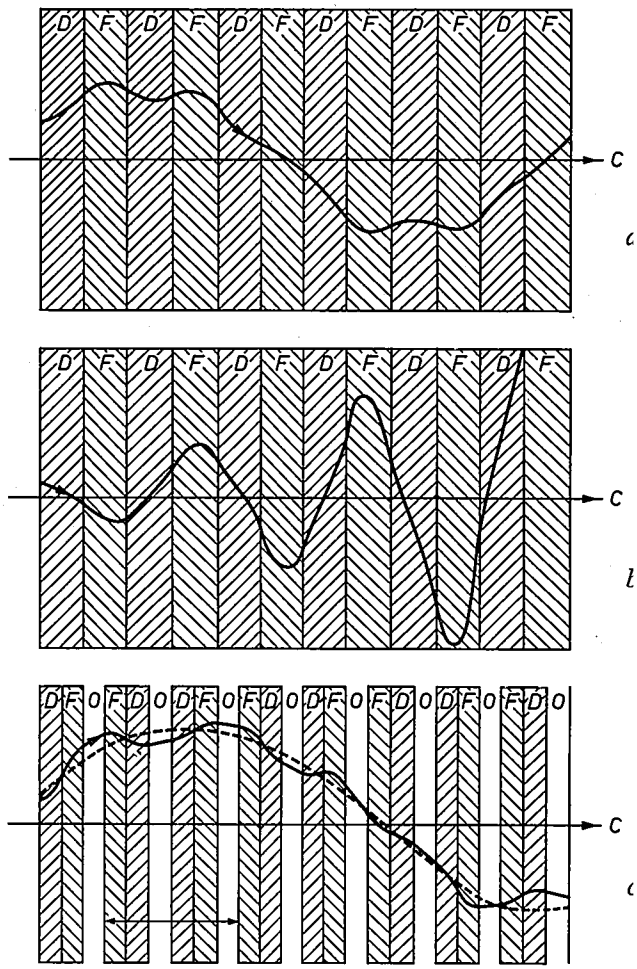


Fig. 12. Particle trajectories in a sequence of D and F magnets of equal length and equal (large) values of n. In (*a*) n is below the critical value, and the orbit is stable. In (*b*) n is too large and the orbit is unstable. *c*) A stable orbit in a sequence of magnets containing straight sections. Superimposed on a slow betatron oscillation there is a "wriggle" corresponding to the passage through the individual D and F magnets. The transverse dimensions are greatly exaggerated in comparison with the longitudinal dimension. The double arrow in (*c*) indicates one period of the magnetic structure.

In a stable orbit (fig. 12*c*) each proton performs large horizontal and vertical oscillatory motions around the nominal orbit, the betatron oscillations, on which is superimposed a small "wriggle" motion, corresponding to the passage through the individual D and F sectors.

[5] The AG focusing method has been proposed independently by L. H. Thomas (for cyclotrons) in 1938 (Phys. Rev. **54**, 580 and 588, 1938), by N. Christophilos in 1950 (unpublished; see however E. D. Courant, M. S. Livingston, H. S. Snyder and J. P. Blewett, Phys. Rev. **91**, 202, 1953) and by E. D. Courant, M. S. Livingston and H. S. Snyder in 1952 (Phys. Rev. **88**, 1190, 1952).

[6] Theoretically it is not necessary for the value of n in the D sectors (n_1) to be equal to that in the F sectors (n_2). In general stable orbits are obtained in well defined regions of the (n_1,n_2)-plane.

The numbers of betatron periods per turn, ν_H in the horizontal plane and ν_V in the vertical plane, are now complicated functions of n.

The values of n are further restricted by the condition that resonances should not occur. This leads to a great number of forbidden values, given by eq. (13).

Owing to the very high values of n, the average restoring force — though very much smaller than the restoring force in the horizontal plane of an individual F sector or in the vertical plane of a D sector — is much larger than in a CG machine. As a consequence the numbers of betatron periods per turn, ν_H and ν_V, are much larger than in the CG case, and turn out to be larger than 1. In the CPS, for example, $n = 288$, $\nu_H = \nu_V = 6.25$. For a given angular deviation of the particles the betatron oscillation amplitude is inversely proportional to ν, and any increase in ν will therefore allow the vacuum chamber and hence the magnet to be made smaller.

The large average restoring force is also expressed in a strong momentum compaction (see eq. 16). For a given range of momenta to be accommodated, this again allows the vacuum chamber to be smaller, although it creates difficulties at injection, as we shall see later.

Zero-gradient edge focusing

In the Zero Gradient Synchrotron (ZGS) in Argonne focusing is obtained in a completely different way. The pole pieces of the bending magnets are flat and parallel, giving a homogeneous field (the gradient is zero, $n = 0$). Focusing is obtained here by using magnets with edges that are *not* perpendicular to the nominal trajectory (see *fig. 13*). It can easily be shown that an arrangement like the one shown in fig. 13 (with $\psi < \vartheta/2$) will focus in the horizontal plane. The path *in the magnet* of a particle with the nominal velocity in a trajectory C', outside the nominal trajectory C, is longer than that along C. It therefore experiences more bending and is thus forced back towards the nominal orbit.

The arrangement will also focus in the vertical plane: this can be seen as follows. *Fig. 14a* shows a cross-sectional vertical plane perpendicular to the edge, as seen from inside the ring. The fringe field is parallel to this plane. It has a horizontal component directed towards the edge above the median plane (see point P) and away from the edge below this plane (see Q). Now resolving the horizontal component (fig. 14*b*, the edge viewed from above) we see that at P the component perpendicular to the particle velocity is directed towards the centre. This leads to a downward component of the Lorentz force. At Q the situation is reversed and there is an upward Lorentz component. More generally, the vertical focusing can be related to that of the D magnet of fig. 10*a*, since, owing to the oblique edges, the total field averaged over an orbit decreases with increasing radius of the orbit. This corresponds to a negative horizontal gradient, as in fig. 10*a*.

As at all edges the forces on the protons are of the restoring kind (not being alternated with repelling forces), this machine must be considered as a *weak-focusing machine*, and it can be shown that the ν's are

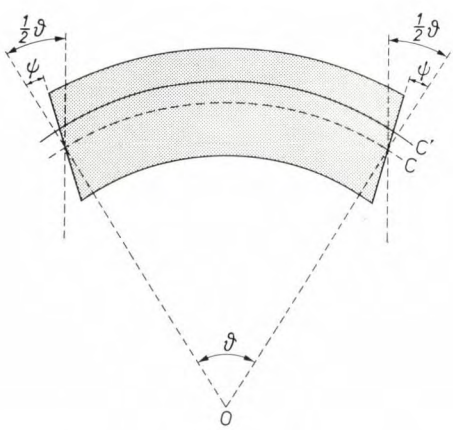


Fig. 13. Horizontal edge focusing. ϑ magnet bending angle. ψ angle between edge and plane normal to the nominal trajectory C. As $\psi < \vartheta/2$, the trajectory C' in the magnet for a particle at the nominal velocity is longer than C, and the particle is deflected further.

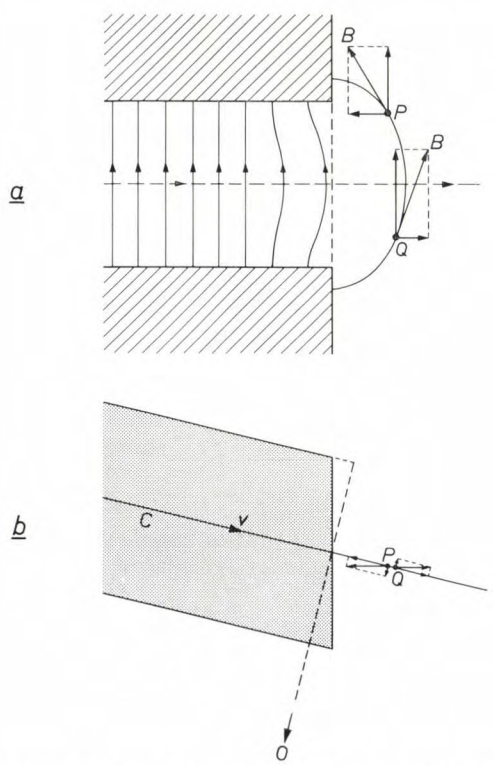


Fig. 14. Vertical edge focusing. *a*) Cross-sectional vertical plane perpendicular to the edge, viewed from inside the ring. *b*) Horizontal plane viewed from above. Resolving the fringe field into components shows that the Lorentz force has a downward component at P and an upward component at Q (see text).

again less than unity. As the bending and the focusing take place in different parts of the magnet field, this machine can in certain respects be considered as a separated-function machine. Bending and focusing characteristics of the ZGS ring can be corrected by means of four special magnets inserted between the main magnet octants (see figs. 6 and 31).

Representation of the particle motion in transverse phase space; acceptance and emittance

The betatron oscillation of the particle around the nominal position is conveniently represented in *transverse phase space* (*fig. 15*). This is defined as the plane in which the abscissa is the deviation y from the nominal orbit C, and the ordinate is $y' = \mathrm{d}y/\mathrm{d}s$, where s is the curvilinear coordinate along C; y' is the angle between the real trajectory and C. Two such planes are used, one for the horizontal and one for the vertical motion.

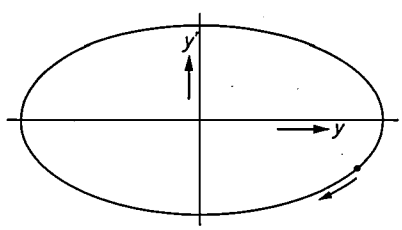


Fig. 15. Transverse phase space. The abscissa y is either the horizontal or the vertical deviation from the nominal orbit C, and the ordinate is $y' = \mathrm{d}y/\mathrm{d}s$, where s is the curvilinear coordinate along C. In the sinusoidal approximation, if the acceleration is neglected, a particle describes an ellipse in this plane; one turn is a betatron period. The size of the ellipse may vary from particle to particle but all the ellipses are geometrically similar (and remain so during acceleration).

In the sinusoidal approximation (the sinusoid of fig. 12*c*, neglecting the "wriggle" motion), the betatron motion of a particle in each plane is described by:

$$y = a\,\bar{\beta}^{1/2} \cos(\nu\vartheta + \delta), \quad \ldots \quad (17)$$

where ϑ is the azimuth angle around the ring ($\vartheta = s/R$) and a and δ are constants. $\bar{\beta}$ is a mean value of the "amplitude function" $\beta(\vartheta)$ [7]. The function $\beta(\vartheta)$ is, like the constant ν, an important characteristic of the synchrotron. It is an oscillating function of ϑ, varying from $\beta_{\max}$ to $\beta_{\min}$ and back in one period of the magnet structure, and it accounts for the wriggles of the particle trajectory (see fig. 12*c*). A "form factor" F is defined by:

$$F = \beta_{\max}/\bar{\beta}. \quad \ldots\ldots \quad (18)$$

F usually has a value between 1 and 1.4. The mean value $\bar{\beta}$ of the amplitude function is given by:

$$\bar{\beta} = R_0/\nu. \quad \ldots\ldots \quad (19)$$

It can be shown that for the sinusoidal approximation:

$$y' = -a\,\bar{\beta}^{-1/2} \sin(\nu\vartheta + \delta). \quad \ldots \quad (20)$$

Thus in transverse phase space the trajectory of the particle is an ellipse of axial ratio $\bar{\beta}$ on which the wriggle motion is superimposed. One turn around the ellipse is a betatron oscillation and there are ν such turns for one revolution around the ring.

The area of that region of phase space that can be used for injection is called the *admittance* or *acceptance A* of an accelerator: particles injected at phase-space coordinates within this area can be successfully accelerated to the end of the acceleration period, and the others are lost to the vacuum-chamber walls. The acceptance is a measure of how many particles can be accepted by the machine from the injector. The beam *emittance* is an area in phase space which contains all the states of the particles which form the beam. No particle is lost at injection if good matching brings the injected beam emittance completely inside the machine acceptance.

The area of the ellipse of fig. 15 is πa^2 (see eq. 17 and 20). If q_H is the semi-aperture of the chamber in the horizontal plane, then, in the sinusoidal approximation, $q_H = a\bar{\beta}_H^{1/2}$ for the largest possible value of a. The acceptance in the horizontal plane, A_H, would therefore be $\pi q_H^2/\bar{\beta}_H$. Taking the wriggle motion into account, we obtain:

$$A_H = \pi q_H^2/\beta_{H\,\max} = \pi q_H^2 \nu_H/FR_0. \quad \ldots \quad (21)$$

The acceptance is proportional to ν. Thus AG machines can be expected to have larger acceptances than CG machines of the same radius and vacuum-chamber cross-section, because of the larger values of ν. An expression similar to eq. (21) can be derived for A_V, the acceptance in the vertical plane. The "total acceptance" is the product $A_H A_V$. Although we usually have $\nu_H \approx \nu_V$ and $\beta_{H\,\max} \approx \beta_{V\,\max}$, this is not necessarily the case.

In the ideal case the beam emittance at injection is an ellipse just filling the machine acceptance. As noted earlier, the beam shrinks when accelerated. The betatron amplitudes decrease as $B^{-1/2}$, and the emittance therefore decreases as B^{-1} or (see eq. 5 and 1) as $(\beta\gamma)^{-1}$. This means that the product of $\beta\gamma$ and the emittance is independent of the energy of the particles in the beam. This quantity is called the *invariant emittance*. Correspondingly the *invariant acceptance* can be defined, which is useful in comparing structures, as will be seen later.

Eq. (21) gives the acceptance in the horizontal plane for the ideal case of a monoenergetic beam in a perfectly rectangular chamber. In practice the acceptance is smaller because of energy spread in the injected

beam, synchrotron oscillations, magnet misalignment, chamber shape, etc. In particular, a deviation Δp from the nominal momentum p_0 (related to the energy spread of the injected beam and the synchrotron oscillations) leads to a deviation ΔR from the radius R_0 of the equilibrium orbit (eq. 14), and thus to a reduction of the maximum betatron amplitude to $q_H - \Delta R$. Also, their betatron frequency becomes slightly different: it is $\nu + \Delta\nu$ where $\Delta\nu/\nu \approx -\Delta p/p$.

Phase focusing, synchrotron oscillations and transition energy

Now that we have had a look at the particle motion in the guiding and focusing magnetic fields where the acceleration need not be taken into account, we turn to the phenomena directly associated with the acceleration.

Let us assume for the moment that the frequency of the accelerating voltage is ideally matched to the magnetic field (so that it satisfies equations 7 and 8). Then an ideal, "synchronous", particle can be imagined, i.e. a particle in the nominal orbit, with the nominal velocity, that always crosses an accelerating gap at the same phase of the accelerating voltage (P_1, P_2, ... or P_1', P_2', ... in *fig. 16*). If M is the number of cavities, or the total number of gaps for multigap cavities, V_m the amplitude of the voltage, and φ the phase when a particle passes the gap ($\varphi = 0$ at the peak voltage), the increase in the particle energy per turn is $eMV_m \cos\varphi$. This must be equal to the energy increase required to keep the particle in the nominal orbit in the increasing magnetic field. This increase is $\dot{E}/f$ where E is related to B by (4) and (5). With the aid of (1), (2) and (6) we obtain:

$$MV_m \cos\varphi_0 = 2\pi R\varrho\dot{B}, \quad \ldots \quad (22)$$

a relation giving the "synchronous phase" φ_0 when $\dot{B}$ and V_m are known.

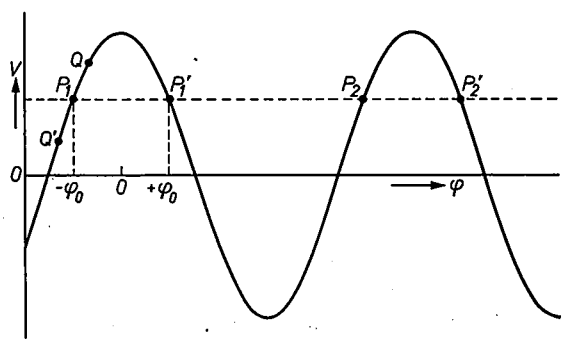


Fig. 16. The accelerating voltage V in a gap as a function of phase φ (or time t, since $\varphi = 2\pi f_a t + \text{const.}$). P_1, P_2, ... and P_1', P_2', ... are "synchronous points". If a particle is at the nominal radius and energy when at Q or Q', it will oscillate about P_1 if $\Delta f/\Delta E$ is positive, i.e. if the nominal energy E_0 is smaller than the transition energy E_t.

Now how does a particle deviating slightly in phase or energy from the synchronous particle behave? If a particle deviates in energy, the corresponding equilibrium-orbit radius and orbital frequency will also deviate from the synchronous values. For reference we note, by differentiating (1) and (2), and using (14), that ΔE, Δv, Δp and ΔR are related by:

$$\Delta E/E_0 = \beta^2\gamma^2\Delta v/v_0 = \beta^2\Delta p/p_0 = (\beta^2/\alpha)\Delta R/R_0. \quad (23)$$

Differentiating (6), we obtain:

$$\Delta f/f_0 = \Delta v/v_0 - \Delta R/R_0, \quad \ldots \quad (24a)$$

or, combining the result with (23):

$$\Delta f/f_0 = (E_t^2/E_0^2 - 1)\,(\alpha/\beta^2)\Delta E/E_0, \quad (24b)$$

where

$$E_t = mc^2/\sqrt{\alpha} = \nu_H mc^2. \quad \ldots \quad (25)$$

The kinetic energy corresponding to E_t is:

$$T_t = (\nu_H - 1)mc^2. \quad \ldots \quad (26)$$

From (26) it can be seen that for a CG accelerator $T_t < 0$ (since $\nu_H < 1$), so that E_t and T_t are of no physical significance. In an AG accelerator, however, E_t and T_t are critical quantities actually reached during acceleration; E_t is called the total and T_t the kinetic *transition energy*. As we shall now see, E_t and T_t are very important constants for AG accelerators.

In the first place it follows from eq. (24b) that $\Delta f/\Delta E$ changes its sign when E_0 becomes greater than E_t. A particle with more energy than the synchronous particle ($\Delta E > 0$) will describe its orbit faster than the synchronous particle ($\Delta f > 0$) below transition ($E_0 < E_t$), but more slowly ($\Delta f < 0$) above transition ($E_0 > E_t$). We can look into this more closely by considering eq. (24a) and recalling that an excess of energy always implies a larger equilibrium orbit ($\Delta R/\Delta E > 0$, eq. 23). If the energies are sufficiently small, the excess of momentum Δp implies an excess velocity Δv that will overcompensate the longer orbit the particle has to travel, resulting in a higher orbital frequency; at large energies, however, when the particle velocity is very close to the velocity of light, the excess of momentum no longer implies an excess of velocity: in eq. (1) it is the increase of γ, and not of β, that becomes the dominant effect, and the larger orbit will result in a lower orbital frequency.

Secondly, let us consider what happens if a particle deviates in *phase* from the synchronous particle. Let us first examine the case $E_0 < E_t$, $\Delta f/\Delta E > 0$. Consider

[7] The now conventional symbols $\bar{\beta}$, $\beta(\vartheta)$, β_{max} and β_{min} will be used in this article. Confusion with the quantity $\beta = v/c$ is not very likely as β is used only in the ways indicated above when referring to the amplitude function.

a particle, corresponding to Q near P_1 in fig. 16, in the nominal orbit with nominal energy, but arriving at the gap late. Beyond the gap ΔE will be positive (as $V_Q > V_P$) and thus Δf will be positive: the particle will be not quite so late at the next gap. After a number of gaps have been passed it will arrive on time. Meanwhile it has gathered an excess of energy, and Δf is still positive. It therefore will arrive earlier and earlier, but, since the voltage is from now on smaller than V_P, our particle acquires less energy than the nominal particle; its excess of energy is used up, until, at Q' say, its energy, radius and frequency are nominal again. As the particle is now early, the process will repeat itself in the reverse direction. The particles are thus seen to perform oscillations (*synchrotron oscillations*) in phase, energy and radius about the synchronous nominal values. The deviations in energy and radius, ΔE and ΔR, are 90° out of phase with $\Delta\varphi$ and therefore in phase with $\mathrm{d}\Delta\varphi/\mathrm{d}t$. It follows that, if $E_0 < E_t$, a range of stable phases has its centre on the *rising* slope. Particles inside this range are said to be *phase-focused*; all particles outside this range are ultimately lost to the vacuum-chamber walls. If, on the contrary, $E_0 > E_t$, the centre of the stable range is on the *descending* slope.

In a *weak-focusing accelerator*, $\alpha > 1$, $E_t < mc^2$, so that $E_0 > E_t$ for all E_0. Thus, in the whole range of energies during acceleration, particles can be stably accelerated in a phase range on the descending slope.

In a *strong-focusing accelerator*, $\alpha < 1$ and $E_t > mc^2$. We now must distinguish between the energy region before transition and the region after transition. Stable acceleration is only possible in a phase range on the rising slope before transition and on the descending slope after transition. Therefore, when the transition energy is reached, the *accelerating voltage has to be forced into a phase jump of* $2\varphi_0$ (*fig. 17*). The accelerated particles, before transition on the rising slope, will find themselves on the descending slope after this phase jump, and can be further accelerated to energies above E_t.

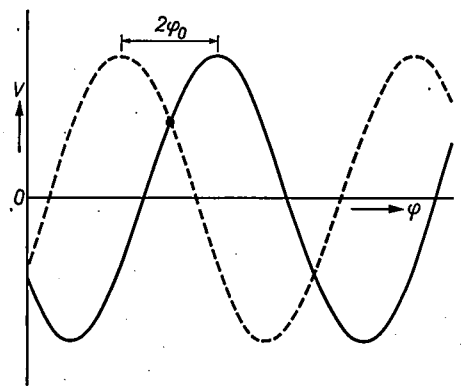


Fig. 17. In a strong-focusing accelerator the accelerating voltage is forced into a phase jump of $2\varphi_0$ at the transition energy. The solid curve applies before transition, the dashed curve after transition.

The first AG synchrotron builders were very anxious to know whether the transition-energy barrier really could be broken. They found that it could, on the 24th of November 1959, when the protons in the CERN machine went through transition for the first time, much to the relief of workers on both sides of the Atlantic.

In general, the transition phenomenon requiring a phase jump of the voltage during acceleration occurs only if $mc^2 < E_t < E_t$, i.e. if

$$1 < \nu < \gamma_t, \qquad \ldots\ldots \quad (27)$$

where γ_t is the γ value of the particles at the end of the acceleration.

In the CPS the transition energy E_t is $6.25 \times 0.938 \approx 5.8$ GeV ($T_t \approx 4.9$ GeV). The rate of increase of the field flux density, $\dot{B}$, is usually constant, and V_m is generally given such a value as to make φ_0 approximately equal to $\pm\pi/3$ (see eq. 22).

If particles are to perform synchrotron oscillations, extra horizontal space will be required because of the variations in equilibrium-orbit radius. This space is additional to that required by betatron oscillations. For the CPS, for instance, a theoretical estimate shows that, somewhere in the middle of the cycle, away from transition, $\Delta E/E \approx 10^{-3}$ and $2\Delta R \approx 6$ mm.

Frequency and variation in amplitude of the synchrotron oscillations

In a "linear" approximation, i.e. for small amplitudes, the frequency f_s of the synchrotron oscillations can be shown to be:

$$f_s = f\sqrt{\frac{h}{2\pi}\tan\varphi_0\left(\frac{1}{\gamma^2}-\frac{1}{\nu^2}\right)\dot{B}/Bf} \qquad \ldots \quad (28)$$

(for f and h, see eq. 7 and 8). The need for a phase jump in the accelerating voltage at the transition energy can be seen from eq. (28): when $1/\gamma^2 - 1/\nu^2$ changes sign, $\tan\varphi_0$ has to change sign too, to keep the frequency real, i.e. the acceleration stable.

In eq. (28), $\dot{B}/Bf$ is the relative increase of the flux density per turn of the particle, which is $\ll 1$. In order of magnitude (somewhere in the middle of the cycle): $\dot{B}/B \approx 1\ \mathrm{s}^{-1}$, $f \approx c/2\pi R \approx 3\times10^8/2\pi\times100 \approx 10^6\ \mathrm{s}^{-1}$, so that $\dot{B}/Bf \approx 10^{-6}$. As $h/2\pi$ and $|\tan\varphi_0|$ are of the order of 1, and away from transition $|1/\gamma^2 - 1/\nu^2|$ is about 1 (before transition) or $1/\nu^2$ (after transition) we have $f_s/f \approx 10^{-3}$ or $10^{-3}/\nu$. Thus the frequency of the synchrotron oscillations is much smaller than the orbital frequency f (and, both for AG and CG accelerators, also much smaller than the betatron frequency νf). It is seen that f_s lies in the kHz range (away from injection or transition).

As with the betatron oscillations, the frequency of the synchrotron oscillations varies during acceleration, and the amplitudes of $\Delta\varphi$, ΔR, ΔE and Δf will also vary. It is now extremely important whether the amplitudes, in particular those of $\Delta\varphi$ and ΔR, will increase or decrease. In the linear approximation it can be shown that, if $\dot{B}$ is taken to be constant and the required phase jump is instantaneously made at transition, the amplitudes will vary as follows:

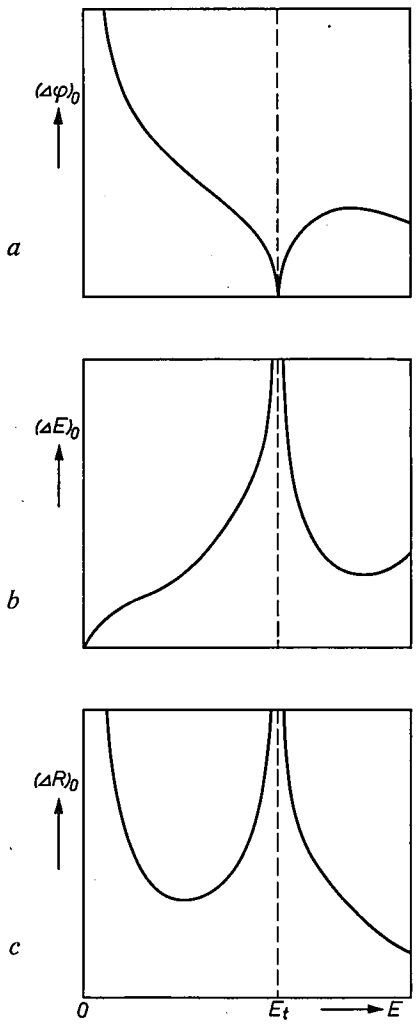


Fig. 18. The variation with energy of the amplitudes $(\Delta\varphi)_0$, $(\Delta E)_0$ and $(\Delta R)_0$ of the synchrotron oscillations, as given by the linear approximation. E_t transition energy.

$$(\Delta\varphi)_0 \propto (|1/\gamma^2 - 1/\nu^2|/\beta^2\gamma)^{1/4}, \quad \ldots \quad (29a)$$

$$(\Delta E)_0 \propto (|1/\gamma^2 - 1/\nu^2|/\beta^2\gamma)^{-1/4}, \quad \ldots \quad (29b)$$

$$(\Delta R)_0 \propto (\beta^6\gamma^3|1/\gamma^2 - 1/\nu^2|)^{-1/4}. \quad \ldots \quad (29c)$$

These variations of the amplitudes as functions of the energy are shown schematically in *fig. 18*. It can be seen that the variation of $(\Delta\varphi)_0$ is satisfactory: $(\Delta\varphi)_0$ tends to zero for $E \to E_t$, and then, after an initial rise, tends to zero again for $E \to \infty$. The variation of $(\Delta R)_0$ shown appears to be much less satisfactory, indeed even catastrophic, since after an initial decrease $(\Delta R)_0$ tends to infinity for $E \to E_t$. This would mean that near transition all the particles would be lost to the vacuum-chamber walls. Fortunately, eqs. (29) do not hold near transition, since the linear approximation cannot be used here; furthermore this effect is slow because the synchrotron frequency tends to zero at transition. Experience has shown that, if there is a good phase jump at transition, the amplitude of ΔR stays within a reasonable range.

The synchrotron oscillations in longitudinal phase space; r.f. buckets

As with the betatron oscillations in transverse phase space, the synchrotron oscillations can conveniently be represented in *longitudinal phase space* (*fig. 19a*), in which the abscissa is the phase φ, and the ordinate is the deviation ΔE from the synchronous particle energy. (We could also use t or s for the abscissa; these are proportional to φ apart from a slow variation of the constants. Similarly, for the ordinate we may also use the deviation ΔR from the synchronous orbit radius, or $\mathrm{d}\Delta\varphi/\mathrm{d}t$.) For small deviations from the synchronous position a particle behaves like a harmonic oscillator, and in longitudinal phase space its trajectory is an ellipse. If, however, the deviations are not very small, the trajectories are no longer elliptical. Fig. 19*a* shows

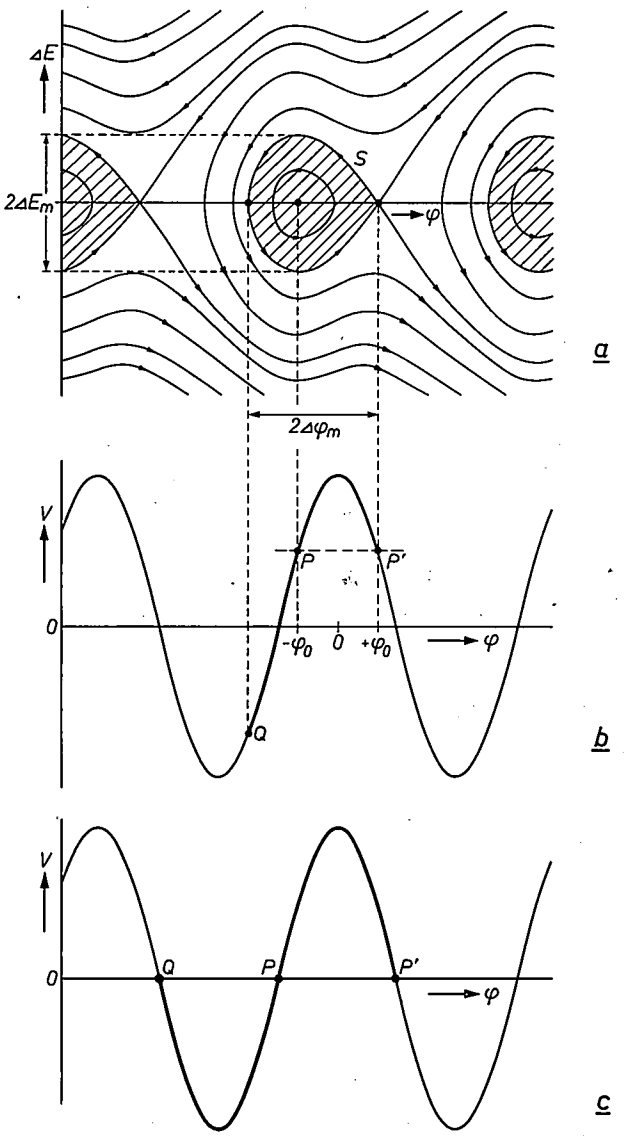


Fig. 19. *a*) Trajectories in longitudinal phase space. Only particles inside the "r.f. bucket", i.e. the hatched area inside the "separatrix" S, are stable. $2\Delta\varphi_m$ bucket length, $2\Delta E_m$ bucket height. The diagram refers to the case where $-\varphi_0$ is the stable and $+\varphi_0$ the unstable synchronous phase (i.e. to strong focusing before transition, $E_0 < E_t$). *b*) Gap voltage V as a function of phase φ. The r.f. bucket length $2\Delta\varphi_m$ is $\varphi(P') - \varphi(Q)$. Q is such that a particle starting at Q with nominal energy will arrive at P' with nominal energy. If $PP' = 0$ ($\varphi_0 = 0$) then $2\Delta\varphi_m = 0$. *c*) If $\varphi_0 = \pi/2$, $2\Delta\varphi_m = 2\pi$ (but the acceleration is zero).

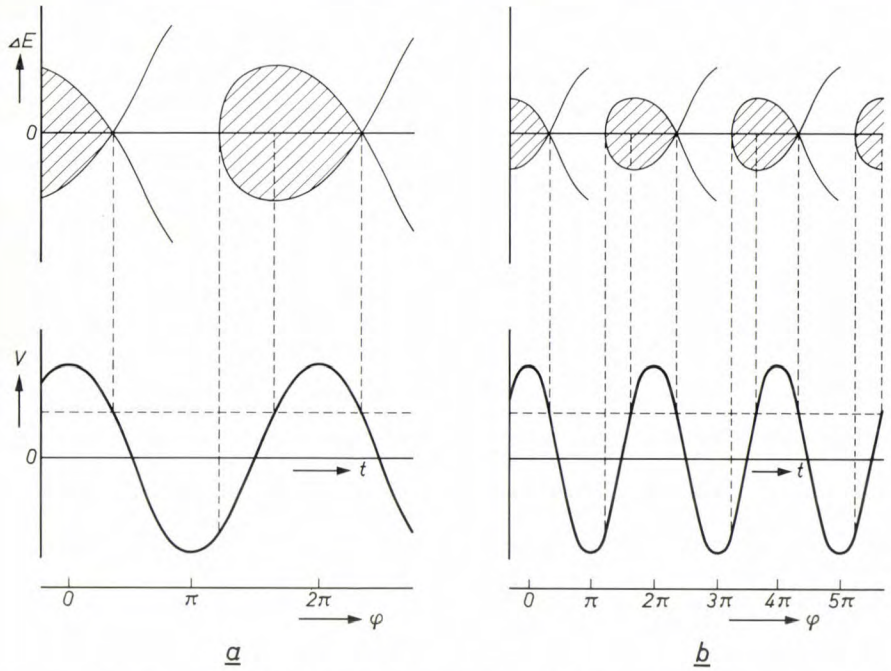


Fig. 20. Gap voltage V and r.f. buckets at two values of the harmonic number $h = f_a/f$, schematic. In (*b*) h is twice as large as it is in (*a*). The values of $\varphi = 2\pi hft$ + const. are indicated below the graphs for both cases.

a few trajectories. One of these is a critical trajectory, the *separatrix S*. In the hatched zone inside the separatrix loop, the trajectories are closed and the acceleration is stable. One synchrotron period is one turn in this space around the stable synchronous point. The particles outside the separatrix loop will diverge indefinitely and be lost.

A particle within a separatrix loop at a particular gap, being stable, will find itself within a separatrix loop in the next gap. The loops can be considered as "r.f. buckets" circulating around the ring, with the particles trapped inside them. There are h of these buckets around the ring. The particles falling inside a bucket at injection are "accepted" for acceleration: the buckets at injection represent a "longitudinal acceptance".

The relative bucket *height*, $2\Delta E_m/E_0$ (for ΔE_m see fig. 19*a*), is a particularly important factor when particles are transferred from one accelerator to another (as from the preaccelerator to the synchrotron).

Generally, the bucket *area* at injection should be as large as possible. A rough estimate of the bucket *length*, $2\Delta\varphi_m$, can be made by considering fig. 19*a* and *b*. The bucket length is $\varphi(P') - \varphi(Q)$, where P' is an unstable synchronous point next to the stable synchronous point P; Q is such that a particle starting at Q with nominal energy will arrive at P' with nominal energy. A particle with nominal energy just to the right of Q will oscillate around P; one just to the left of Q will shoot past P' and be lost. It is now clear that if $\varphi_0 = 0$ (i.e. if P' coincides with P) then $2\Delta\varphi_m = 0$, and, for reasons of symmetry (see fig. 19*c*), if $\varphi_0 = \pi/2$, then $2\Delta\varphi_m = 2\pi$. In the latter limit we have a maximum length, but the acceleration is zero. Using eq. (22), we see that $2\Delta\varphi_m$ decreases from 2π to 0 if $\dot{B}/V_m$ increases from 0 ($\varphi_0 = \pi/2$) to $M/2\pi R\varrho$ ($\varphi_0 = 0$). For the usual value $\varphi_0 = \pi/3$ the bucket length can be shown to be π. It can be shown that the bucket height also decreases for increasing $\dot{B}/V_m$. *Therefore, at injection, small values of* $\dot{B}/V_m$ *are favourable because the bucket area will be large.*

Another interesting point is the dependence of bucket size upon h. The bucket length, measured in phase, does not change very much if h is changed (see *fig. 20*). If now h is increased, let us say by a factor of 2, then, measured in time, or in length along the circumference, the bucket length will roughly be halved, but there will be twice as many buckets. The only major effect is therefore the decrease in bucket height. It can be shown that the bucket height is proportional to $h^{-1/2}$. This may influence the choice of h to be discussed later.

A number of turns after injection, the unstable particles are lost, and the stable particles are grouped in h bunches circulating around the ring (*fig. 21*). At injection these bunches fill the buckets completely, and therefore have a length of $2\Delta\varphi_m$. For $2\Delta\varphi_m = \pi$

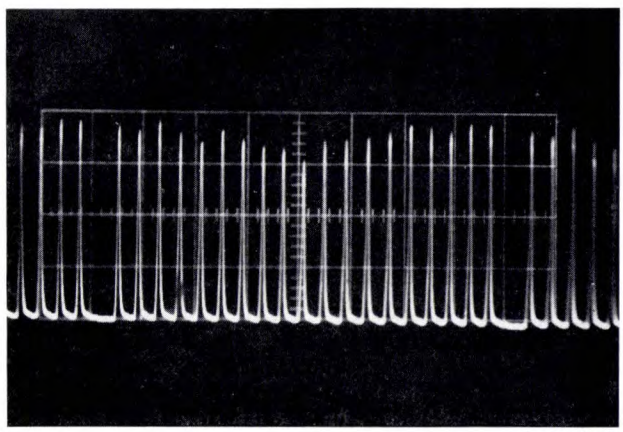


Photo CERN

Fig. 21. Pips on an oscilloscope representing bunches of particles in the CPS. Originally there were 20 bunches ($h = 20$) 105 ns apart, but one of the bunches has been ejected for experimental use and there is therefore no corresponding pip.

(corresponding to $\varphi_0 = \pi/3$), they occupy half of the circumference. During acceleration the length of these bunches will vary approximately in the way given by fig. 18*a*.

Space-charge effects

When the beam intensity is high, and particularly when the particles are tightly bunched, the Coulomb repulsion between the protons can have two noticeable effects:

a) In the transverse direction it counteracts the focusing forces of the magnet. This shifts ν, and if ν is drawn into a nearby resonance, a large loss of particles is the result. The shift in ν varies as $\beta^{-2}\gamma^{-3}$, and can be kept within bounds by choosing the injection energy sufficiently high.

b) In longitudinal phase space it tends to reduce the bucket area.

Radio-frequency control

For a given programme for the magnetic field the frequency of the accelerating voltage must be kept accurately equal to h times the frequency given by eq. (7). (As mentioned before, its amplitude, which should be constant if $\dot{B}$ is constant (eq. 22), is not exceptionally critical; a slight amplitude shift will just cause a slight shift of the synchronous phase φ_0 which is not disastrous.) It is impossible in practice to keep the frequency sufficiently accurately to the value given by eq. (7), so that it follows the magnetic field, merely by using an analogue process for reproducing eq. (7). Instead, once the particle bunches are formed, the phase of the voltage is locked to the bunches passing the gap by means of a feedback loop.

Synchronizing the voltage to the particle bunches, however, is not enough to produce a stable beam, as the beam might drift from the nominal orbit to neighbouring equilibrium orbits, the accompanying drift in frequency being followed by the voltage. To prevent this, a "radial" feedback loop is provided in which a radial deviation of the beam is translated into a relative phase shift of the voltage, thereby adjusting the amount of acceleration and readjusting the energy and orbit radius of the beam.

Two methods are in use for locking the accelerating voltage to the bunches. In one method the frequency of the voltage is made to follow the magnetic field given by eq. (7) by an analogue method and then corrected by a feedback loop. The second method makes use of a "bootstrapping" process: a signal derived from the bunches is used to generate the voltage across the gap (which, in turn, generates the bunches). Experience shows that there is no significant difference in performance between the two methods. In both cases there is of course a short period after injection during which the bunches are not yet properly formed. In this period an analogue method for reproducing the variation of eq. (7) must be used for generating the voltage.

Fig. 22 shows two simplified block diagrams to illustrate the two methods. In G_1 in fig. 22*a* a signal is generated whose frequency follows the flux density B in the way given by eq. (7). If the frequency f_p of this signal is not accurately equal to the frequency of the bunches of the beam b passing through the cylindrical pick-up electrode P_φ, it is corrected to $f_p + \varepsilon$ by the frequency-control unit F, and is then fed to a power amplifier A whose output is applied to the cavity gap g. The correction ε is proportional to the output of a phase discriminator D, which detects any phase difference between the signal from P_φ and the signal from the cavity. Radial deviations of the bunches are detected by the pick-up electrode P_R (an obliquely cut cylinder); in the phase shifter PS a signal from P_R will shift the phase of the input signal of D from the cavity, and thereby lead to an

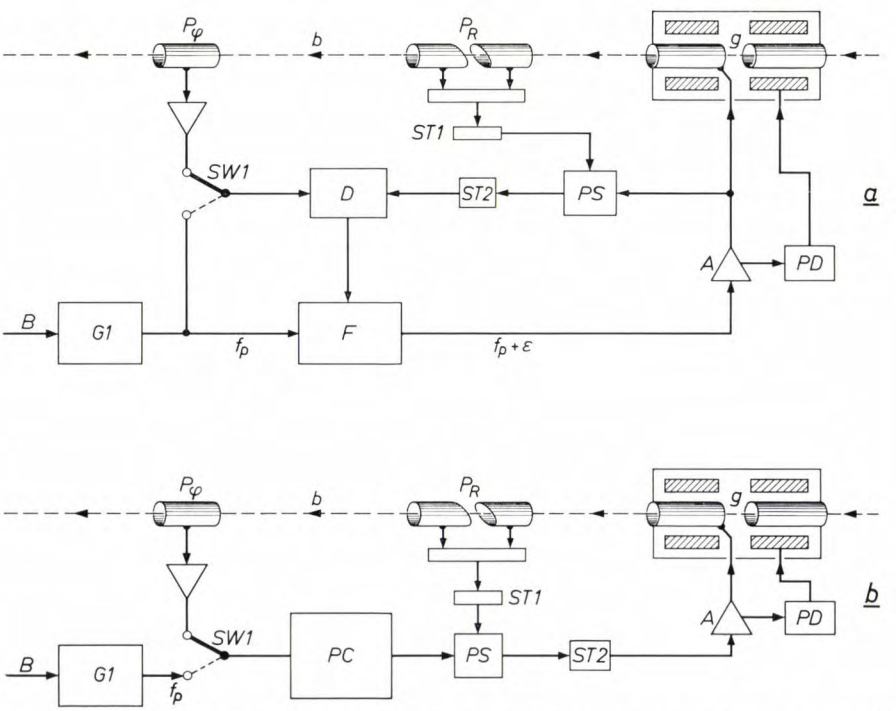
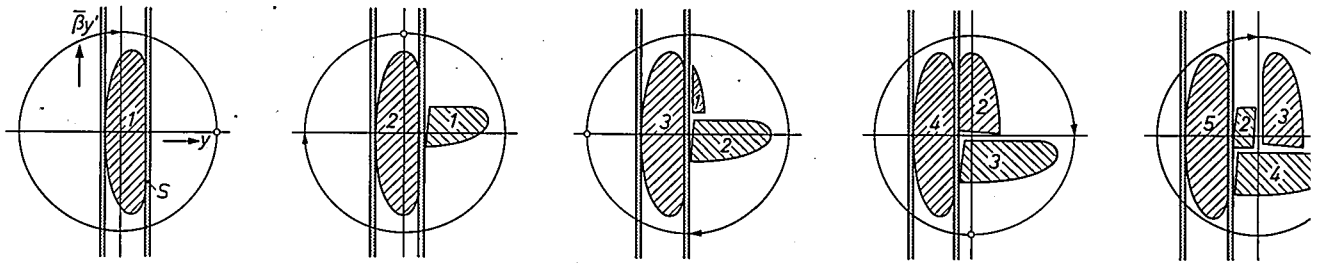


Fig. 22. Very schematic block diagram of beam-control systems. In (*a*) a signal is generated whose frequency varies with the field in such a way that eq. (7) is satisfied approximately; proper synchronization is obtained by a corrective feedback from the bunches. In (*b*) the voltage across the cavity gap is derived from the signals from the bunches (bootstrapping). A more detailed explanation is given in the text.



additional correction of the frequency. At transition, the signal from P_R is inverted by a switch ST_1 (as required by eq. 23-24), and the required jump of $2\varphi_0$ in the voltage phase is obtained by a switch ST_2 which adds $2\varphi_0$ to the phase. Before the bunches are formed the switch SW_1 is in the lower position.

Fig. 22*b* illustrates the bootstrapping method. After processing, the signal from P_φ is applied to the cavity. Corrections for radial deviations are made as above using an electrode P_R, and the switches ST_1, ST_2 and SW_1 have the same function as above. With the frequency varying so rapidly, it is by no means a simple problem to keep the phase difference between P_φ and the cavity gap at the required constant value. This requires a highly complicated multiheterodyne equipment, which is incorporated in the "phase-lag compensation unit" *PC*.

In both cases the tuning of the cavity to the frequency of the signal is performed by means of a magnetic field which polarizes a ferrite core in the cavity (see the preceding article). This field is produced by coils that are energized by a phase detector *PD* that responds to the phase difference between plate and grid of the last stage of the power amplifier *A*; this phase difference is directly related to the frequency of the signal.

Injection

A beam of particles, originating from a source and a preaccelerator, is injected into the ring. To obtain a sufficient number of particles that can be accelerated successfully, the beam to be injected should satisfy certain requirements. Another problem to be solved is that of injecting the beam into the ring in such a way that the circulation will not be affected by the presence of the injector.

The requirements for the beam are as follows. To be accepted by the ring, its emittance must fall within the synchrotron acceptance, both in the horizontal and the vertical plane (eq. 21). Secondly the spread in particle energy must be sufficiently small for the beam to be accepted by the r.f. accelerating voltage, i.e. it must be no larger than the "bucket height" $2\Delta E_m$ (fig. 19*a*). It should also be as concentrated as possible, the number of protons per pulse in the output of the synchrotron being directly proportional to the density of particles in the beam. Finally, the particle energy at injection should be sufficiently large.

Two different lower limits can be quoted for the particle energy. In the first place the magnetic field has a minimum value, given by the remanence of the magnet steel and by the condition that the field should be well defined, i.e. large compared to the stray magnetic fields around the ring. Eq. (5) then gives a minimum value for the particle momentum. For the CPS, with $\varrho \approx 70$ m and $B_{min} = 10$ milliteslas (100 gauss), we would obtain $T_{min} \approx 20$ MeV. Another lower limit to the particle energy is set by the space-charge effects (p. 347). In the CPS this limit is of the order of 50 MeV for 10^6 protons/cm^3.

The injection proper

The particles are injected into the ring by means of an electrostatic or magnetic inflector, providing a horizontal electric or a vertical magnetic field steering the particles away from their original path and into a direction approximately parallel to the equilibrium orbit in the ring (see *fig. 23*). The major problem is to prevent the particles from undergoing undue influence from the inflector at their subsequent transits. To solve this problem either the inflector must be "withdrawn" or the particles must be made to miss the back of the inflector at their subsequent transits. In principle it should be possible to achieve the second solution by making use of the fact that *at constant particle energy* the equilibrium orbit will shrink in the increasing magnetic field (the r.f. voltage is usually switched on

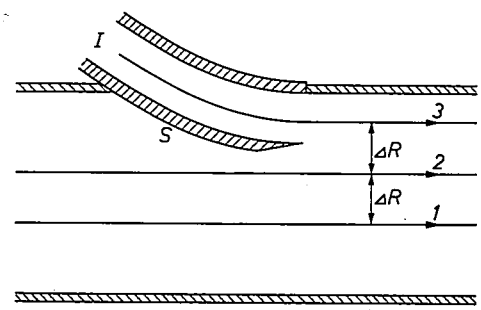


Fig. 23. Injection into a CG accelerator. *I* inflector, *S* septum of the inflector screening the vacuum-chamber space from the inflector fields. During injection (with zero accelerating voltage) the equilibrium-orbit radius shrinks by an amount ΔR for each turn in the increasing magnetic field. A particle injected when the equilibrium orbit is at *3* is at *2* after the first turn, and at *1* after the second turn. If the injection energy is boosted up during injection so that each injected particle enters the ring at *3* when the corresponding equilibrium orbit is at *3*, a beam can be continuously injected, giving a spiral of particles in the ring. Disadvantages of this method are a) the large momentum dispersion of the particles in the ring after injection, b) a severe loss of particles as ΔR is usually considerably smaller than the width of the beam. An alternative method is to make use of the betatron oscillations of the particles injected at constant energy (see text).

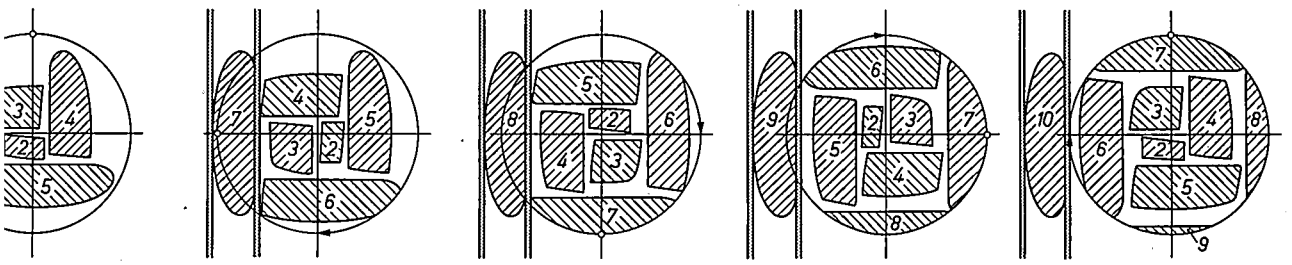


Fig. 24. Very schematic representation in horizontal transverse phase space of filling the horizontal acceptance of a CG accelerator (with $\nu_H = 0.75$) during nine turns, at constant energy. The ordinate is $\beta y'$, so that the betatron oscillations of the particles are represented by circles (see eq. 17 and 20). *S* inflector septum (see fig. 23). At each turn, the transverse phase space (y, $\beta y'$) with the horizontal acceptance (the circle) moves an amount ΔR to the right with respect to the inflector, and the injected particles (hatched areas) make three quarters of a turn in phase space ($\nu_H = 0.75$). The first and the ninth turns are nearly completely lost, leaving the protons of about seven turns in the ring after injection.

only after injection), so that a particle following the equilibrium orbit will spiral inwards. The change of radius of the equilibrium orbit per turn, ΔR, is given by:

$$\frac{\Delta R}{R} = -\alpha \frac{\dot{B}}{Bf}. \quad \ldots\ldots \quad (30)$$

If this method were used, particles would have to be injected at the instant that the equilibrium orbit coincides with the inflector axis (*3* in fig. 23); after one turn they would be in the orbit *2* and after two turns in the orbit *1*. This principle of injection cannot be used when the orbit shrinkage ΔR turns out to be too small, so that a large fraction of the particles are lost, hitting the back of the inflector after the first turn. But in fact as the equilibrium orbit moves away from the inflector axis the injected particles start to perform betatron oscillations around this orbit. Particles injected at *3* (fig. 23) at the instant that the equilibrium orbit is at *2* will perform betatron oscillations with an amplitude ΔR. Since ν is not an integer, they will be in a different phase at the second transit and miss the back of the inflector. With careful setting up it is possible to arrange that the particles will also miss the back of the inflector at a number of subsequent transits, and a number of turns can be injected in this way before the r.f. voltage is switched on.

Multiturn injection on this principle is typical for *CG* accelerators. The number of turns injected may vary from about 10 to more than 200. The method is illustrated for a hypothetical case in *fig. 24*, which shows how the machine horizontal acceptance in the horizontal transverse phase space is filled by the injected beam in nine turns; the successive injected turns are arranged in a spiral in this phase space. The inflector can be compared to a fixed gun shooting at a moving target with bullets that undergo betatron oscillations.

It has been assumed above that the beam is injected in the horizontal median plane. Injection *not* in the median plane can similarly be employed to make use of vertical betatron oscillations.

The fact that injection *in* the equilibrium orbit (without making use of betatron oscillations) is only possible at one particular instant can be circumvented by boosting up the injection energy during the injection period. After injection the machine is then filled with particles of different momentum in different equilibrium orbits. Owing to the momentum dispersion, however, only a fraction of the particles will be accepted by the r.f. buckets when the r.f. voltage is switched on. Another variation is to slowly accelerate the particles *in* the ring during the injection period. In fact all these methods have been tried with more or less success.

In *AG* accelerators the injection is complicated by the high momentum compaction, the small value of α giving a small rate of shrinkage of the equilibrium orbit (eq. 30). One solution to the problem is to omit the septum in the vacuum chamber, and to use an inflector consisting of two plates for electrostatic deflection, one on each side of the vacuum chamber. Any effect of the inflector on the particles after their first turn is eliminated by switching off the deflection field: after one revolution period the inflector is short-circuited in a time (≈ 0.1 μs) short with respect to the revolution period (≈ 7 μs), leaving one string of protons in the chamber. This is called *single-turn injection.*

If the beam emittance or its high-density core is much smaller than the accelerator acceptance, the latter is far from fully employed by single-turn injection. *Multiturn injection in AG accelerators* can be established by local manipulation of the equilibrium orbit, instead of relying upon the "natural" spiralization due to the radius shrinkage of eq. (30). The method is illustrated in *fig. 25*. A coil M_1 kicks an ideal particle in the equilibrium orbit *C* into a large betatron oscillation, and a coil M_2, half a betatron wavelength away from M_1, kicks it back into the original equilibrium orbit. The result is a new equilibrium orbit with a bump between

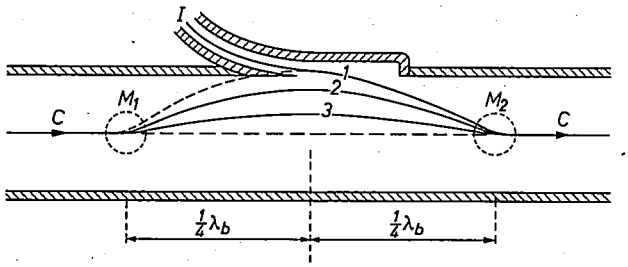


Fig. 25. Manipulation of the equilibrium orbit in AG accelerators. A coil M_1 kicks an ideal particle in the equilibrium orbit C into oscillation and M_2 kicks it back into the orbit C. The distance M_1M_2 is half of the betatron wavelength $\lambda_b = 2\pi R/\nu$. The result is a new equilibrium orbit with a bump, which is made to decrease (*1,2,3,...*) during the injection period by adjusting the coil currents. The inflector mouth lies centrally between M_1 and M_2.

M_1 and M_2, but coinciding with the original equilibrium orbit elsewhere. By varying the coil currents the size of the bump can be varied, and the bump is made to decrease (*1*, *2*, *3* in fig. 25). Injection then proceeds as described above. Particles are shot into the ring by an inflector on one side of the vacuum chamber. The injection is started when the (new) equilibrium orbit is in the inflector axis. As the equilibrium-orbit bump decreases, the injected beam will undergo a series of betatron oscillations of increasing amplitudes. With careful arrangement a number of turns can be made to miss the inflector and fill up the synchrotron acceptance. A 3-turn injection is illustrated in *fig. 26*.

R.f. trapping

Once a beam of particles has been properly injected, the ring is full of particles, corresponding to a continuous ribbon N in the longitudinal phase space (*fig. 27a*); then the r.f. voltage is switched on. The particles inside the r.f. buckets in longitudinal phase space (cross-hatched in fig. 27*a*) are trapped and accelerated to the end of the cycle; the others are lost. The size of the buckets, i.e. the trapping efficiency is largely determined by $\dot{B}/V_m$, as we have seen before (p. 346). For the normal value of $\dot{B}/V_m$, for which the synchronous phase angle $\varphi_0 \approx \pi/3$, the bucket length is π, so that half of the length of the ribbon N can supply useful particles. As to the height of the buckets, an estimate shows that $2\Delta E_m/E_0$ is quite often of the order of 10^{-3}; this is the accuracy to which the particle energy in the injected beam must be defined.

R.f. trapping is illustrated rather neatly by the oscilloscope traces in *fig. 28*.

Efforts to increase the trapping efficiency are often directed towards using a smaller $\dot{B}$ at injection than during acceleration. Changing $\dot{B}$, however, is a tricky operation. The frequency programme should follow it carefully. Moreover, if the later increase of $\dot{B}$ to the normal value and the accompanying "shrinkage" of the buckets are not sufficiently slow, particles may be lost, because the r.f. buckets become smaller than the beam bunches in longitudinal phase space (whose dimensions shrink during acceleration in the way given by fig. 18*a*).

Another way of increasing the trapping efficiency is a kind of trick discovered by accident at CERN. Sup-

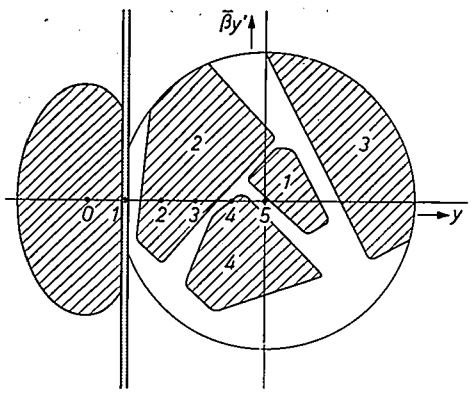


Fig. 26. Filling the horizontal acceptance of an AG accelerator (with $\nu = 6.2$) during 4 turns at constant energy in a way similar to that of fig. 24. The particles now make 6.2 turns in phase space for each turn in the ring. The injected beam appears relatively wider than in fig. 24 because q is smaller in AG accelerators (see eq. 21), and less elongated as the factor $\bar{\beta} = R_0/\nu$ is smaller.

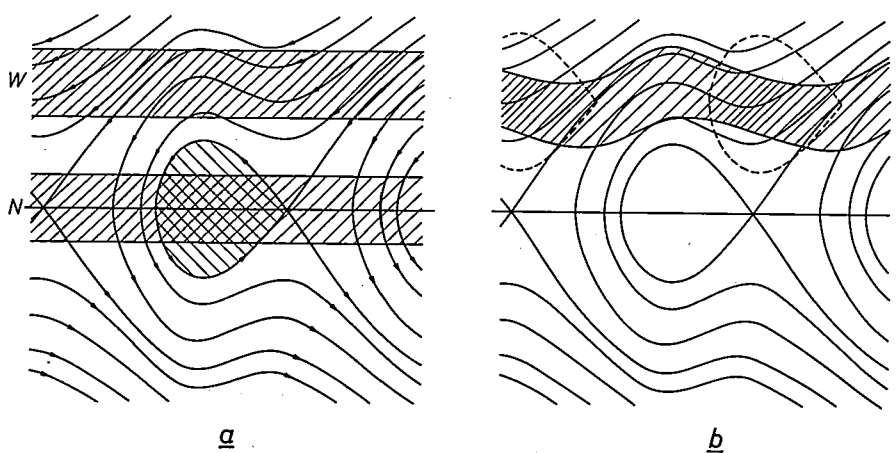


Fig. 27. *a*) If the r.f. voltage is switched on at the appropriate frequency, the proton beam covers a central band N in longitudinal phase space (φ, ΔE), and the particles inside the buckets (the particles in the cross-hatched area) are trapped. The feedback loop is then closed. If the r.f. voltage is switched on at too low a frequency, the beam covers a band above the central line (W). The particles then move along the trajectories, and after some time there is slight bunching (*b*). If the feedback loop is now closed the buckets move to enclose the bunches.

pose that the r.f. voltage is switched on without feedback from the beam (SW_1 down in fig. 22), and at a slightly wrong frequency (1.5% too low, for instance). The string of protons (W in fig. 27*a*) is in the unstable region. It will start moving along the unstable trajectories. As the motion is not uniform, a slight bunching effect will start to appear and consequently a signal will arise at the pick-up electrodes (fig. 27*b*). If now after an appropriate time (the "take-over time", of the order of 50 μs, a fraction of the period of the synchrotron oscillations) the feedback loop is closed (SW_1 up in fig. 22), the r.f. frequency and phase will be readjusted in such a way that the r.f. bucket encloses the region of higher density. When using this method, an r.f. trapping efficiency of 70% can be expected.

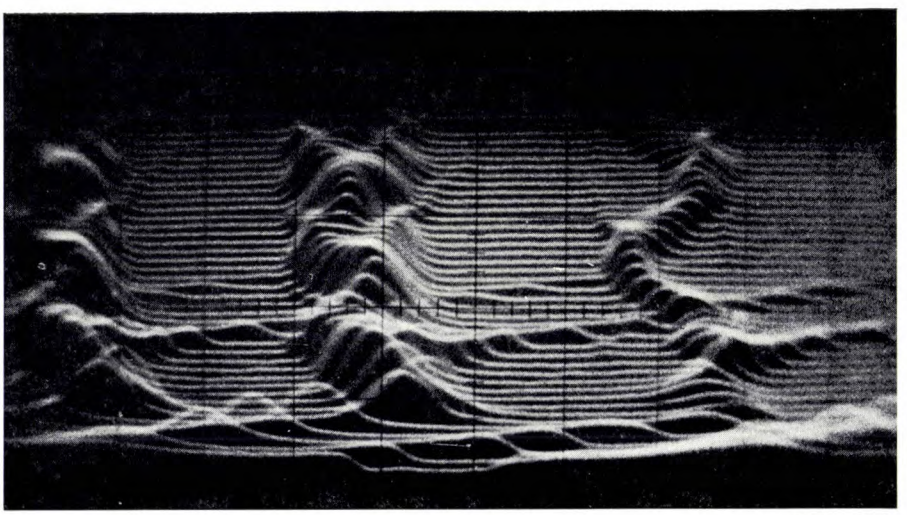


Photo CERN

Fig. 28. R.f. trapping shown on an oscilloscope. The behaviour of the beam during its first forty turns in the machine is shown. The vertical signal is from a pick-up station in the ring; the oscilloscope trigger is synchronized to the accelerating voltage. A hump in the trace represents a bunch passing the pick-up station. After each turn of the beam around the ring the trace is moved up a little. After the first few turns bunches start to appear, induced by the voltage. After six turns the feedback loop is closed, synchronizing bunches and r.f. voltage.

Final beam dimensions

In the accelerator the vertical width of the beam is largely determined by the betatron amplitudes; it will therefore shrink in proportion to $B^{-1/2}$. If, after injection, the beam emittance fills the machine acceptance properly, we may then expect the vertical width of the beam after acceleration to be:

$$W \approx 2q_v \sqrt{B_i/B_f}, \quad \ldots \quad (31)$$

where $2q_v$ is the vertical vacuum-chamber aperture, and B_i and B_f are the initial and final flux densities respectively. Owing to momentum dispersion of the injected beam and synchrotron oscillations the horizontal width is always larger than the vertical width. In the machines under discussion the ratio B_f/B_i is of the order of 100, and the final beam width would be expected to be roughly 1/10 of the chamber diameter. For the same reasons (energy spread, synchrotron oscillations, misalignment, etc.) that the acceptance is usually smaller than eq. (21) indicates, the width should be smaller, but on the other hand it is enlarged by scattering in residual gas, beam expansion at transition, space-charge effects, various non-linearities, etc.

II. Detailed information about the five accelerators

We have collected various data characterizing the five synchrotrons. Each of the machines has its own very special features, and we shall try to present the data in such a way that these features are brought out clearly.

We ought to warn the reader that although some quantities, such as the ring diameter, are fixed for each of the five accelerators, others may be altered from time to time, since accelerator technique is advancing rapidly with practically day-to-day improvements. Consequently, some of the data relating to intensity, ion-source type, injection procedure, regulation proced-

Photo CERN

Fig. 29 Aerial view of the CERN Proton Synchrotron with experimental halls and associated buildings (1967).

ure, etc., may very well be out of date at the time of printing.

Broadly speaking, the purpose of an accelerator is to provide protons of kinetic energies as high as possible, in quantities as large as possible, at a cost which is economically acceptable. However, even for modest energies, knowledge of the possible processes is still far from complete and low-energy data can bring interesting information to the nuclear world. As physicists are often seeking very rare events, high-intensity proton beams — giving a relatively high probability for the occurrence of such events in a given period — are more and more in demand. Of the two related aims, high energy and high intensity, the first takes precedence with the AGS and the CPS, and the second with the ZGS, Nimrod and the PPA.

Table II. Energy, magnetic field, cycle period and beam intensity (late 1969).

		AGS	CPS	ZGS	Nimrod	PPA
Final kinetic energy (T_f)	GeV	33	28	12.7	7	3
Flux density of final magnetic field (B_f)	teslas [*]	1.3	1.4	2.15	1.4	1.4
Intensity (protons per pulse)		3×10^{12}	2×10^{12}	3.3×10^{12}	1.8×10^{12}	4×10^{10}
Typical cycle period [**]	s	2.4	2.5	2.3	2.7	0.054
Corresponding average current	μA	0.20	0.1	0.23	0.11	0.12
Injection kinetic energy (T_i)	MeV	50	50	50	15	3.0
Flux density of injection field (B_i)	milliteslas	12.1	14.7	48.2	29.9	27.0

[*] 1 tesla = 10 kilogauss.
[**] These values are purely indicative, the period may be varied considerably (except at PPA).

Trying to characterize each of the five machines in a few words by their most original features we might say that the AGS and the CPS were the first proton accelerators to apply the AG principle, to pass through transition and to reach the 30 GeV range, that the ZGS is the largest machine with a zero gradient, allowing it to be very compact, that Nimrod has a huge vacuum chamber for super-high intensities and that the PPA is the first fast-cycling high-energy proton accelerator.

Fig. 29 gives an aerial view of the CPS. The ZGS has been shown earlier in fig. 1, the PPA in fig. 2, and the AGS ring in fig. 3 (and its interior in fig. 2 of the preceding article). Fig. 33 shows systems for inflection and ejection at Nimrod.

In Table III it can be seen that the fraction of the ring occupied by magnets does not vary widely from one machine to another, it is roughly 0.7-0.8. The magnet arrangement, however, is quite different, as we have already seen in fig. 6. Some details of this arrangement are given in *Table IV*. In the AGS the individual magnets are so short that they can be straight; in all the other machines the units are curved. In view of the rapid field variation all the units are laminated.

The magnet arrangement for the ZGS is shown in *fig. 31* and a magnet octant for this machine in *fig. 32*. Parts of the magnet ring of Nimrod are shown in *fig. 33a* and *b*. An impression of the magnet units for the other machines is obtained from photos shown

Table III. Size of the ring. Fraction occupied by magnets.

		AGS	CPS	ZGS	Nimrod	PPA
Circumference ($2\pi R_0$)	m	807	628	172	148	78.58
Mean radius (R_0)	m	128.5	100	27.4	23.6	12.51
Radius of curvature (ϱ)	m	85.37	70	20.8	18.8	9.145
Fraction occupied by magnets ($\approx \varrho/R_0$)		0.63	0.68	0.76	0.80	0.731
Orbital frequency at velocity c ($f_c = c/2\pi R_0$)	MHz	0.372	0.478	1.75	2.02	3.815
Mass of magnet steel	t	4000	3000	4700	7000	400

Radius, energy and magnetic field

Table II shows the initial and final values of the particle energy and the magnetic field (flux density), as well as the beam intensity and the cycle period. The size of the ring and the fraction occupied by magnets is given in *Table III*. Size, kinetic energy and magnetic field are conveniently summarized in *fig. 30*. Two things may be noted in this figure: a) the final field of the ZGS is considerably larger than the other final fields, giving a relatively compact machine at a given final energy; b) the frequency swing for small machines, in particular for the PPA, is large (as discussed earlier).

In all cases except one, B_f, the flux density of the final field at the nominal orbit, is about 1.4 teslas (at Nimrod this can be pushed to 1.6 teslas to reach 8 GeV). This value is mainly determined by the consideration that, even in high-quality magnet steel, serious fluctuations in the magnetic field will appear if the material is saturated. The maximum acceptable flux density B_{max} in this respect is usually around 2.0 teslas. The fact that the condition $B < B_{max}$ must be satisfied over the cross-sectional area of the vacuum chamber means that, if a field gradient is present, B_f must remain considerably smaller than B_{max}. The ZGS, with zero gradient, is the exception. We shall return to this point when considering the "useful aperture".

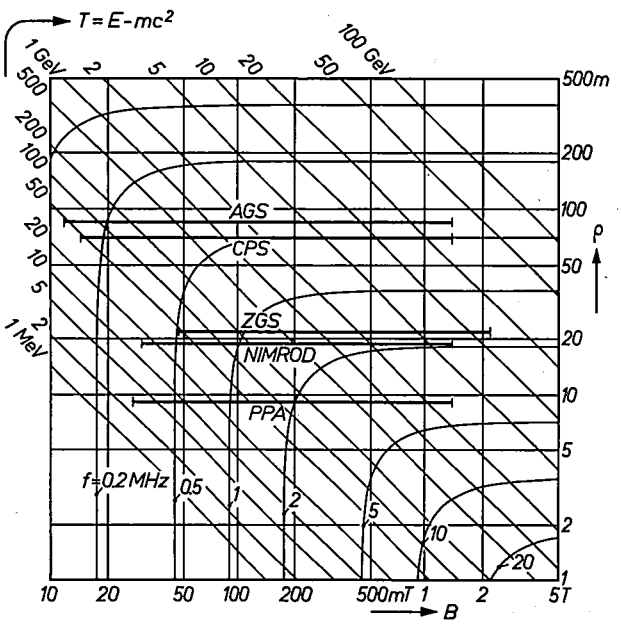


Fig. 30. Radius of curvature ϱ, flux density B of the magnetic field and kinetic energy T in the five accelerators. B is plotted horizontally, ϱ vertically, both on a logarithmic scale. Each accelerator is represented by a horizontal line between the field at injection and the final field. Lines of constant T given by the relations (3), (4) and (5) are shown, and lines of constant orbital frequency f given by eq. (6). (The value of f in the B-ϱ diagram depends on the value of ϱ/R; for simplicity we have assumed one single value, $\varrho/R = 0.7$.) The bending of the lines of constant frequency marks the transition from "pre-relativistic" to "relativistic". The diagram shows clearly that acceleration in the pre-relativistic range implies a large frequency swing.

earlier (PPA, fig. 2; CPS, fig. 4; for the AGS, see fig. 2 of the preceding article).

All the magnets have a C-shaped cross-section, except the ZGS magnet. To construct a magnet of sufficient mechanical strength is in itself a problem, in particular for the heavier magnets (ZGS, Nimrod). From this point of view the ZGS design has great advantages. On the other hand, the C shape allows much easier access to equipment in the vacuum vessel and it does not restrict the path for extracted proton and secundary beams, and for this reason the C shape has been adhered to in Nimrod (see fig. 5).

The AG and the CG principle; combined and separate functions

In the AGS and the CPS strong focusing (large ν values) is obtained by applying the AG principle; the other three machines are weak-focusing machines, with ν values less than unity. The focusing characteristics are given in *Table V*.

As we have seen before, large values of ν are advantageous for two reasons. a) The number of particles that can be accelerated is proportional to $A_H A_V$, the product of the acceptances ($A = \pi q^2 \nu / FR$) in the horizontal and the vertical planes, and thus proportional to $\nu_H \nu_V$. An increase in the ν's leads directly to an increase in the accelerated beam intensity. An increase of the acceptances can otherwise only be obtained by increasing the vacuum-chamber cross-section, which is an expensive undertaking since the magnet cross-section then has to be increased in about the same proportion. b) A large ν_H goes hand-in-hand with

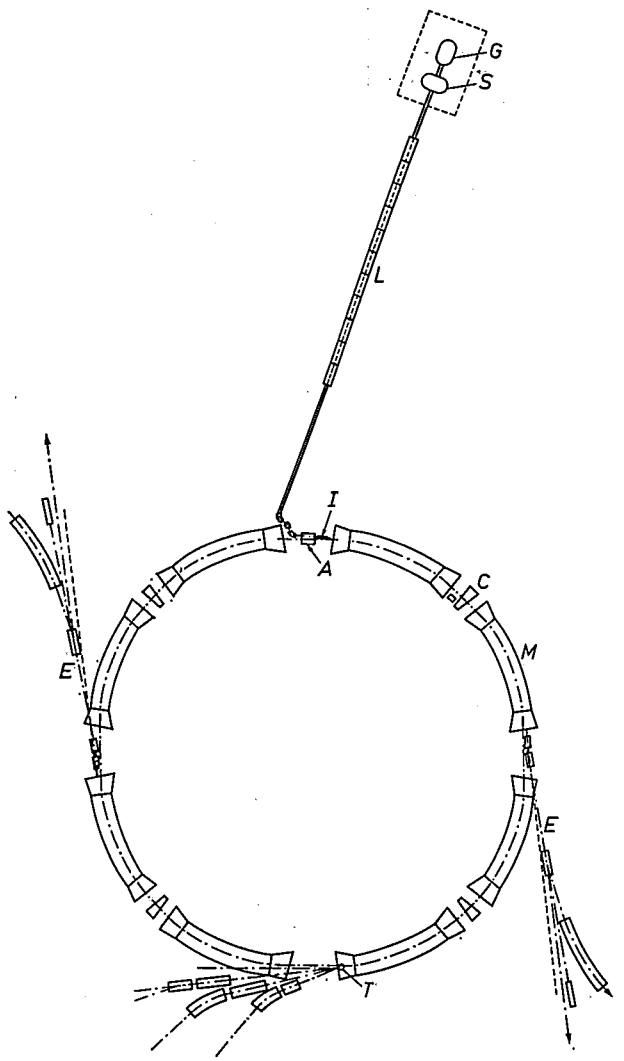


Fig. 31. Plan view of the magnet arrangement of the ZGS. Note the angles of the magnet edges. *M* main bending and edge-focusing magnets. *C* correcting magnets. *G* high-voltage generator. *S* ion source. *L* linear accelerator. *I* inflector. *A* acceleration station. *E* ejected beams. *T* internal target.

Table IV. Structure of the ring.

F : horizontally focusing magnet (fig. 10*b*).
D : horizontally defocusing magnet (fig. 10*a*).
O : straight field-free section.
O′: very short straight section.

AGS	D and F magnets, 60 periods (FO′D)O(DO′F)O	
	144 F or D of 2.29 m	= 329.6 m
	96 F or D of 1.90 m	= 182.4 m
	240 magnets	512.0 m [*]
	24×O of 3.15 m	= 75.6 m
	72×O of 1.63 m	= 117.4 m
	144×O′ of 0.71 m	= 102.2 m
	$2\pi R_0$	= 807 m
	$2\pi\varrho$	= 536 m [*]
CPS	DF and FD units, 50 periods FDODFO	
	50 DF of 4.26 m	= 213 m
	50 FD of 4.26 m	= 213 m
	100 units	= 426 m [*]
	80×O of 1.74	= 139.2 m
	20×O of 3.14	= 62.8 m
	$2\pi R_0$	= 628 m
	$2\pi\varrho$	= 440 m [*]
ZGS	flat-pole magnets	
	8 octants of 16.3 m	= 130.4 m
	8 straight sections, total length	= 41.45 m
	$2\pi R_0$	= 172 m
	$2\pi\varrho$	= 130.4 m
Nimrod	D type magnets	
	8 octants of 14.75 m	= 118.0 m
	8 straight sections, total length	= 30.4 m
	$2\pi R_0$	= 148 m
	$2\pi\varrho$	= 118.0 m
PPA	D type magnets	
	16 D of 3.591 m	= 57.46 m
	8 O of 1.981 m	= 15.85 m
	8 O′ of 0.660 m	= 5.28 m
	$2\pi R_0$	= 78.58 m
	$2\pi\varrho$	= 57.46 m

[*] The difference between the total length of the magnets and $2\pi\varrho$, where ϱ (Table III) is determined by p and B (eq. 5), is due to fringe fields.

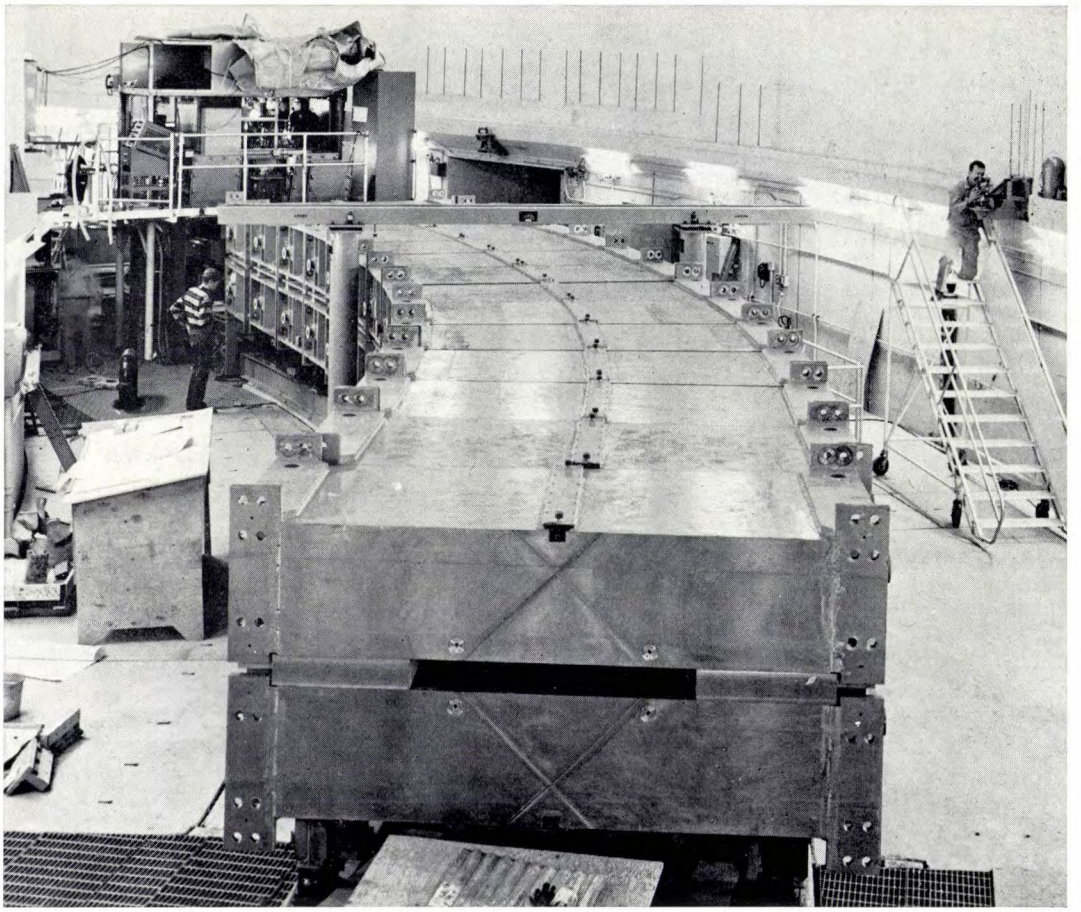


Photo Argonne National Laboratory

Fig. 32. A magnet octant under construction for the ZGS.

strong momentum compaction. If the particles in a beam have a certain spread in energy (injected beam energy dispersion, synchrotron oscillations), the accompanying spread in equilibrium orbits is small if the momentum compaction is strong. From eq. (14), using the relations (1) to (4), we obtain:

$$\frac{\Delta R}{R} = \frac{1}{\nu_H^2} \frac{\gamma}{\gamma + 1} \frac{\Delta T}{T}.$$

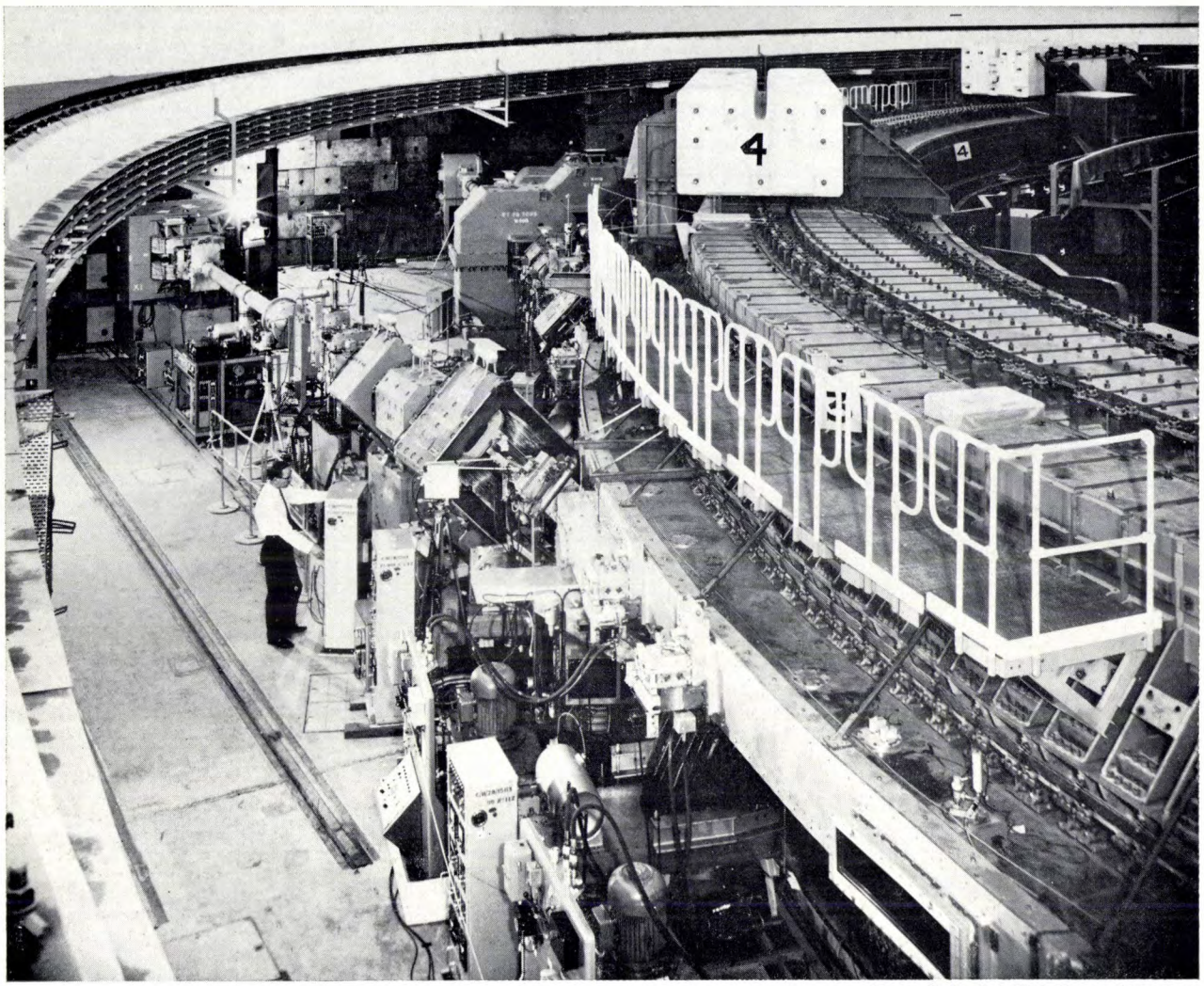
If $\Delta T/T = 10^{-3}$, $\gamma \approx 1$ and $R = 100$ m (CPS at injection), then, for $\nu = 6.25$, we have $\Delta R = 1$ mm. For $\nu = 0.7$ the spread in radii would be no less than 10 cm. In AG accelerators the compaction is of particular importance as a counterbalance to the beam expansion near transition (see fig. 18*c*).

However, if strong focusing is obtained by a large (alternating) field gradient n of the individual bending magnets (as in the AGS and the CPS), then there is a limitation to the method, which shows up as a limit to the *maximum useful aperture*. Before going into this briefly, we note that this limitation for strong focusing applies to combined-function machines only, where the same magnetic fields are used for focusing and bending and where ν and n are directly related.

If q is the semi-aperture and B_f is the flux density of the final field at the nominal orbit, then $B_f + |q(\partial B/\partial r)_f|$ must remain below the maximum acceptable flux den-

Table V. Focusing.

		AGS	CPS	ZGS	Nimrod	PPA
Focusing type		AG	AG	edge	CG	CG
Edge angles				13.26°; 9.16°		
Field index n		± 357	± 288	0	0.6	0.65-0.58
ν_H		8.75	6.25	0.83	0.71	0.69-0.78
ν_V		8.75	6.25	0.81	0.87	0.94-0.87
$\alpha \approx 1/\nu_H^2$		0.013	0.026	1.7	2.0	1.84-2.10
Transition energy						
total : $E_t = \nu mc^2$	GeV	8.2	5.8			
kinetic: $T_t = (\nu - 1)mc^2$	GeV	7.3	4.9			
$\bar{\beta} = R_0/\nu$	m	14.7	16	{ H 33 / V 24	H 33 / V 27	H 18 / V 14
Form factor $F = \beta_{max}/\bar{\beta}$		1.35	1.35		1	1



Photos Rutherford High Energy Laboratory

Fig. 33*a*. Part of the magnet ring of Nimrod, with equipment for beam ejection.

sity B_{max} (around 2.0 teslas). Using eq. (9) we obtain:

$$q < q_{max} = (B_{max}/B_f - 1)\varrho/n, \quad . \quad . \quad (32)$$

where q_{max} is the "maximum useful semi-aperture"; n/ϱ is called the "profile parameter". *Table VI* gives values of n/ϱ (as derived from Table V and Table III), of q_{max} (with $B_{max} = 2.0$ teslas, $B_f = 1.4$ teslas, $B_{max}/B_f - 1 = 0.43$) and of the actual vacuum-chamber horizontal semi-aperture q_H. It can be seen that the AGS and the CPS have used up more than half of their maximum useful aperture, whereas in the CG accelerators the maximum useful aperture is very far from setting a practical limit to the dimensions of the cross-section. Thus, if combined-function machines only are considered, CG machines can be increased in aperture more easily than AG machines. Of course, the price to pay for such an increase in aperture would be an increase in the required amount of magnet steel.

The exceptional position of the ZGS is evident from Table VI: $n = 0$, $q_{max} = \infty$, and there is no limit to the aperture on field-gradient considerations. This can be considered as a result of the fact that the ZGS is, in one respect, a separated-function (CG) machine: bending and focusing are established in different parts of the field, and n and ν are no longer directly related. As a consequence B_f can be chosen arbitrarily close to B_{max} (see eq. 32). In the ZGS this has been exploited by taking special measures to push B_{max} as far as pos-

Table VI. Maximum useful and real semi-aperture (horizontal), q_{max} and q_H; q_{max} has been calculated for $B_{max} = 2.0$ teslas, $B_f = 1.4$ teslas.

		AGS	CPS	ZGS	Nimrod	PPA
n/ϱ	m^{-1}	4.2	4.1	0	0.03_2	0.06_6
q_{max}	cm	10	10	∞	1300	650
q_H	cm	6.7	7.3	41	50	9

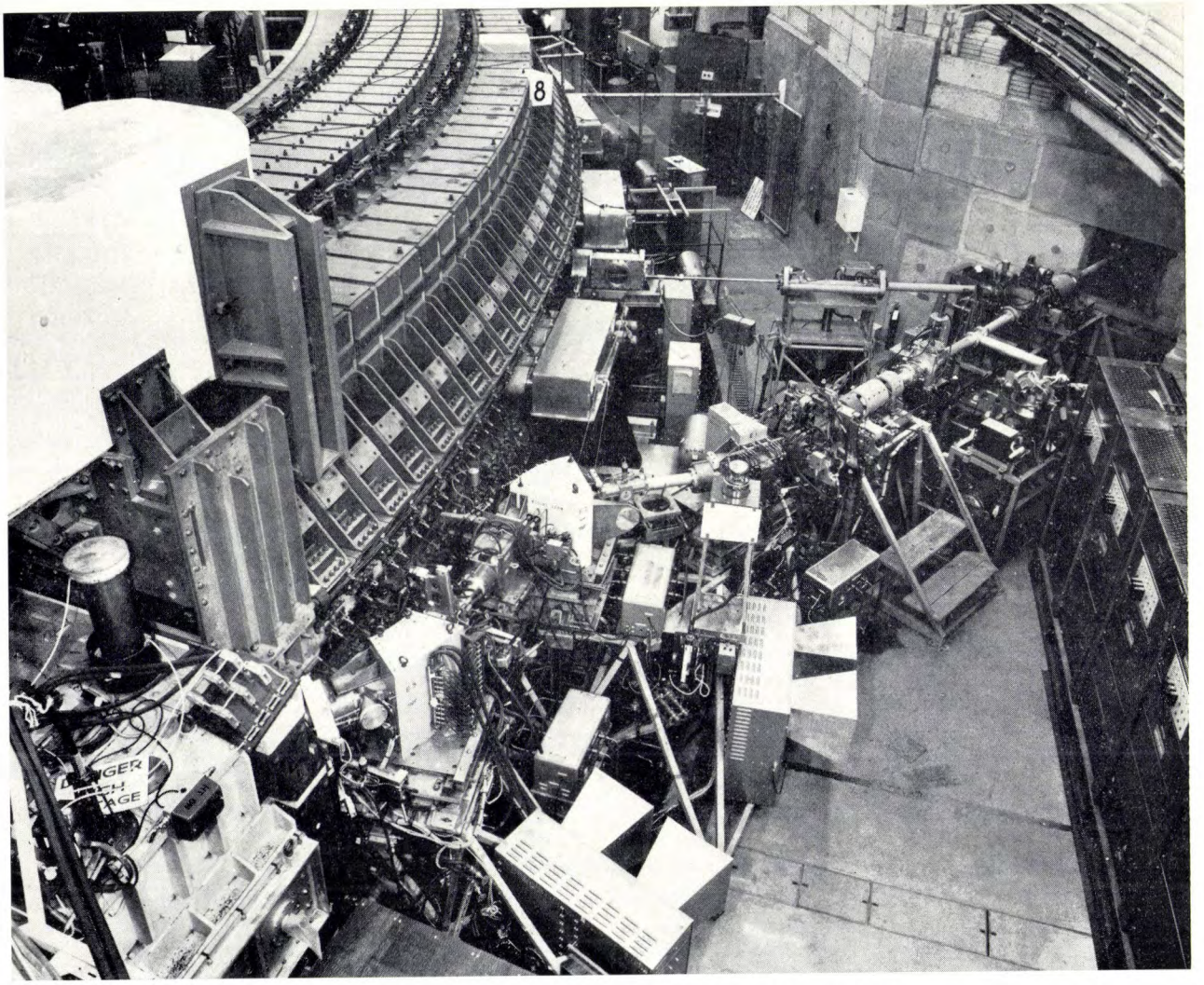


Fig. 33*b*. Part of the magnet ring of Nimrod, with injection line.

sible. *Fig. 34* is a cross-section of the magnet. It was found that, in general, the field becomes inhomogeneous above 1.8 teslas, because of saturation of the magnet steel near the coils; however, by drilling holes in the steel in this region, the field distribution was changed favourably, giving a maximum acceptable field of 2.15 teslas. The result is, as mentioned before, a relatively compact machine.

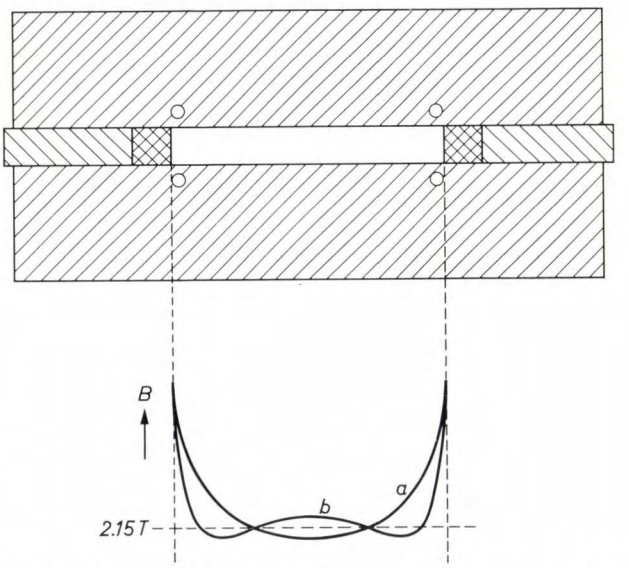


Fig. 34. Cross-section of the ZGS magnet, schematic. By drilling holes in the magnet steel near the coils, the field has been made more uniform in the high-field region where saturation occurs. Lower diagram: the magnetic field distribution for a central field value of 2.15 teslas, *a* without holes, *b* with holes.

The potential beam intensity

Let us consider for a moment the number of particles that can be accelerated by the various machines. This we can do by considering a beam from the same hypothetical preaccelerator, "tapped off" at the required injection energy, that (whether in one or in many turns) will completely fill the machine acceptance, in both the horizontal and the vertical plane. The number of particles accelerated will then be proportional to both the horizontal and the vertical *acceptance* of the machine.

In the preaccelerator the beam will shrink during

acceleration, and the particle density in transverse phase space will increase with the energy (this density is proportional to $\beta\gamma$, see p. 342). Thus, for a given acceptance, more particles will be accepted if they are "tapped off" from our hypothetical preaccelerator at a higher energy. This means that the *invariant acceptance*, i.e. the acceptance normalized from the individual injection energy to the same energy for all cases, is a better quantity for comparing structures than the acceptance pure and simple. The acceptance is generally normalized to the energy at which $\beta\gamma = 1$ (i.e. $\gamma = \sqrt{2}$, $\beta = 1/\sqrt{2}$); the invariant acceptance is then $\beta_1\gamma_1 A$.

The theoretical ideal horizontal acceptance given by eq. (21), $A_H = \pi q_H^2 \nu_H / F R_0$, and its normalized value are shown in *Table VII*. (The theoretical vertical acceptances will vary less from one machine to another as the variation in the vertical dimension is smaller than in the horizontal dimension of the cross-section of the vacuum chamber.) As discussed earlier, the practical acceptances will be considerably smaller than the theoretical ones, since space is needed for momentum dispersion of the injected beam, for synchrotron oscillations, and for lateral displacements of the central orbit caused by mechanical or magnetic defects. This extra space is relatively large in CG machines, the momentum compaction being weak.

Table VII. Acceptances (in mm × mrad).

	AGS	CPS	ZGS	Nimrod	PPA
Theoretical horizontal acceptance ($A_H = \pi q_H^2 \nu_H / F R_0$)	230 π	250 π	5100 π	7500 π	450 π
Theoretical invariant horizontal acceptance $\beta_1\gamma_1 A_H$, at injection	76 π	82 π	1700 π	1300 π	40 π

By far the largest theoretical acceptances are those for Nimrod and the ZGS, two machines that have been designed for high intensities. The main factor in obtaining these high values is the *large vacuum-chamber cross-section*, on which the acceptance is strongly dependent (eq. 21).

In the AGS and the CPS the q's are much smaller and, together with the larger R (and, much less important, the larger F), this would lead to very much smaller acceptances; this is compensated to some extent by the larger values of ν.

The favourable influence of a high injection energy on the "potential beam intensity" (as expressed by normalization of the acceptance) is illustrated by comparing the AGS and the PPA in Table VII: in the PPA the acceptance is decreased by a factor of more than 10 by the normalization, in the AGS by a factor of only 3. A high injection energy also pushes up the limit imposed on the particle density by space-charge effects. It should of course be realized that in practice the price that has to be paid for the advantage of a high injection energy is that a more powerful injector has to be used [8].

In fact, the product of the invariant acceptances in the horizontal and the vertical plane is not yet a very accurate measure of the "potential beam intensity". The number of accelerated protons per pulse (apart from those lost in r.f. trapping) is equal to the number of particles filling the ring after injection. Therefore the circumference of the ring will be of significance, as will also the longitudinal particle density in the beam, which decreases as the injection velocity increases. More important is the fact that it is the number of accelerated particles *per second* that is really interesting. Thus the PPA, a machine of only modest vacuum-chamber cross-section and injection energy, can nevertheless compete with the others, because its *repetition rate* — to which the number of accelerated particles per second is proportional — is at least 40 times as large as in the other machines (see Table II).

The ratios of the intensities that have been achieved up till now are not represented very well by the ratios of the "potential beam intensities" calculated as indicated above, because of differences between the injected beams (quality of the source and the preaccelerator, effectiveness of the injection procedure). In addition space-charge defocusing effects, beam instabilities, nonlinearities, etc. still affect the intensities to a varying extent.

The vacuum chamber

The problems met in designing the vacuum chamber are again different from machine to machine. Fig. 7 is an illustration of how widely the cross-sectional sizes vary. *Table VIII* gives some details of the size, shape, wall material and means of maintaining the vacuum.

Apart from being vacuum-tight, the vacuum-chamber wall should have various other properties. It should be non-magnetic so that it has no undue influence on the magnetic field; as the magnetic field is rapidly changing it should preferably be non-conducting to prevent eddy currents from changing the field; it must be strong enough to withstand the atmospheric pres-

Table VIII. Vacuum chamber.

	AGS	CPS	ZGS	Nimrod	PPA
Type	single	single	double	double	single
Wall material	Inconel	stainless steel	stainless steel	epoxy resin + fibreglass	copper nickel ribs, epoxy + fibreglass
Cross-section { size $2q_H \times 2q_V$ (cm × cm)	13.3 × 6.35	14.6 × 7.0	81.3 × 13.3	100 × 20	18 × 6
Cross-section { shape	elliptical	elliptical	rectangular	rectangular	rectangular
Main vacuum pumps (10^{-6} torr) [*]	72 Evaporion pumps	53 oil diffusion pumps, 21 getter-ion pumps	16 oil diffusion pumps	40 oil diffusion pumps	24 oil diffusion pumps
Interspace vacuum pumps (1-10^{-2} torr)			16 rotary pumps	16 rotary pumps	

[*] Not counting the roughing pumps.

sure, yet as thin as possible to economize on the expensive magnetic field region; no charge should accumulate on its surface; it must be safe against radiation damage, and it must not have outgassing properties that spoil the vacuum.

The problem of making a chamber wall of sufficient strength — of particular importance for large chambers — without losing too much magnetic volume has been solved in Nimrod and in the ZGS by using a double-walled chamber. The inner chamber (10^{-6} torr) has a thin wall inside the magnet gap. It does not have to stand much pressure as in the outer chamber the pressure is some 10^{-2} or 1 torr. The outer chamber is partly built around the magnet, it contains the magnet pole tips and some of the coils, and it derives its strength partly from the magnet. In Nimrod the inner chamber wall is made of an insulator, in the ZGS of panels of stainless steel, semi-insulated from each other.

In the PPA the eddy-current problem is particularly severe because of the high rate of change of the field. The chamber wall is made of laminated copper-nickel ribs in the shape of a "U", which are clamped to a stainless-steel plate outside the magnet gap (*fig. 35*); they are covered with epoxy and glass-fibre skins.

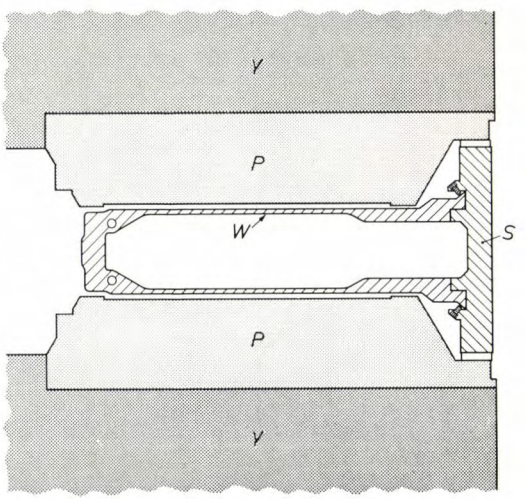


Fig. 35. Cross-section of the PPA vacuum chamber. The wall consists of copper-nickel ribs *W* clamped to a stainless steel plate *S* outside the magnet gap. The ribs are covered with epoxy and glass-fibre skins. They are laminated to prevent eddy currents in the wall. *P* pole tip, *Y* magnet yoke.

In the AGS and the CPS the vacuum chamber is elliptic in cross-section; in the AGS it is of Inconel, in the CPS of stainless steel. In the AGS the wall resistivity is marginal, and in the initial period of the cycle the field would be too low not to be seriously perturbed by eddy currents. An inductor is therefore included in the magnet circuit; this slows down the rate of change of B in the initial period but becomes saturated and hence ineffective at higher fields. This inductor is an impressive piece of equipment of no less than 200 tons.

Magnetic field cycle and magnet power supply

In all the machines except the PPA, the magnetic field is made to vary in accordance with a programme of the kind shown in fig. 9. In such a programme a large voltage must be applied to the magnet coils in the rise time to produce the desired $\dot{B}$. Only a small fraction of this voltage is necessary to maintain the peak current during the flat-top period. During this period a large magnetic energy is stored in the magnetic field; during the low-field period this energy is stored as mechanical energy in one or more flywheels (in the CPS the rotor of the alternator to be mentioned in a moment functions as flywheel).

The kind of arrangement used to establish the variation of magnetic field is schematically indicated in *fig. 36*. During the rise time the flywheel drives an alternator, whose output after rectification is applied to the magnet. To remove the field the rectifier system is switched into inversion and the alternator, now acting as a motor driven by the energy stored by the magnet, returns the energy to the flywheel. The energy loss per

[8] For the AGS and the CPS new injectors are planned to replace the present 50 MeV injectors (Table XI): an 800 MeV booster synchrotron for the CPS and a 200 MeV linear accelerator for the AGS, in order to reach the 10^{13} protons per pulse region (see Table II).

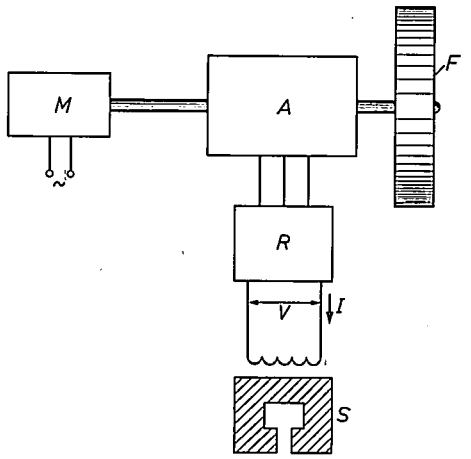


Fig. 36. Magnet current supply, schematic (AGS, CPS, ZGS Nimrod). During the rise time the flywheel F drives the alternator A, whose output is rectified by the rectifier R, giving a d.c. current in the coils of the synchrotron magnets S. To remove the field the rectifier set is switched into inversion, and A acts as a motor driven by the magnet current and accelerating the flywheel. Energy loss is made up for by the motor M.

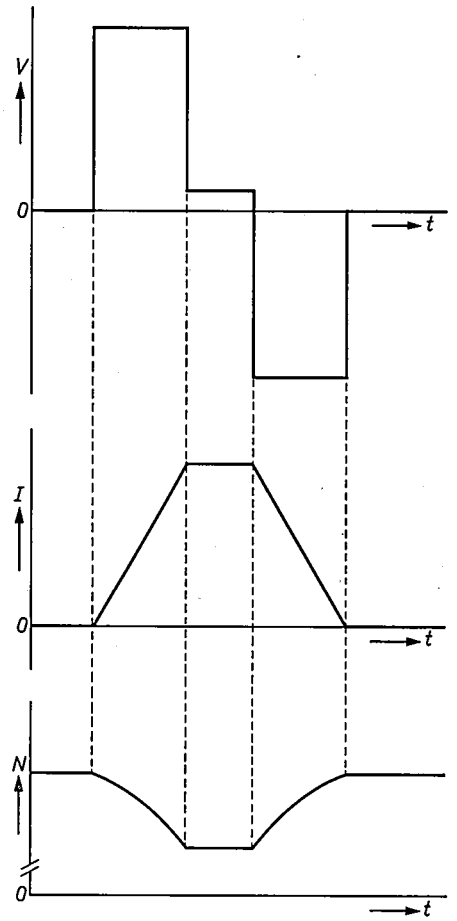


Fig. 37. Variation of the voltage V across the coils, the current I through the coils and the rotational speed N of the flywheel as a function of time (AGS, CPS, ZGS, Nimrod).

cycle is nevertheless a considerable part of the peak magnetic energy, and is made up for by a motor driven by the mains. The energy loss is mainly Joule heat developed in the coils, which is carried away by cooling water. The variation of voltage across the coils, the current in the coils and the rotational speed of the flywheel are shown very schematically in *fig. 37*. The variation in speed of the flywheel is usually less than 10%. Typical values relating to the magnetic field cycle and power supply are shown in *Tables IX* and *X*.

In the PPA the variation of the magnetic field and the means of obtaining it are completely different. The method described above would be impracticable for performing 20 magnet cycles per second. Instead, the magnet coils are built together with capacitors into a resonant system resonating at 19.5 Hz providing a sinusoidally varying magnet current. By adding a d.c. bias current nearly equal to the a.c. amplitude, a total current is obtained varying between nearly zero and some maximum value. The resulting field variation is shown in *fig. 38*.

The PPA magnet resonant power supply, as it would be for a four-section magnet, is shown schematically in *fig. 39*; in reality the system consists of 16 sections. The system must contain a choke short-circuiting the capacitors for the d.c. component but offering sufficient impedance to the a.c. component (L_2 in fig. 39). A large 130-ton air-core reactor, subdivided into 16 sections, is used for this. Together with some primary windings, it also constitutes the transformer for generating the a.c. current. At low field the energy is stored partly in the capacitors and partly in the air-core reactor. The capacitor bank and the reactor, built on top of the main building, are shown in *fig. 40* (see also fig. 2). The Q factor of the resonant system is 70.

We note that in the PPA, which is of relatively modest dimensions, the peak magnetic energy stored is relatively small, but the peak power flowing into the

Table IX. Magnetic field cycle. The time values indicated are typical for acceleration to the final energy given.

		AGS	CPS	ZGS	Nimrod	PPA
Final energy	GeV	30	25	12.5	7	3
Period	s	2.4	2.5	2.3	2.7	0.054
Rise time	s	0.8	0.75	1	0.7	} biased sinusoidal
Flat top	s	0.5	0.5	0.3	0.55	} variation

Table X. Magnet power supply. Peak values refer to a final field of the indicated flux density (end of 1969).

		AGS	CPS	ZGS	Nimrod	PPA
Final-field flux density	teslas	1.3	1.4	2.15	1.4	1.38
Peak magnetic energy stored	MJ	14	15	47.7	40	1.5
Power input peak	MW	30	32	110	160	
Power input mean	MW	2.4	1.6	10	2.5	2.5
Energy during low-field period stored in		flywheel 47 t 850 rev/min	flywheel [*] 1000 rev/min	flywheel 90 t	2 flywheels 40 t each 930 rev/min	capacitor + air-core reactor 87000 kvar at 19 Hz
Peak current	A	7000	6400	11000	8600	697 (a.c.) + 715 (d.c.)
Peak voltage	kV	6	7.6	9.5	15.5	7.5

[*] The alternator rotor functions as flywheel.

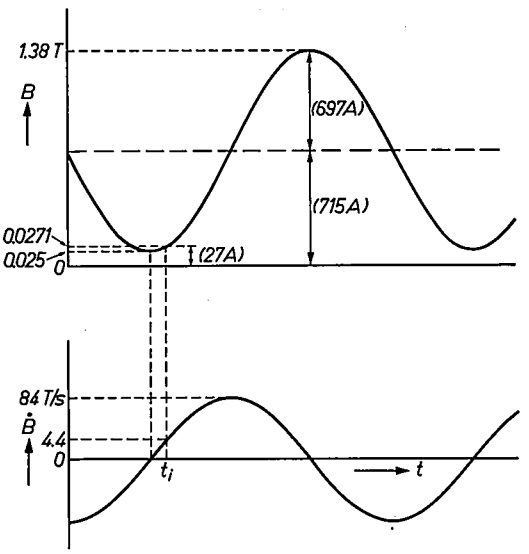


Fig. 38. Variation of B and $\dot{B}$ in the PPA, schematic. t_1 instant of injection.

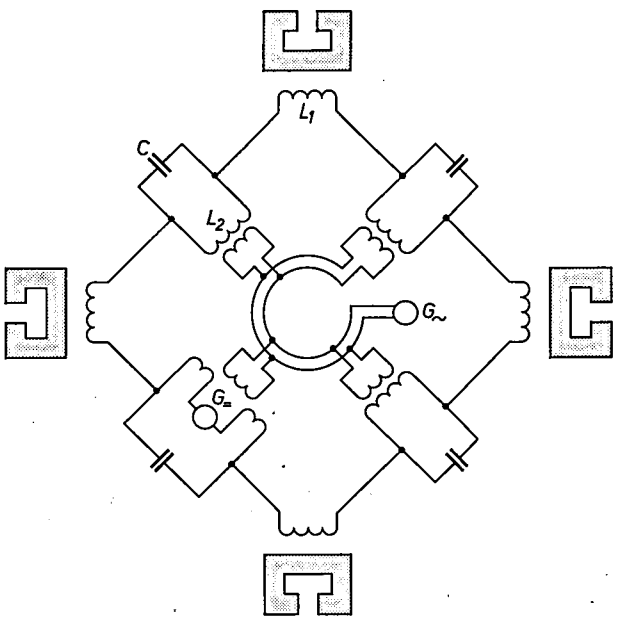


Fig. 39. Magnet resonant power supply of the PPA, drawn for a four-section magnet. In reality there are 16 sections. L_1 magnet coil. L_2 section of air-core reactor (d.c. choke). C capacitors. G_- d.c. generator. $G_\sim$ a.c. generator.

magnet and the average power dissipated are comparable to those of the larger machines; this again is due to its high repetition rate. An important part of the dissipated energy is hysteresis loss in the magnet steel; the heat is carried off by air flowing through slots in the steel.

Beam injection and r.f. trapping

A few data concerning beam injection have been summarized in *Table XI*. Injection systems have been shown earlier (for Nimrod in fig. 33*b*; for the AGS in fig. 2 of the preceding article).

The number of particles accelerated per second is in principle given by the brightness of the preaccelerator, the "potential beam intensity" of the synchrotron as discussed earlier (see p. 357), the efficiency of the injection procedure for filling the machine acceptance, and the r.f. trapping efficiency. Improving the preaccelerator brightness is a matter to which much attention is paid [8], but we shall not deal with this here.

As to the *injection efficiency*, a first point to remember is that the injection energy should be large enough to make the space-charge effects (p. 347) negligible. We note that the ZGS and Nimrod (both designed for high intensities) have relatively high injection fields and energies (see also fig. 30), and this is in fact mainly in connection with space charge. In the AGS and the CPS the injection magnetic field could be chosen not too far from the limit imposed by remanence and stray fields [9]. In the PPA it is also determined by the $\dot{B}$ required (see below).

The number of injected turns should be roughly equal to the ratio of the acceptance to the injected-beam emittance. As the emittances at injection are of the same order of magnitude the number of turns varies (very roughly) as the theoretical acceptance (Table VII). In the ZGS and Nimrod multiturn injection is applied with the aid of betatron oscillations (see fig. 24). In the PPA multiturn injection is used without betatron

[9] When the new injectors (see footnote [8]) are used, B_1 will increase correspondingly.

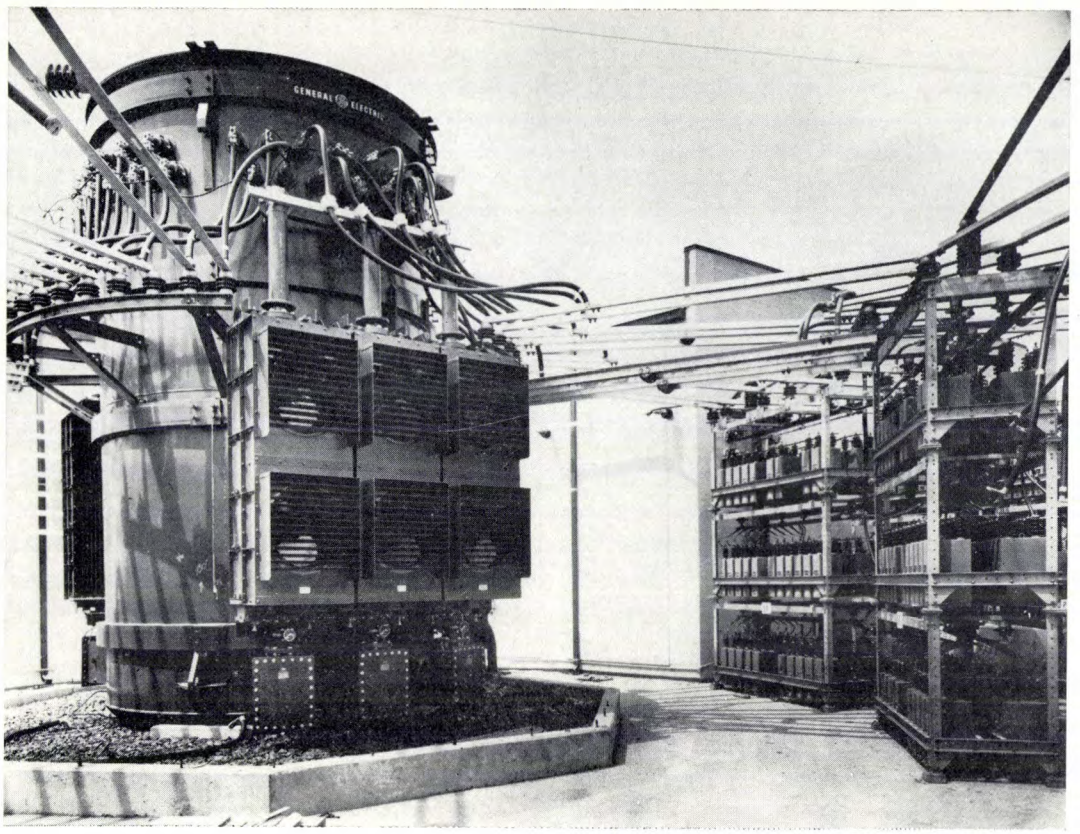


Photo Princeton-Penn Accelerator

Fig. 40. The capacitor bank and air-core reactor to resonate with the PPA magnet, on top of the PPA building.

oscillations, injecting *in* the equilibrium orbit. When this method is used the rate of orbit shrinkage must be sufficient, requiring (see eq. 30) a relatively high value of $\dot{B}_i$, and the injected-beam energy must be gradually modulated upwards during injection (in this case by a total of 35 keV).

Originally it was thought doubtful whether multiturn injection could be applied in an AG machine, and the CG principle was adhered to if high intensity was the primary purpose. Meanwhile multiturn injection has proved to be possible in AG machines too (p. 349); this is of importance when the injected-beam emittance

Table XI. Injection.

		AGS	CPS	ZGS	Nimrod	PPA
Injection kinetic energy (T_i)	MeV	50	50	50	15	3
Injection field (B_i)	milliteslas	12.1	14.7	48.2	29.9	27.0
$\beta_i = v_i/c$		0.31	0.31	0.31	0.175	0.081
$f_i = \beta_i f_c$	MHz	0.125	0.145	0.55	0.354	0.31
Injection period (number of turns)		10	1 to 3 [*]	60	120	9
$\dot{B}$ at injection ($\dot{B}_i$)	teslas per s	0.3	1.4 [*]	2.2	0.8	4.4
Normal $\dot{B}$ during acceleration	teslas per s	1.5	1.7	2.2	2.0	84 (max)
Radius shrinkage per turn (ΔR)	mm			3.2	3.6	10
Proton source [**]		D	D	D	RF	PIG
Voltage of high-voltage generator	kV	750	500	750	600	—
Preaccelerator type [***]		L	L	L	L	G
Preaccelerator frequency	MHz	202	202	202	115	
Emittance (horiz. and vert.) [****]	mm × mrad	10 π	30 π	28 π	50 π	2 π

[*] $\dot{B}_i$ can be reduced to 0.35 teslas per second. For the time being it is difficult to say which injection scheme is the most efficient.
[**] D = Duoplasmatron (Von Ardenne source), RF = radio-frequency source, PIG = Philips ionization gauge.
[***] L = linear accelerator, G = Van de Graaff accelerator.
[****] The beam emittance indicated is the area in phase space containing 95% of the beam.

is small compared with the chamber acceptance. In the AGS a 10-turn injection is regularly used; this, coupled with an increased source brightness, gives a gain in intensity of a factor of about five compared with the original single-turn injection. In the CPS a 3-turn injection has been tested succesfully but is not yet operational; a gain of 1.6 to 2 compared with single-turn injection is expected.

The *r.f. trapping efficiency* is proportional to the r.f. bucket size at injection. As discussed earlier, a small $\dot{B}/V_m$ at injection is favourable for bucket size. In Table XI it is seen that $\dot{B}_i$ is considerably smaller than the normal $\dot{B}$ in the AGS, in Nimrod, in the PPA and also in the multiturn tests in the CPS, leading to a bucket width of nearly 2π (instead of π for a normal synchronous phase angle of 60°). In the PPA this is a natural consequence of the biased sinusoidal variation of B (fig. 38). In the AGS a small $\dot{B}_i$ is caused by the saturating inductor that was originally incorporated to reduce the eddy currents in the vacuum-chamber walls at injection (p. 359); the advantage for r.f. trapping was discovered later. In the CPS and in Nimrod the reduction of $\dot{B}$ at injection is obtained by reducing the voltage applied to the magnet. The r.f. trapping efficiency can be increased further by applying the method illustrated in fig. 27 (p. 350). This method is often used, but finding the best solution is often still a matter of trial and error.

Acceleration and r.f. stations

Some details of the acceleration and the r.f. stations are given in *Table XII*. Given the magnetic-field cycle, the $\dot{B}$ during acceleration can be determined at any moment and eq. (22) then gives the required energy gain per turn. The election of the synchronous phase angle φ_0 and the number of gaps per turn M then establishes the peak voltage per gap V_m.

Except in the PPA, the accelerating cavities are all simple variants of the type indicated in fig. 5 of the preceding article (page 315) and they have been shown in some more detail in fig. 7 (the same article, page 316). Some more data are given in Table XII.

Table XII. Acceleration and r.f. stations (late 1969).

		AGS	CPS	ZGS	Nimrod	PPA
$\dot{B}$ during acceleration [*]	teslas per s	1.5	1.7	2.2	2.0	84 (max)
Energy gain per turn ($MeV_m \cos \varphi_0$) [*]	keV	100	70	10	5.5	61 (max)
Synchronous phase angle (φ_0) [*]		± 60°	± 60°	60°	60°	60°
Number of cavities		12	14	1	1	4
Number of gaps per cavity		2	1	3	2	2 [**]
Number of gaps per turn (M)		24	14	3	2	8 [**]
Gap peak voltage (V_m)	kV	8	10	8.0	5	5.5 (max)
Total r.f. power input peak	kW	300	70	60	45	320
Total r.f. power input mean	kW		60	30	20	80
Cavity length	m	3	2.3	2.03	2.1	1.5
Gap length	cm	3	3	7.6	7.6	7.5
Gap insulation		ceramic	glass	ceramic	polythene	alumina
Orbital frequency at injection (f_i)	MHz	0.117	0.145	0.55	0.354	0.31
Final orbital frequency (f_f)	MHz	0.372	0.478	1.75	1.995	3.75
Harmonic number ($h = f_a/f$)		12	20	8	4	8
Accelerating voltage frequency at injection ($f_{ai} = hf_i$)	MHz	1.40	2.90	4.4	1.42	2.5
Final accelerating voltage frequency ($f_{af} = hf_f$)	MHz	4.46	9.56	14.0	7.98	30
Frequency swing (f_{af}/f_{ai})		3.2	3.3	3.2	5.6	12
R.f. control		bootstrapping, phase lock	phase lock	phase lock	phase lock	phase lock

[*] $\dot{B}$ varies slightly, and consequently the energy gain per turn and the synchronous phase angle also vary.
[**] In the PPA there are in fact four gaps per r.f. station, but only two operate at a time; see fig. 7 of the preceding article.

The r.f. system of the PPA is considerably more complex than those of the other machines since this machine has the special features of a rapidly and sinusoidally varying magnetic field and a low injection energy. In the first place, since its maximum $\dot{B}$ is much larger than that of the other machines (see Table XII), the energy gain required per turn is comparable with that of the largest machines, even though its radius is small. This is why a relatively high proportion of the large straight sections (4 out of 8) are occupied by r.f. stations. Secondly, the frequency swing required is very high because of the low injection energy (see also fig. 30), and the customary method of tuning the cavities is not feasible. The acceleration is achieved in two steps, with a drift tube for the first frequency region and a cavity

for the second. This complicated acceleration station has also been shown in fig. 7 of the preceding article. Finally, as $\dot{B}$ varies rapidly during acceleration, the gap-voltage amplitude must vary rapidly (see eq. 22) if the synchronous phase angle is to keep its normal value of 60° during acceleration.

The harmonic number h and the number of cavities M

The choice of the harmonic number $h = f_a/f$ and of the number of cavities M is a delicate compromise as there are arguments in favour of both large as well as small values. M and h are closely connected. As long as the cavities are kept in phase with each other [10], h must be equal to or larger than M. If h is larger than M, the journey from cavity to cavity is not isochronous, and this will be a source of synchrotron oscillations. From this point of view $h = M$ is the best choice. Small M (and thus small h) is favoured by the consideration that each r.f. station is an expensive unit taking valuable straight-section space; large M (i.e. large h), on the other hand, reduces the total r.f. power required (see p. 337). Other arguments in favour of *small h* values are that a) the r.f. bucket at injection will be large, giving a large r.f. trapping efficiency (page 346); b) a large distance between the bunches facilitates beam ejection. A *large h* value may be favoured for the following (further) reasons: a) the frequency and thus the "restoring forces" of the synchrotron oscillations are large (eq. 28), leading to small synchrotron oscillation amplitudes; b) the cavities can be made small, the length being inversely proportional to f_a; c) a large number of bunches approximates to a homogeneous distribution of particles around the ring, which for certain reasons may be considered desirable. Table XII shows that the value chosen for h in the five machines varies from 4 (at Nimrod) to 20 (at the CPS).

Final beam dimensions

A high intensity of the accelerated beam is one of the main goals of an accelerator. It may however also be of importance to have the accelerated particles concentrated in a *narrow* beam. When the beam is intercepted by a target inside the vacuum chamber, the experimental efficiency does not depend too much on the beam size since in a large number of turns the beam is completely "peeled off" by the target. When the beam is ejected, however, it strikes an external target only once over its full area, and in this case it is desirable to have many particles per cm², i.e. a narrow beam. (For this reason it is better to increase intensity by increasing the brightness of the injected beam than by increasing the acceptance of the machine.)

If we neglect a number of disturbing effects (beam expansion at transition, instabilities, field distortion, non-constancy of n, etc.), the final beam size should be determined by the vacuum-chamber aperture and the amount of shrinkage during acceleration (eq. 31). The shrinkage factor for width is $\sqrt{B_i/B_f}$. The flux density of the final magnetic field B_f has a fairly well defined value of 1.4 to 2.0 teslas (determined by saturation). The flux density at injection B_i, on the other hand, is determined by many considerations, and varies considerably. $\sqrt{B_i/B_f}$ is about 1/10 for the AGS and the CPS, and 1/7 for the ZGS, Nimrod and the PPA. In fact, the real final beam dimensions are not very well known.

Concluding remarks

In this article we have reviewed five proton synchrotrons, ranging from 3 to 30 GeV. The two largest of these, the CERN Proton Synchrotron at Geneva and the Alternating Gradient Synchrotron at Brookhaven, have for a long time constituted the frontier of particle energies. Since 1967 they have been surpassed by the 80 GeV proton synchrotron in Serpukhov in the USSR.

Nuclear physicists continue to be very interested in much higher particle energies and it appears that the AG proton synchrotron is still the most promising type of machine for the 300 GeV range. There seems to be no reason why the design of the existing AG proton synchrotrons should not be used, with a particle orbit ten times as large, to reach an energy region ten times greater. The cost would only be an approximately linear function of the desired particle energy. (With a classical cyclotron the cost would have been proportional to something like the cube of the energy.)

In fact, the building of accelerators in the 300 GeV range is in an advanced state of preparation. Barring unexpected events, the plans for a 200-400 GeV AG synchrotron in Battavia near Chicago will be carried through. This will be a separated-function machine: the bending will be established by flat-pole bending magnets and the focusing by quadrupole lenses inserted between the magnets. Very strong focusing can thus be obtained using strong gradients without affecting the useful aperture (p. 356). The European project for the next big machine is a 300 GeV synchrotron with a diameter of 2.4 km. The design has been worked out in detail at CERN in Geneva, and several possible locations in Europe have been studied carefully (e.g. for soil solidity and stability, but a decision on whether it will actually be built has not been taken yet (December 1969).

Another line of development, making available even

[10] In the AGS six cavities are in phase, and the remaining six are in antiphase. In the CPS the phase of half of the cavities can be varied with respect to the others.

[11] See for instance G. K. O'Neill, Storage rings, Science **141**, 679-686, 1963.

higher "effective" collision energies, is that of the "storage rings", but we shall have to leave this aside here [11].

The five machines discussed are representative for the range of energies covered by proton synchrotrons up till the present, and also of the techniques used in this kind of machine. These particular machines have been chosen for discussion because Philips supplied the ferrites for reducing the accelerating cavity dimensions and for tuning the cavities to the changing orbital frequency of the accelerated particles, as discussed in the preceding article. In the next generation of machines, for energies of 100 GeV and more, ferrite will not be required for tuning the cavities: since the particles will be injected with a velocity very close to c, the orbital frequency will hardly change, and, as in the classical cyclotron but for a completely different reason, tuning of the acceleration voltage is not necessary. However, it is quite likely that the next-generation machines will use a medium-energy booster synchrotron for the injector: this would have a frequency swing, and would require ferrite material for this purpose.

Literature

The reader may care to refer to the following books and articles about proton synchrotrons.

E. D. Courant and H. S. Snyder, Theory of the alternating-gradient synchrotron, Ann. Physics **3**, 1-48, 1958.
H. Bruck, Les accélérateurs circulaires de particules, Presses Universitaires de France, 1966.
K. G. Steffen, High energy beam optics, Interscience Publ., New York 1965.
M. S. Livingston and J. P. Blewett, Particle accelerators, McGraw-Hill, New York 1962.
R. Gouiran, Particles and accelerators, McGraw-Hill, New York 1967.
G. K. Green and E. D. Courant, The proton synchrotron, Handbuch der Physik, Part 44, 218-340, Springer, Berlin 1959.
J. P. Blewett, The proton synchrotron, Rep. Prog. Phys. **19**, 37-79, 1956.
Proceedings of the 1965 and 1967 Particle Accelerator Conferences in Washington D.C., U.S.A., IEEE Trans. **NS-12**, No. 3, 1965 and **NS-14**, No. 3, 1967.

Further information about a particular machine may often be obtained from the PS laboratories concerned in the form of internal reports on some particular subject. Articles on the AGS have been published in Science **128**, 1393-1401, 1958 and in IRE Int. Conv. Rec. **8**, Part 9, 3-27, 1960.

Summary. Philips have supplied the large amounts of ferrite material required in five major proton synchrotrons, which accelerate protons to energies in the range of 3 to 30 GeV. This prompted a more general article about all five of these machines, which, though basically of the same design, are all quite different in appearance and in detail. The comparisons allow the special features of each machine to be brought out clearly. The introductory first part gives a general description of a proton synchrotron. After reviewing the relation between the final value of the guiding field, the final proton energy and the radius of the ring, and discussing the synchronization of the accelerating voltage to the circulating proton bunches, the stability of the proton orbit and the accelerator focusing characteristics are discussed in detail. The (transverse) betatron oscillations of the protons around an equilibrium orbit are related to the transverse focusing, which is established either by a transverse gradient in the guiding field (in "combined-function machines") or by other fields (in "separated-function machines"). During acceleration the beam shrinks in width (i.e. the betatron amplitudes decrease). The synchrotron oscillations (oscillations in orbit, phase and energy around the "synchronous particle") are related to the longitudinal or phase focusing. The strong-focusing or AG accelerators (with large alternating gradients) differ from the weak-focusing or CG accelerators (with a small constant gradient) in that they have *fast* betatron oscillations of *small* amplitude, and a *small* spread in equilibrium orbits for a given spread in momentum (strong "momentum compaction"). They also have a "transition energy" at which the accelerating voltage must make a phase jump to maintain phase stability. In the (weak-focusing) zero-gradient (ZG) accelerators, focusing is obtained by shaping the magnet edges (separated functions). The protons are injected into the ring by making use of the shrinkage of the equilibrium orbit in the increasing magnetic field at constant energy, and also of the betatron oscillations around this orbit. Special measures are taken to ensure that a large fraction of the injected particles is captured by the accelerating field.

In the second part the special features of the five synchrotrons are discussed with the aid of many numerical data. 1) The Alternating Gradient Synchrotron at Brookhaven (33 GeV) and 2) the CERN Proton Synchrotron at Geneva (28 GeV) were the first proton accelerators to apply the AG principle, to carry particles through transition and to reach the 30 GeV range. 3) The Zero Gradient Synchrotron at Argonne near Chicago (12.7 GeV) is the largest machine with a zero gradient; it has a strong final field (2.1 teslas instead of 1.4 teslas elsewhere) and is relatively compact. 4) Nimrod, the 7 GeV synchrotron at Chilton (England), has a large vacuum chamber for super-high intensities. 5) The Princeton-Pennsylvania Accelerator at Princeton, N.J. (3 GeV) is the first fast-cycling high-energy proton accelerator, which can accelerate many protons per second although it is relatively small.

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

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.

G. A. Acket & J. J. Scheer: Relaxation oscillations due to impact ionisation in epitaxial sheet-type Gunn oscillators.
Electronics Letters **5**, 160-161, 1969 (No. 8). *E*

G. A. Acket & M. T. Vlaardingerbroek: Physical properties of transferred-electron and avalanche microwave devices.
Festkörperprobleme **9**, 280-299, 1969. *E*

G. Arlt & K. Reiber: Ultrasonic amplification in tellurium along the *z*-axis with longitudinal and transverse electric drift fields.
Solid State Comm. **7**, 1155-1157, 1969 (No. 16). *A*

J. Basterfield: The chemical polishing of yttrium iron garnet.
Brit. J. appl. Phys. (J. Physics D), ser. 2, **2**, 1159-1161, 1969 (No. 8). *M*

V. Belevitch & Y. Kamp: Theory of monolithic crystal filters using thickness-twist vibrations.
Philips Res. Repts. **24**, 331-369, 1969 (No. 4). *B*

G. Blasse & A. Bril: Luminescence of phosphors based on host lattices ABO_4 (A is Sc, In; B is P, V, Nb).
J. chem. Phys. **50**, 2974-2980, 1969 (No. 7). *E*

G. Blasse & A. Bril: Luminescence of lead tungstate and related centres.
Philips Res. Repts. **24**, 275-283, 1969 (No. 4). *E*

R. Bleekrode: A study of the spontaneous emission from CO_2-N_2-He-H_2 laser discharges: $C^3\pi_u$-$B^3\pi_g$ emission bands of N_2.
IEEE J. Quantum Electronics **QE-5**, 57-60, 1969 (No. 2). *E*

R. Bleekrode: Near-infra-red spontaneous emission spectra of high-power CO_2 lasers.
IEEE J. Quantum Electronics **QE-5**, 297-298, 1969 (No. 6). *E*

R. Boite (Faculté Polytechnique de Mons) **& J. P. V. Thiran:** Synthesis of filters with capacitances, switches, and regenerating devices.
IEEE Trans. **CT-15**, 447-454, 1968 (No. 4). *B*

P. F. Bongers, C. Haas, A. M. J. G. van Run & G. Zanmarchi: Magnetoresistance in chalcogenide spinels.
J. appl. Phys. **40**, 958-963, 1969 (No. 3). *E*

H. van den Boom: Millimeter-wave electron spin resonance spectrometer with homodyne detection.
Rev. sci. Instr. **40**, 550-555, 1969 (No. 4). *E*

J. van den Boomgaard: On the *P-T-x* phase diagram of the Mn-Te system.
Philips Res. Repts. **24**, 284-298, 1969 (No. 4). *E*

G. A. Bootsma: Calorimetric study of the adsorption of oxygen and gases H_xA on germanium and silicon.
Surface Sci. **15**, 340-344, 1969 (No. 2). *E*

A. Broese van Groenou, H. G. Beljers & F. C. M. Driessens: Anisotropy and ferromagnetic resonance in $NiFe_2O_4$ with [111] texture.
J. appl. Phys. **40**, 1424-1426, 1969 (No. 3). *E*

K. H. J. Buschow & A. S. van der Goot: The crystal structure of rare-earth cobalt compounds of the type R_3Co.
J. less-common Met. **18**, 309-311, 1969 (No. 3). *E*

H. B. G. Casimir: De natuurwetenschappelijke achtergrond van de moderne industrie.
Akademiedagen XX (Kon. Ned. Akad. Wetensch.), 7-17, 1969. *E*

D. M. Caughey: Computer simulation of u.h.f.-transistor small-signal behaviour at high frequencies.
Electronics Letters **5**, 165-167, 1969 (No. 8). *E*

W. A. Crossley, R. W. Cooper, J. L. Page & R. P. van Stapele: Faraday rotation in rare-earth iron garnets.
Phys. Rev. **181**, 896-904, 1969 (No. 2). *E, M*

L. Crousel & J. Neirynck: Polynomial Chebyshev approximations of the ideal filter.
IEEE Trans. **CT-15**, 307-315, 1968 (No. 4). *B*

P. R. Dijksterhuis & T. Verhey (N.V. Philips' Phonographische Industrie, Baarn): An electronic piano.
J. Audio Engng. Soc. **17**, 266-271, 1969 (No. 3).

E. S. Eilley, J. B. Hughes & N. D. Richards: Design of an 800 000 bit, 150 nsec magnetic thin film store.
Proc. Conf. on Computer science and technology, Manchester 1969, p. 124-132. *M*

D. G. J. Fanshawe: A computer analysis of the circle diagram of an ultrasonic transducer.
Proc. Joint Conf. on Industrial Ultrasonics, Loughborough 1969, p. 211-219. *M*

L. Fraiture: Limits on the load for bandpass two-ports with one dc transmission zero.
IEEE Trans. **CT-16**, 126-128, 1969 (No. 1). *B*

N. J. Freedman & L. K. Brundle: Magnetoelastic wave propagation in single-crystal lithium ferrite.
IEEE Trans. **MAG-5**, 132-136, 1969 (No. 2). *M*

S. Garbe: Photoemission from $GaAs_{1-x}P_x$, covered with low work function layers.
Phys. Stat. sol. **33**, K 87-91, 1969 (No. 2). *A*

R. Genève: Introduction à la thermographie médicale.
Acta electronica **12**, 7-19, 1969 (No. 1). *L*

A. H. Gomes de Mesquita & A. Bril: The afterglow of some old and new Ce^{3+}-activated phosphors.
J. Electrochem. Soc. **116**, 871, 1969 (No. 6). *E*

G. Groh: Multiple imaging by means of holograms of correlated objects.
Appl. Optics **8**, 967-969, 1969 (No. 5). *H*

G. J. van Gurp: Flux-flow noise in a type-I superconductor.
Phys. Rev. **178**, 650-656, 1969 (No. 2). *E*

W. van Haeringen & H. de Lang (Technical University, Delft): Role of linear phase anisotropy in a Zeeman laser.
Phys. Rev. **180**, 624-625, 1969 (No. 2). *E*

P. Hansen: Temperature dependence of ferromagnetic resonance linewidth in NiZn-ferrite.
Phys. Stat. sol. **33**, K 5-8, 1969 (No. 1). *H*

J. Hasker: A simple approximate method for calculating the space-charge-limited beam current of rotation-symmetrical electron guns.
Philips Res. Repts. **24**, 249-262, 1969 (No. 4). *E*

J. Hasker: Anomalous electron velocity distributions.
Philips Res. Repts. **24**, 263-274, 1969 (No. 4). *E*

E. P. Honig: Solid phase properties and colloid chemistry.
Trans. Faraday Soc. **65**, 2248-2259, 1969 (No. 560). *E*

F. N. Hooge & A. M. H. Hoppenbrouwers: Amplitude distribution of $1/f$ noise.
Physica **42**, 331-339, 1969 (No. 3). *E*

M. Jatteau: Thermographe fonctionnant à 4 ou 10 microns et permettant le tracé précis des isothermes.
Acta electronica **12**, 21-62, 1969 (No. 1). *L*

H. Jonker *et al.*: Physical development recording systems:
II. Mercury nucleation and the amalgam effect **(H. Jonker, C. J. Dippel, H. J. Houtman, A. Molenaar & E. J. Spiertz)**;
III. Physical development **(H. Jonker, A. Molenaar & C. J. Dippel)**;
IV. PD photoplating **(H. Jonker, C. J. G. F. Janssen, C. J. Dippel, Th. P. G. W. Thijssens & L. Postma)**.
Photogr. Sci. Engng. **13**, 33-37, 38-44, 45-49, 1969 (No. 2). *E*

G. Klein & K. E. Kuijk: Generating the sum or difference of two signal frequencies.
Electronic Engng. **41**, 378-380, 1969 (No. 493). *E*

P. N. Kuin: Activation analysis of the impurities in silicon carbide.
Mat. Res. Bull. **4**, S 273-283, 1969 (SiC issue). *E*

F. A. Lootsma: Hessian matrices of penalty functions for solving constrained-optimization problems.
Philips Res. Repts. **24**, 322-330, 1969 (No. 4). *E*

F. A. Lootsma: Synchronisatie van verkeerslichten met behulp van lineaire programmering.
Operationele Research in Nederland, publ. Het Spectrum, Utrecht 1969 (Marka No. 104), p. 133-148. *E*

M. H. van Maaren: On the superconductivity, carrier concentration and the ionic model of Sn_4P_3 and Sn_4As_3.
Physics Letters **29A**, 293-294, 1969 (No. 6). *E*

R. Memming: Mechanism of the electrochemical reduction of persulfates and hydrogen peroxide.
J. Electrochem. Soc. **116**, 785-790, 1969 (No. 6). *H*

N. St. J. Murphy, F. Berz & I. Flinn: Carrier mobility in silicon MOST's.
Solid-State Electronics **12**, 775-786, 1969 (No. 10). *M*

J. M. Noothoven van Goor & H. M. G. J. Trum: Distribution coefficient and acceptor valency of tin in bismuth.
J. Phys. Chem. Solids **30**, 1636-1638, 1969 (No. 6). *E*

R. C. Oldfield, D. J. Lishman & B. Gibson: Making and testing vacuum-deposited magnetic thin-film storage elements.
Proc. Conf. on Computer science and technology, Manchester 1969, p. 133-140. *M*

C. J. M. van Opdorp: Si-Ge isotype heterojunctions.
Thesis, Eindhoven 1969. *E*

G. H. Plantinga (Philips Semiconductor Development Laboratories, Nijmegen): Influence of dislocations on properties of shallow diffused transistors.
IEEE Trans. **ED-16**, 394-400, 1969 (No. 4).

J. Polman: Nonlinear effects in the positive column of a modulated gas discharge.
Physics of Fluids **12**, 1128-1129, 1969 (No. 5). *E*

A. Rabenau & H. Rau: Spectral reflectance measurements on ternary compounds of copper with group VI and VII elements.
Solid State Comm. **7**, 1281-1282, 1969 (No. 17). *A*

O. B. P. Rikkert de Koe & P. van der Wurf: On some extensions of Nyquist's telegraph transmission theory.
Proc. IEEE **57**, 701-702, 1969 (No. 4). *E*

P. Schnabel: Frequency equations for *n* mechanically coupled piezoelectric resonators.
Acustica **21**, 351-357, 1969 (No. 6). *A*

G. J. Scholten: A novel type of bidirectional helix seal, the R-L seal.
Preprint Proc. 4th Int. Conf. on fluid sealing, Philadelphia 1969, 6 pp. *E*

E. Schwartz: Zur Theorie der Anpassung mit zwei Reaktanzen.
Archiv elektr. Übertr. **23**, 169-176, 1969 (No. 4). *A*

A. Stegherr: Das Dreistoffsystem Zinn-Antimon-Tellur.
Thesis, Aachen 1969. *A*

P. J. Strijkert: The effect of 5-fluorouracil on induced enzyme synthesis in yeast.
Biochim. biophys. Acta **182**, 262-263, 1969 (No. 1). *E*

J. S. Sussenbach & P. J. Strijkert: Arginine metabolism in Chlamydomonas reinhardi. Arginine deiminase: the first enzyme of the catabolic pathway.
FEBS Letters **3**, 166-168, 1969 (No. 3). *E*

C. van Trigt: Analytically solvable problems in radiative transfer, I.
Phys. Rev. **181**, 97-114, 1969 (No. 1). *E*

Q. H. F. Vrehen: Absorption and dispersion in porous and anisotropic polycrystalline ferrites at microwave frequencies.
J. appl. Phys. **40**, 1849-1860, 1969 (No. 4). *E*

W. L. Wanmaker & J. W. ter Vrugt (Philips Lighting Division, Eindhoven): Luminescence of gallates.
J. Electrochem. Soc. **116**, 871-872, 1969 (No. 6).

K. Weiss: Untersuchungen an kubischem Kupfer(I)sulfid (Digenit), I. Fehlordnung in Digenit, II. Nichtstationäre elektrische Flüsse.
Berichte Bunsenges. phys. Chemie **73**, 338-344, 344-348, 1969 (No. 4). *E*

R. M. G. Wijnhoven: Attenuationless voltage or current propagation on a line with ohmic losses.
IEEE Trans. **C-18**, 280-281, 1969 (No. 3). *E*

W. J. Witteman: Mode competition in lasers with homogeneous line broadening.
IEEE J. Quantum Electronics **QE-5**, 92-97, 1969 (No. 2). *E*

S. Yoshioka (Research Laboratory Matsushita Electronics Corp., Takatsuki, Osaka, Japan): Investigation of the chemical properties of stain films on silicon by means of infra-red spectroscopy and omegatron mass analysis.
Philips Res. Repts. **24**, 299-321, 1969 (No. 4).

G. Zanmarchi & P. F. Bongers: Infra-red Faraday rotation in ferrites.
J. appl. Phys. **40**, 1230-1231, 1969 (No. 3). *E*

L. E. Zegers: Common bandwidth transmission of information signals and pseudonoise synchronization waveforms.
IEEE Trans. **COM-16**, 796-807, 1968 (No. 6). *E*

H. Zijlstra & F. F. Westendorp: Influence of hydrogen on the magnetic properties of $SmCo_5$.
Solid State Comm. **7**, 857-859, 1969 (No. 12). *E*

Contents of Mullard Technical Communications **11**, No. 101, 1969:

J. C. Beckley: Integrated sound intercarrier amplifier and transformerless audio stage for colour tv receiver (p. 2-5).
B. J. M. Overgoor: Matched field effect transistors in the differential input stage of an operational amplifier (p. 6-16).
D. R. Hyde: Economical 3 W record-player amplifier (p. 17-20).
A. Garskamp: Portable a.m./f.m. radio receiver (p. 21-24).

Contents of Mullard Technical Communications **11**, No. 102, 1969:

D. J. King: Monochrome tv power supplies for 117 and 220 V mains inputs (p. 26-28).
D. J. King: Thyristor power supplies for colour tv receivers (p. 29-34).
M. C. Gander & L. Hampson: Luminance, colour-difference output, synchronising and a.g.c. circuits (p. 35-45).
S. W. Gray: Thermostatic fan control for central heating (p. 46-48).

Subject index, Volumes 21-30

Figures in bold type indicate the volume number, and those in ordinary type the page number. Subjects dealt with in volumes **1-20** are listed in the indexes at the end of volumes **10, 15, 20** and **25.**

Author index, Volumes 21-30

Figures in bold type indicate the volume number, and those in ordinary type the page number. Articles published in volumes **1-20** are given in the author indexes at the end of volumes **10, 15, 20** and **25.**

Volume 30, 1969, No. 11/12 pages 309-384 Published 27th February 1970