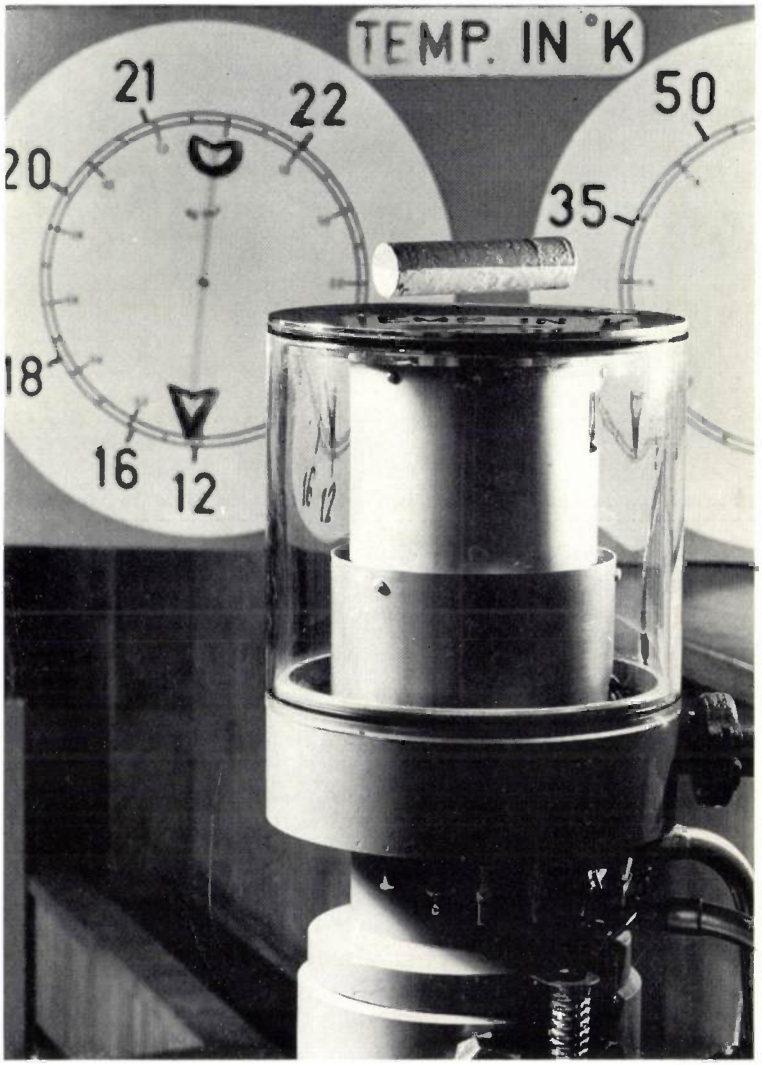
PHILIPS TECHNICAL REVIEW

VOLUME 26, 1965, No. 1

Scientific research and technical development work at temperatures around the boiling points of hydrogen and helium — e.g. in some branches of solid state research, the development of masers, certain aspects of computer component research, etc.—are still restricted to a considerable extent by the size and complexity of the machines required to produce these extremely low temperatures. Consequently this work has until now been mainly confined to specially equipped cryogenic laboratories. The appearance of a small, easily operated gas refrigerating machine that can work with good efficiency at 20 °K, can even reach temperatures in the region of 12 °K, and can produce about 5 litres of liquid hydrogen per hour, will therefore undoubtedly be welcomed by many research workers.



A gas refrigerating machine for temperatures down to 20 °K and lower

G. Prast

621.573

For some considerable time now, Philips Research Laboratories have been investigating the possibilities which the Stirling cycle offers as a refrigeration process. It soon became clear that this process has very attractive properties, particularly when the temperature at which the "cold" must be supplied is substantially lower than that of the environment (i.e. as a rule room temperature). This led to the construction of a very compact cryogenerator — the well-known *gas refrigerating machine* [1] — which can work with good efficiency in

Ir. G. Prast is a research worker at the Philips Research Laboratories, Eindhoven.

[1] See J. W. L. Köhler and C. O. Jonkers, Fundamentals of the gas refrigerating machine, Philips tech. Rev. **16**, 69-78, 1954/55, and J. W. L. Köhler and C.O. Jonkers, Construction of a gas refrigerating machine, Philips tech. Rev. **16**, 105-115, 1954/55.

the temperature range from —100 °C to —200 °C (173 °K to 73 °K). This range includes the boiling points of oxygen and nitrogen. With this machine — of which there is now a four-cylinder version in addition to the original single-cylinder type — installations have been made for producing liquid air and for the separation of nitrogen from air, which are now in fairly wide use. The gas refrigerating machine has also found numerous other applications [2].

The following are the principal advantages of the gas refrigerating machine over other types:

1) The machine operates with a fixed, small quantity of gas, either hydrogen or helium, which is contained in an enclosed space where it cannot be contaminated.
2) The machine has no valves.
3) Considering its large refrigerating capacity, the dimensions of the machine are very small.
4) The machine has a high efficiency.
5) The machine requires hardly any attendance or maintenance.

After this initial success, one of the main objects of research was to discover whether the Stirling cycle could be used to attain much lower temperatures, e.g. 10 °K, still with a good efficiency. The temperature range of a machine capable of this would then also contain the boiling points of neon (27 °K) and hydrogen (20.4 °K) — that of helium lies at 4.2 °K. This continued research proved successful and indeed led to the construction of a machine of this kind. This machine, which was designed for 20 °K but is capable of operating with good efficiency in the region between 14 °K and 60 °K, is the subject of the present article. The new machine is not simply an improved version of the familiar gas refrigerating machine, but operates on a principle that may be described as a modification of the Stirling cycle. It retains, however, all the merits of the original gas refrigerating machine as summarized above.

It may be recalled that in a machine based on the "normal" Stirling cycle, the compartment containing the working gas consists of two spaces in open communication with each other via a regenerator (*fig. 1a*). One part is at the temperature T_E at which the cold is produced, and the other is at the ambient temperature T_C. The working gas is compressed in the latter space — the heat of compression being removed by cooling water in the "cooler" — after which it passes through the regenerator where it is cooled to T_E and then expanded in the other space. In this expansion the cold is generated and supplied to the "freezer". The gas then returns to the first space — the compression space — and in passing through the regenerator it is again heated to T_C, by reabsorbing the heat it left behind in its previous passage through the regenerator. The lowest temperature that can be reached is determined by the losses, which cannot of course be entirely reduced to zero.

In the final form of the gas refrigerating machine there are not *two* pistons but only one, and the gas is transported from the one space to the other by what is termed the displacer [1]. In this case the regenerator, like the cooler and the freezer, is annular in shape (fig. 1*c*).

The modified cycle in the new machine consists essentially of two such cycles combined, whereby one cycle compensates partly for the losses of the other, so that a lower temperature can be reached than is possible with a single cycle [3]. *Fig. 2* gives a schematic representation of this combined process. As can be seen, there are now *three* spaces: one compression space and *two* expansion spaces; there are also *two* regenerators. Upon expulsion from the compression space, part of the gas goes to the first expansion space and another

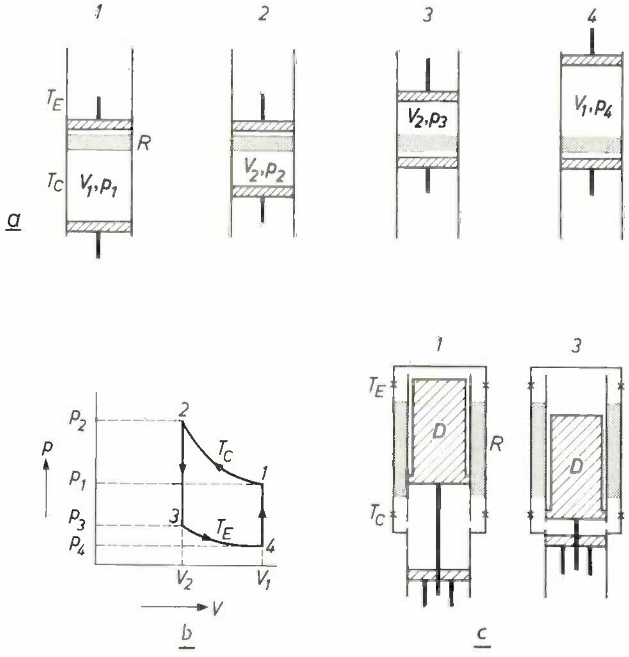


Fig. 1. *a*) Principle underlying the use of the Stirling cycle for refrigeration. Two pistons move in a cylinder which contains one space at room temperature T_C and one at low temperature T_E, and between them a regenerator *R*. The regenerator is a space filled with a mass of finely divided metal (e.g. thin wires). The filling has a high heat capacity and excellent heat-transfer properties. The cycle comprises four phases, whose initial states (1-4) are shown in the diagram.
b) Variation of the pressure and volume of the working gas. The phases 1 and 2 occur isothermally, phases 3 and 4 are isochoric.
c) Diagram of the actual construction of a conventional gas refrigerating machine. There is only one piston, and the cylinder space is divided into two by a displacer *D*. The regenerator is annular in shape. The crosses on the walls indicate schematically the location of the cooler (below) and the freezer (above) which are also annular. Piston and displacer are driven by a crankshaft and have an almost harmonic motion. Consequently the gas cycle is not exactly identical with that shown in fig. 1*b*.

part simultaneously to the second expansion space (the splitting into two fractions is represented schematically by the dashed line). The latter part passes through both regenerators. The spaces a_1 and b, together with the part of the regenerator R_1 situated below the dashed line (yellow), can as a whole be regarded as the space in which the one part of the process takes place, and the spaces a_2 and c together with R_2 and the section of R_1 above the dashed line (red) are where the other part of the process occurs. The first part of the process is a normal Stirling cycle, which supplies cold at the temperature T_{m} of the middle space. This cold is not removed but compensates the losses of R_1.

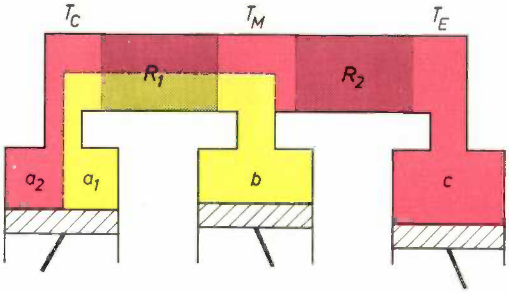


Fig. 2. Diagram of a machine with *two* expansion spaces (three-space machine). The process here can be represented as divided into two normal Stirling cycles; the first occurs in the yellow part of the space and the other in the red part. In principle the pistons in the two expansion spaces b and c need not move in phase; in our machine they do, however. The cold supplied by the first process is entirely or partly used to compensate the losses of R_1.

The lowest temperature we have hitherto been able to reach with the machine described in this article (designed for 20 °K) is 10.5 °K. Obviously, the only eligible working gas is helium. With a machine specially designed for the purpose it would certainly be possible to reach an *even lower* temperature. On the other hand, it is evident that the boiling point of helium cannot be reached directly with such a machine, simply because no gas exists that has a lower boiling point under atmospheric pressure.

The fact that the machine can attain the extremely low temperatures mentioned above may be elegantly demonstrated by a disc of niobium-tin on the head of the machine. At a temperature lower than 18 °K, NbSn is a superconductor, and the superconductivity is demonstrated, as shown in the title photograph, by the magnet floating over the disc [4]. For the purpose of the experiment, the normal insulation space of the head was replaced by one in which the distance between the NbSn disc and the outer wall was only 1 millimetre.

The new machine was designed entirely on a theoretical basis. Our starting point was the theory of the normal gas refrigerating machine, as earlier developed by us. Although this theory takes account of all important effects — such as the various kinds of losses, imperfect heat transfer in the regenerator, the non-ideal behaviour of the working gas, the almost harmonic motion of the piston, etc. — it is simple enough for practical application. The results have been found in all cases to be within 10% of measurements on the actual models. This theory was now extended to the case of a machine with three spaces. The extended theory is of course more complicated, and in order to use it for designing a machine it is necessary to resort to a computer. Once the required programmes are available, however, the design of a three-space machine does not take much more time. We shall return to this point later. A discussion of the theory itself does not enter into the scope of this article.

In the following we shall give a somewhat more detailed description of the modified Stirling cycle on which the operation of the new machine is based, and briefly consider the difference between the use of such a machine as a "pure refrigerator" and as a liquifier; furthermore we shall give a description of the machine and compare its performance with that of other installations, and finally review the principal possibilities for application. We shall begin by considering the normal Stirling cycle at somewhat greater length, with particular reference to the lowest temperature that can be reached with it.

The minimum value of T_{E} for the normal Stirling cycle

To determine the lowest temperature that can in practice be reached using a normal Stirling cycle, we have to find how the (gross) cold production and the cold losses vary with T_{E}. The lowest attainable value of T_{E} is of course that at which the losses are equal to the cold production. These cold losses include those due to regeneration, conduction and insulation; the regeneration loss is the most important of the three. We can do the calculation by representing the process in a highly schematic form. We assume that a given mass m of gas is isothermally compressed at room temperature to a pressure p_1, then isobarically cooled in a regenerator, and finally expanded again isothermally at low temperature (T_{E}) from the pressure p_1 to a pressure p_2. In this process the gross cold production Q_{E} is given by:

$$Q_{\mathrm{E}} = \int_{p_1}^{p_2} p\mathrm{d}V = mRT_{\mathrm{E}} \ln p_1/p_2 . \quad (1)$$

[2] For a concise review, discussing the construction and various applications of the gas refrigerating machine, see J. W. L. Köhler, Progress in Cryogenics **2**, 41, 1960. An air fractionating plant recently put into operation and using the four-cylinder machine was shown in Philips tech. Rev. **25**, 340, 1963/64 (No. 11/12).

[3] See also G. Prast, Cryogenics **3**, 156, 1963. The idea of modifying the Stirling process in this way is due to H. Fokker and J. W. L. Köhler, both of Philips Research Laboratories (see e.g. British patent No. 749 815).

[4] See for example A. H. Boerdijk, Levitation by static magnetic fields, Philips tech. Rev. **18**, 125-127, 1956/57.

The gross cold production is thus proportional to the (absolute) expansion temperature T_E and is therefore smaller the lower the value of T_E.

The amount of heat which the regenerator must extract from this quantity of gas in order to cool the gas from T_C to T_E is $mc_p(T_C - T_E)$. In practice this does not happen completely; the regenerator extracts only a part η_r, and there is thus a cold loss ΔQ_R, called the regeneration loss, whose magnitude is given by:

$$\Delta Q_R = (1 - \eta_r) mc_p(T_C - T_E) \; . \; . \; . \; . \; . \quad (2)$$

The quantity η_r is called the regenerator efficiency. In practice η_r is in the region of 0.99, so that the loss is roughly 1%. Since η_r and c_p are nearly independent of temperature, the regeneration loss is proportional to $(T_C - T_E)$. For $T_E = T_C$ the loss is zero and becomes greater the lower the value of T_E. *Fig. 3* presents this variation of Q_E and ΔQ_R in the form of a graph. The lowest obtainable temperature is given by the abscissa of the point where the two lines intersect. Its value can be calculated by equating the two expressions just arrived at for Q_E and ΔQ_R. Taking $R = 2$J/g, $c_p = 5$ J/g (the value for helium), $\ln p_1/p_2 = 0.7$, $T_C = 300$ °K and $\eta_r = 0.99$, we arrive at approximately 10 °K as the minimum value of T_E. As can be seen, the result depends markedly on the value of η_r. In practice there are other losses than those mentioned and the compression and expansion are not in fact isothermal. The lowest temperature we have hitherto reached with the normal system is therefore higher, being in the region of 20 °K.

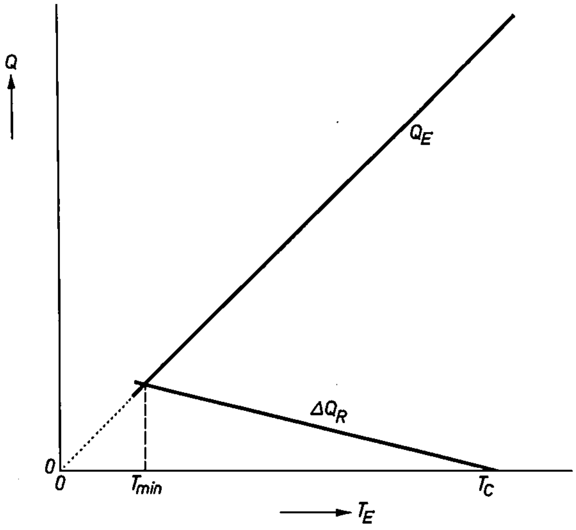


Fig. 3. Variation of the gross cold production Q_E and the regeneration loss ΔQ_R as a function of the temperature T_E of the expansion space in a normal Stirling cycle. Both curves are straight lines; extended, the first passes through the origin, the other passes through the point $T_E = T_C$. The lowest temperature T_{min} attainable is given by the abscissa of the point where the two lines intersect. At this temperature the entire cold produced is needed to compensate the losses.

From (1) and (2) it can be deduced that the minimum temperature obtainable with an ordinary Stirling cycle is lower:

a) the higher is the efficiency η_r of the regenerator,

b) the greater the pressure ratio p_1/p_2, and

c) the lower the compression temperature T_C.

Scarcely any practical use can be made of the first two properties for lowering the minimum value of T_E. In general, to increase the efficiency of the regenerator we need to increase its volume: this brings us into conflict with b), which calls for a large pressure ratio and hence for a small dead space. The machine that can reach 20 °K has in fact been designed with an almost optimum combination of η_r and p_1/p_2.

The third possibility, that of lowering the compression temperature T_C, presents difficulties in connection with the removal of compression heat. Where T_C is below room temperature, a second refrigerating machine is then needed to remove the heat of compression. This can of course be done, and we have thus been able to produce a temperature as low as 8 °K by using a cascade system of two gas refrigerating machines. A system of this kind, however, is not nearly as easy to operate, and does not have such a high efficiency as a single machine. Prospects are better with a multi-space machine, and this has been confirmed in practice.

Operation and characteristics of a multi-space machine

To give a better insight into the merits of the modified Stirling cycle underlying the new gas refrigerating machine, we shall return for a moment to fig. 2. We saw there that the regenerator losses of the section composed of a_2-R_1-R_2-c are partly compensated by the cold production of section a_1-R_1-b, which produces cold at the temperature T_M between T_C and T_E. It is known from thermodynamics that cold is produced most efficiently at the temperature at which it is needed. This means that theoretically the best way of compensating the regeneration losses of a Stirling cycle taking place between T_C and T_E is by using an infinitely large number of cold sources of different temperatures, together spanning the whole interval T_C-T_E. By using one or two in practice, however, a substantial proportion of the possible gain is already achieved. *Fig. 4* shows the diagram, analogous with that of fig. 2, of a machine which has *three* expansion spaces, called a four-space machine; here the cold required to compensate the regeneration losses is thus produced at *two* of the temperatures lying between T_C and T_E. In the following we shall confine our considerations to the three-space machine.

Referring to *fig. 5*, we shall now try to explain how and to what extent the regeneration losses of a three-space machine are compensated by the cold production of the middle cylinder (b in fig. 2) and the manner in which this shifts the minimum attainable temperature to a lower value. For this purpose we assume the same schematic simplifications as used in the reasoning that led to equations (1) and (2) and to fig. 3. Both solid lines in fig. 5 are the same as those in fig. 3 and thus represent the variation with T_E of the gross cold production Q_E and of the regeneration losses ΔQ_R of a normal Stirling cycle. Looking again at the situation at $T_E = 75$ °K, we see that the losses there are about 20% of the total production of cold. These losses, then, can be compensated by the cold production of only 20% of the total quantity of gas. Let us now turn to the process illustrated in fig. 2 and assume that this 20% is exactly the quantity of gas taking part in the (normal) cycle in the spaces a_1-R_1-b with T_M being 75°K. The remaining 80% of the gas is similarly cooled in the regenerator R_1 to 75 °K, and we shall now see what can be done with this in the second (normal) cycle, into which we imagine the modified cycle to be split, i.e. the cycle taking place in a_2-R_1-R_2-c. The curve representing the gross cold production Q_E' of this cycle, like the curve of Q_E, is a straight line through the origin, but of course its slope is 20% less steep. In the second cycle the curve of losses cuts the abscissa at 75 °K — the starting temperature for the part of the cycle taking place in R_2 — and has a slope which is 20% less than that of the ΔQ_R line. Expressed mathematically (see equation 2):

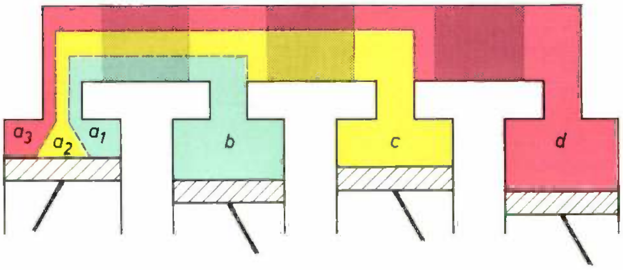


Fig. 4. Representation, analogous to fig. 2, of the operation of a machine with *three* expansion spaces (four-space machine). The process can be regarded as divided into three normal Stirling cycles.

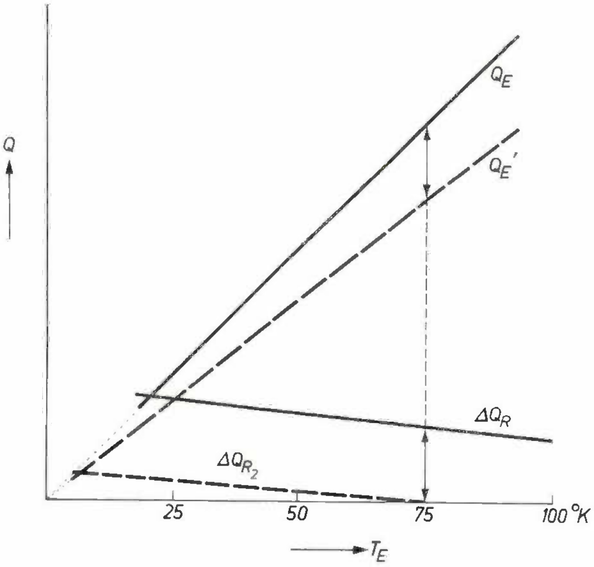


Fig. 5. Illustration of the fact that, using a three-space machine, a lower minimum value of T_E can be reached (abscissa of the point of intersection of the dashed curves) than with a machine in which the same quantity of gas goes through a single Stirling cycle (intersection point of the solid curves, cf. fig. 2).

$$\Delta Q_{R_2} = 0.8(1 - \eta_r) m\, c_p (T_M - T_E) . \quad . \quad . \quad (3)$$

As can be seen, the point of intersection of the two dashed curves is at a lower temperature than that of the curves relating to the normal cycle; at the lowest temperature attainable with the normal cycle, the modified cycle still has a reasonably high net yield.

Refrigerator or liquefier

So far we have spoken only of the possibility of using the cold production in the middle space for the compensation of regeneration losses; the production of externally useful cold was considered as taking place only in the end space at the temperature T_E. Where a three-space machine is employed in this way we shall say it is used as a "pure refrigerator". We can also, however, let more gas expand in the middle space than is needed for compensating the regeneration losses, so that in this space, too, useful cold is available for other purposes. This can be of practical importance, for example in the liquefaction of gases, where it is necessary, prior to the actual condensation, first to cool the gas down from room temperature to the temperature of condensation. In this case, cold is needed at all temperatures between room temperature and the condensation temperature. To effect this cooling most advantageously we should again, as noted in regard to the compensation of the regeneration losses, need to have a series of sources of cold, each operating at one of the temperatures within the interval. If a Carnot cycle took place in all these sources, the refrigeration needed for liquefaction would be obtained in the most efficient way. It is evident that here, too, we can benefit by arranging for the cold needed for cooling the gas to be produced not only at the lowest temperature of the interval, but also in the middle stage of the machine. Calculation shows, for example, that the cooling and condensation of hydrogen — condensation temperature 20.4 °K — requires:

a) with an infinite series of Carnot cycles . . . . . 14×10^6 J/kg;

b) with an ideal, pure refrigerator . . 62×10^6 J/kg;

c) with one intermediate temperature 27×10^6 J/kg.

If we calculate the intermediate temperature at which the useful effect is greatest, we find for $T_E = 20$ °K that this should be roughly 80 °K both when the machine is used as a "pure refrigerator" or as a liquefier. From the equality of both T_M values it is obvious that the design must depend on the purpose for which the machine is to be used; the middle stage of one and the same machine, with a given T_E and T_M, obviously cannot have two different values of cold production.

When cooling by means of an infinite series of Carnot cycles, the energy $Q_{\min}$ needed per gram to liquefy a gas with specific heat c_p, heat of condensation r and boiling point T_{liq}, is given to a good approximation by the formula:

$$Q_{\min} = -c_p \int_{T_0}^{T_{\text{liq}}} \frac{T_0 - T}{T} \, dt + r \frac{T_0 - T_{\text{liq}}}{T_{\text{liq}}} . \qquad (4)$$

In this expression $(T_0 - T)/T$ is the reciprocal value of the efficiency of a Carnot cycle taking place between the temperature T_0 (in our case the ambient temperature) and T. (The approximation employed consists in considering c_p as a constant; at temperatures just above T_{liq} this is not the case.) Denoting T/T_0 by τ and integrating, we find from (4):

$$Q_{\min} = c_p(T_0 - T_{\text{liq}}) \left(\frac{\tau}{\tau - 1} \ln \tau - 1 \right) + r(\tau - 1) . \qquad (5)$$

Using a pure refrigerator, however — which must produce all the cold at T_{liq}; this is the most unfavourable situation —, cooling plus condensation costs per gram an energy $Q_{\max}$ given by:

$$Q_{\max} = \{c_p(T_o - T_{\text{liq}}) + r\} (\tau - 1).$$

With one intermediate temperature the energy Q_1 required per gram for cooling plus liquefaction is:

$$Q_1 = c_p(T_0 - T_m) \frac{T_0 - T_m}{T_M} + \\ + c_p(T_M - T_{\text{liq}}) \frac{T_0 - T_{\text{liq}}}{T_{\text{liq}}} + r \frac{T_0 - T_{\text{liq}}}{T_{\text{liq}}} . \qquad (6)$$

Calculating the value of T_M at which this expression is minimum, we find:

$$T_M = \sqrt{T_0 T_{\text{liq}}} . \qquad (7)$$

For this value of T_M we have:

$$Q_1 = c_p \left\{ 2 T_0 \sqrt{\frac{T_0}{T_{\text{liq}}}} - 3 T_0 + T_{\text{liq}} \right\} + r \frac{T_0 - T_{\text{liq}}}{T_{\text{liq}}} . \qquad (8)$$

Applying the above to the liquefaction of hydrogen ($T_{\text{liq}} =$ 20.4 °K) we then find the numerical values for the three examples discussed.

Since the cooling and condensation take place under constant pressure, the production of cold required from both stages in a given case can easily be computed from the mass m and the relevant enthalpy change of the gas to be condensed (according to the thermodynamic definition, the change in enthalpy H is equal to the heat supplied under constant pressure). The middle stage, which is at a temperature of about 80 °K, must therefore deliver $m(H_{300} - H_{80})$, and the end stage: $m(H_{300} - H_{\text{liq}})$. For hydrogen the first amount is roughly three times as high as the second, i.e. for cooling and condensing hydrogen the net cold production of

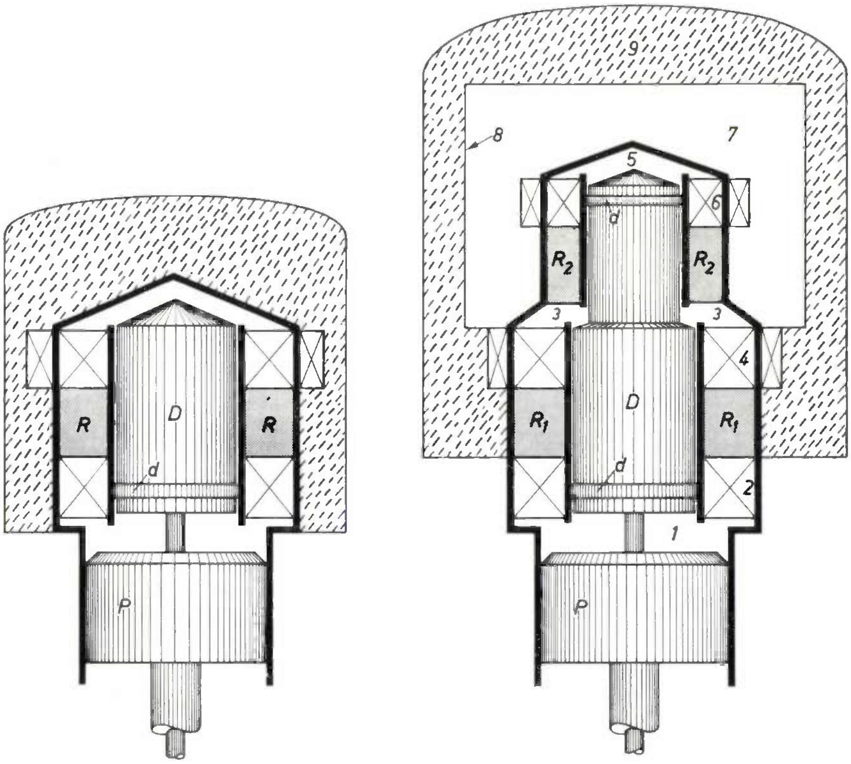


Fig. 6. Schematic cross-section of the new machine. A cross-section of the normal gas refrigerating machine is shown on the left for comparison. P piston. D displacer; in the new machine this consists of two sections with different diameters. d piston rings of D. *1* compression space. *2* appertaining cooler. R_1 first regenerator. *3* intermediate expansion space. *4* appertaining freezer; this is used only when the middle stage has also to supply cold. R_2 second regenerator. *5* expansion space. *6* appertaining freezer. *7* high-vacuum space which insulates R_2, *5* and *6*. *8* screen at temperature T_M to restrict radiation losses. *9* low-vacuum space filled with powder for thermal insulation.

the middle stage at 80 °K should be roughly three times as high as that of the end stage at 20 °K.

Design and features of the new machine

In designing the prototype of the three-space machine the working temperature of the end stage was taken to be 20 °K, the requirement being that the machine should have a high efficiency as a pure refrigerator and at the same time meet reasonable demands as a hydrogen liquefier.

We have just seen that these two applications call in principal for different machines if the maximum of efficiency is wanted. A compromise therefore had to be found. The procedure for this purpose was first of all to design, with the aid of a computer, a pure refrigerator having the highest possible efficiency at 20 °K. To this purpose we started from a design produced on the basis of estimates, and successively changed all principal parameters in the direction of higher efficiency. The characteristics of the design thus obtained were then computed. These are the curves that show the net cold production q_E and q_M of both spaces as a function of the temperature of the middle space.

From the considerations discussed in the previous section it may be deduced that the characteristics can be modified by changing the dimensions of the intermediate expansion space. Taking now the optimized case as our basis, we altered the dimensions so as to obtain a q_E characteristic which was almost horizontal. The net cold of the end stage of the new machine is thus hardly influenced at all by the value of T_M, that is to say by the amount of cold removed from the middle stage.

The construction of the machine is quite different from that possibly suggested by fig. 2. There is only one piston and one displacer, but the latter consists of two parts of different diameter. In this way both expansion spaces could be obtained using a single displacer body. *Fig. 6* shows side by side the schematic cross-section of the new machine and that of the single expansion-space machine with displacer.

The virtues of the single-space machine with the displacer construction, i.e. compactness and a low pressure drop over the displacer seal, are also present in the new machine. A further advantage of this design is that, since there is only one displacer, the driving mechanism is the same as that of the former machine.

The other details of the construction, with the two freezers and insulation spaces around the colder freezer, can clearly be seen in the drawing.

Characteristics

In order to be able to measure the characteristics, we fitted both freezers in the prototype with an electrical heating wire. This device makes it possible to measure the net cold production in a wide temperature range. In fact one really does the opposite: the choice of the current in each of the heater wires establishes q_E and q_M, and it is then a matter of determining the corresponding values of T_E and T_M. With this method of measurement a relatively high degree of accuracy can be achieved, because the freezers can in principle be perfectly insulated.

The results of the measurements are presented in *fig. 7*. It can be seen that the $q_E(T_M)$ curves are almost horizontal; the cold produced by the end stage is virtually unaffected by that of the middle stage.

From these characteristics it is also easy to find what the cold production is when the machine is used as a "pure refrigerator". All one has to do is to find from the graph, for a given value of T_E, what the value of q_E is at the T_M value at which q_M is zero — the left end of the q_E curves in fig. 7. In *fig. 8* these q_E values are plotted versus T_E. At 12 °K the value of q_E is zero, which is evidently the lowest temperature attainable. Above this the curve is fairly steep, and at 20 °K as much as 100 W of cold is already being produced. Because of this high production the time the machine takes, after being switched on, to reach a certain low temperature is extremely short.

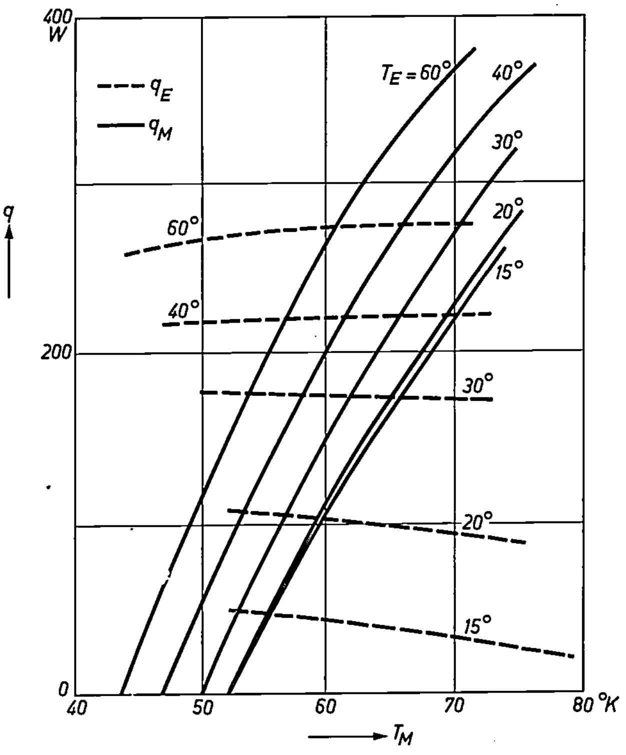


Fig. 7. Characteristics of the machine. The solid lines represent the variation of q_M at the indicated values of T_E, the dashed lines the variation of q_E. The latter curves are almost horizontal, which is an advantage for the liquefaction of gases. The curves were obtained with the working gas (helium) at an average pressure of 30 atm and with a crankshaft speed of roughly 1500 r.p.m.

When the characteristics for each adjusted value of q_E were measured the shaft power N of the machine was measured at the same time, and therefore the efficiency η could also be calculated and from this the ratio of η to the efficiency η_C of a Carnot cycle operating between the same temperatures. This relative efficiency is given by:

$$\frac{\eta}{\eta_C} = \frac{q_E}{N} \times \frac{T_0 - T_E}{T_E}.$$

The results of these calculations are also plotted in fig. 8 (dashed curve). It can be seen that at the working temperature on which the design was based (20 °K), the machine supplied the above-mentioned 100 W of cold with a relative efficiency of 17 %.

Finally, it should be noted that the curves in fig. 7 and fig. 8 were obtained with the working gas at an average pressure of 30 atm and a crankshaft speed of 1500 r.p.m. If these values are reduced somewhat the efficiency is slightly higher but the production lower.

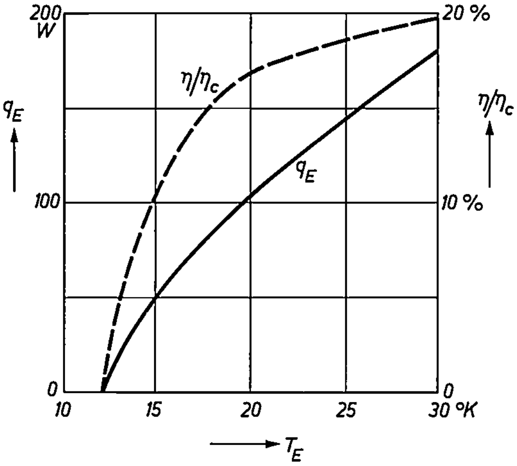


Fig. 8. Variation with T_E of the net cold production q_E and the coefficient of performance η (as a fraction of the efficiency of the Carnot cycle η_C) when cold is taken from the middle freezer (the machine being used as a pure refrigerator). The net production is zero at $T_E = 12$ °K, which is thus the lowest attainable temperature. Above this value the cold production increases rapidly with rising temperature.

Comparison with other plants

The main advantage of the new machine *(fig. 9)* compared with other refrigeration plants operating in the same temperature range is its simplicity. It requires none of the various components that make other refrigeration plants complicated, such as compressors, counterflow heat-exchangers, expansion engines, precooling baths and throttle valves, that always need control. In our machine the compression, expansion and heat exchange are combined in one unit, all processes thus being automatically adjusted to one another. Finally, for the liquefaction of hydrogen it is not necessary to compress the gas very much, or if so wished not at all, which is a great advantage.

The merits summarized above have not been acquired at the expense of the efficiency. As mentioned, the efficiency of our plant working as a pure refrigerator is 17 % of the efficiency of the Carnot cycle at 20 °K; if we had not chosen the above-mentioned compromise, but had designed the machine optimally as a refrigerator, our calculations show that the efficiency would have been about 21 %. These figures compared favourably with those of similar plants. For example, a refrigerating machine has been described [5] which was specially built for compensating the heat losses of a hydrogen bubble chamber and gives 300 W of cold at 27 °K. Calculations, which the designers themselves refer to as rather optimistic, put the relative efficiency at 20 %. The relative efficiency of our machine at this temperature is 19 %. A three-stage cooling system has also been described [6] which has a relative efficiency of 4.1 % at 14 °K. At this temperature our machine has a relative efficiency of 8 %.

As a hydrogen liquefier too, our machine compares very favourably. The previously built liquefiers have a power consumption 8 to 10 times the theoretical minimum [7]. This amounts to a value from 1.85 to 2.3 kWh/l. The consumption of our machine is 1.9 KWh/l.

Applications

We have shown in the foregoing that the new gas refrigerating machine constitutes an exceptionally simple refrigerator that can operate at temperatures down to about 12 °K, and has a particularly high efficiency in the region from 14 to 60 °K. In principle a machine of this kind can find application where a temperature is to be maintained in this region or where it is required to cool something to that temperature.

With regard to the use of the new machine as a pure refrigerator, an application that springs to mind is the recondensation of hydrogen (or neon) which has evaporated from a cooling bath, a storage tank, a bubble chamber, etc. In the case of hydrogen this can be of particular importance where during the liquefaction the ratio of the concentrations of orthohydrogen and parahydrogen have not been brought into line with the new temperature (at 300 °K 25 % para-, at 20 °K 99 %). The conversion then continues for hours in the liquid produced, in which process heat is liberated [8]. The machine can be used as a recondenser at any temperature between 14 and 30 °K. The cold production at a given temperature can be read from fig. 8.

In low temperature research it is not absolutely necessary to work with a cryostat kept at the temperature required by the evaporation of a liquid. It is also possible to use a cryostat mounted on the head of the ma-

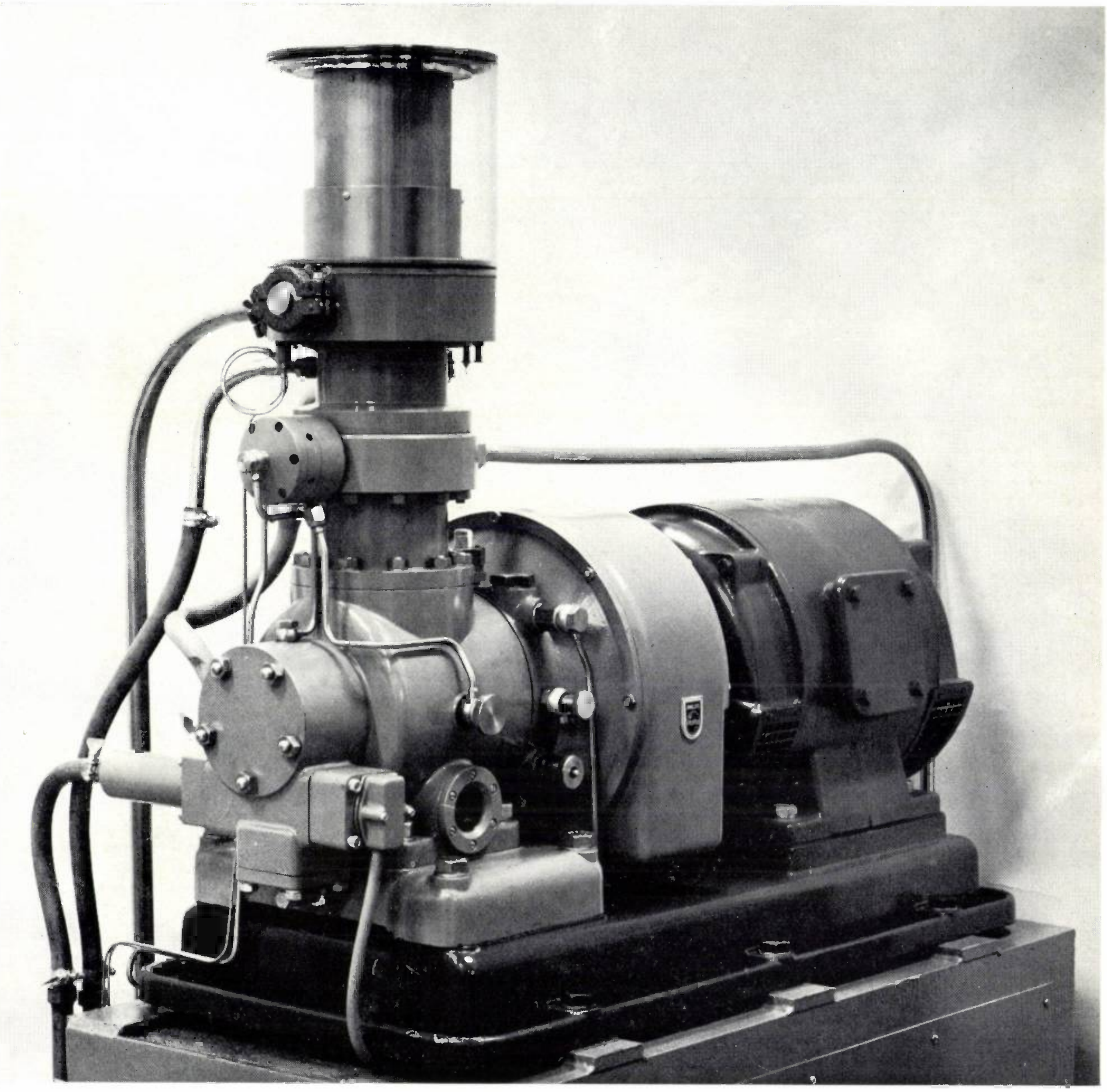


Fig. 9. The new machine. The insulating wall (9 in fig. 6) of the evacuated space which surrounds the colder expansion space has been replaced by a wall of Perspex so that the stepped portion of the cylinder can be seen.

chine itself and kept cold by conduction. The advantages of this are that the cryostat can be used throughout the entire range from 12 to 300 °K and that the cooling is very much faster. A drawback is that the cryostat is rigidly mounted on the machine and is thus subjected to mechanical vibrations. There will be many measurements, however, where this will not be an objection. A cryostat of this type is shown schematically in *fig. 10*, and a photograph is to be seen in *fig. 11*. Other important advantages are that the installation is very simple, that the cold space is easily accessible, because there is no need to be so sparing with the cold, and that the temperature can quickly be varied, e.g. from 80 to 12 °K in 11 minutes.

A relatively recent low-temperature application is what is termed "cryopumping", i.e. the evacuation of an enclosure by freezing gases. For this purpose, too,

[5] D. B. Chelton, J. W. Dean and B. W. Birmingham, NBS Technical Note 38.

[6] H. O. McMahon, Cryogenics **1**, 65, 1960.

[7] See e.g. Cryogenic Technology, edited by R. W. Vance, Wiley, New York 1963, Chapter II.

[8] This effect is described by R. B. Scott in Cryogenic Engineering, Van Nostrand, Princeton 1959, and also by S. Flügge, Handbuch der Physik XIV, Springer, Berlin 1956.

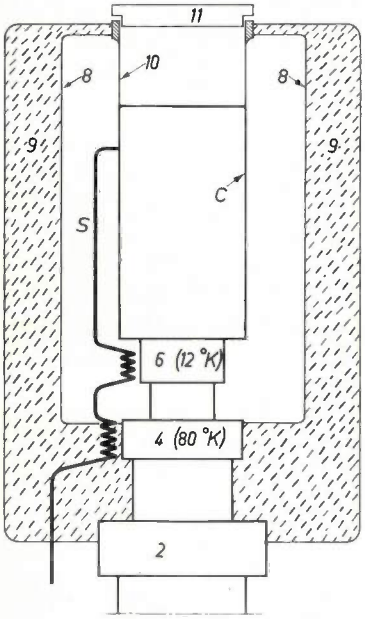
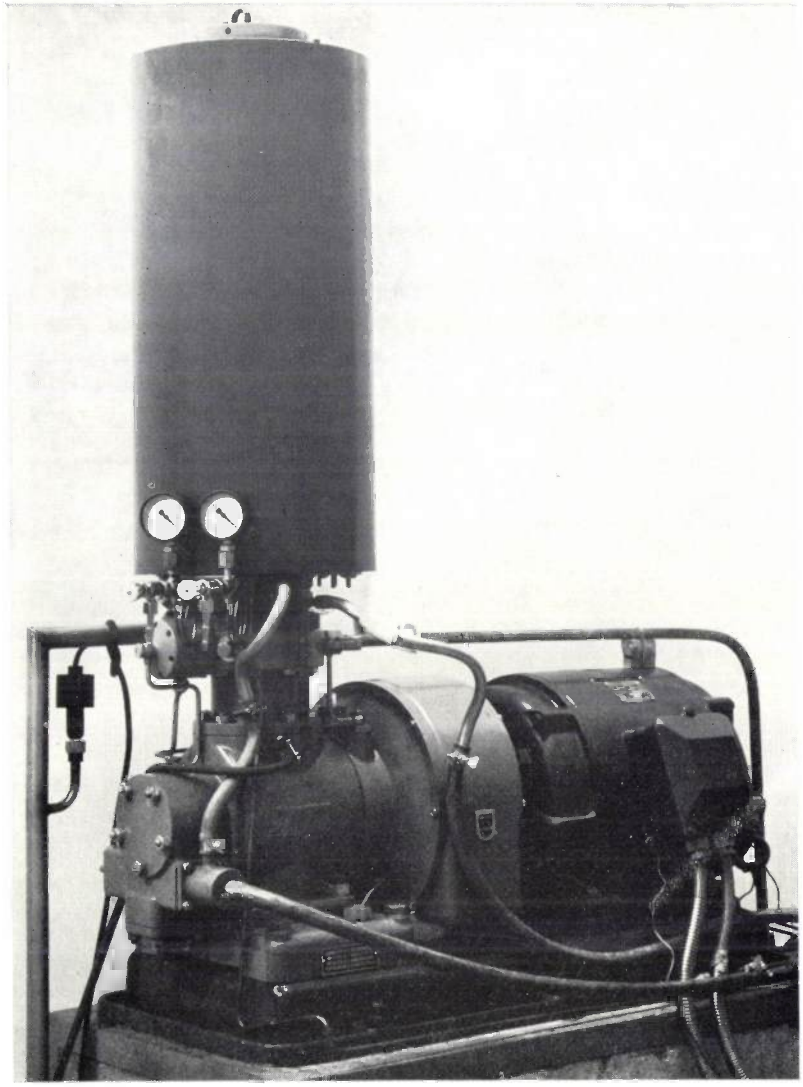


Fig. 10. Schematic cross-section of a cryostat for measuring purposes, mounted directly on the machine. The figures have the same meaning as in fig. 6. The measuring space *C* is formed by a copper vessel of 11 cm diameter and 30 cm length. *S* gas supply line (hydrogen or neon). *10* cylinder of material with low thermal conductivity. *11* cap.



→

Fig. 11. The cryostat of fig. 10 mounted on the three-space machine.

the new machine is eminently suitable. Since nitrogen and many other gases have a vapour pressure at 20 °K of about 10^{-9} mm Hg or less, the machine can be used to produce a very high vacuum. Cryopumping is superior where the evacuation should be fast, or where large quantities of desorbed gas are to be removed while maintaining a pressure of 10^{-3} to 10^{-6} mm Hg; the pumping speed required for the latter purpose cannot normally be reached using conventional pumping methods. A great advantage of cryopumping is that the "pump" does not introduce any contamination (such as oil or mercury) into the space to be evacuated.

Mention has already been made of the machine's application as a hydrogen liquefier, and of the considerable importance in this connection of the middle stage. Since the machine can supply cold at 20.4 °K, the hydrogen can be condensed at a pressure of 1 atm; thus that compression is not needed, as has already been pointed out. The machine can supply a quantity of about 5 litres of liquid hydrogen per hour. It can also, of course, condense neon gas.

Another application is as a precooler for a helium liquefier. The two stages of the machine can then be used instead of the nitrogen and hydrogen baths of the conventional helium liquefiers (*fig. 12*).

An aggregate of this kind can also be used as a refrigerating plant, as distinct from a liquefier. It is difficult to indicate how much cold a plant can supply at 4.2 °K, because this depends primarily on the helium compressor and on the quality of the heat exchangers. In theory, however, at 4.2 °K almost as much cold can be produced as is supplied by the head of the machine, the reason being that in principle a Joule-Thomson system is simply a system for transport of cold. If the head is held at 20 °K, then, the production is 100 W. Since, however, the maximum integral Joule-Thomson effect at 20 °K is 11.4 J/g, it is necessary for this purpose to pump the helium around the circuit at a rate of about 9 grams per second. To do this a fairly large compressor is required.

A simple experimental arrangement on this principle has been made using a small compressor and very simple heat exchangers. This apparatus produces about 4 W of cold at 4.2 °K. The relative efficiency is 3.75 %, which is a most satisfactory value for such a simple set-up [6]. This simple helium system was designed in the form of a cryostat and is used for keeping masers, cryotron memory devices, superconducting coils, etc. at the required low temperatures.

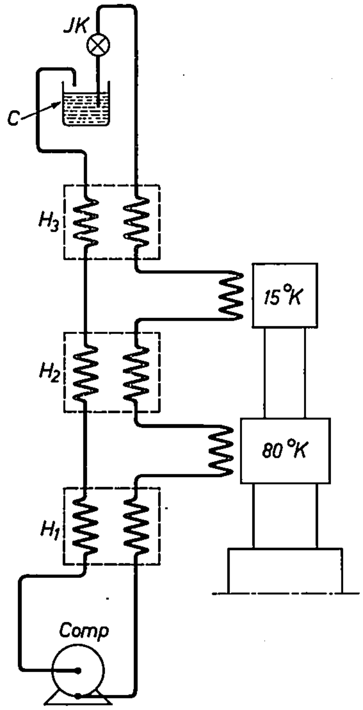


Fig. 12. Diagram of a helium liquefier — or refrigerating plant for 4.2 °K — consisting of a three-space gas refrigerating machine (right), several heat exchangers *H*, a Joule-Thomson system (throttle valve) *JK* with helium vessel *C* and a compressor *Comp*. In the compressor helium gas is compressed to e.g. 20 atm and cooled when passing through the two freezers — by the first to about 80 °K, by the second to about 15 °K. In the throttle valve *JK* the gas expands to 1 atm, in which process it is further cooled so that a part condenses. This part arrives in the vessel *C*. The gaseous part flows back to the compressor where, in the heat exchangers H_1 to H_3, it largely transfers its cold to the gas which is on its way in the other part of the circuit o the throttle valve.

Summarizing, it can be said that a new type of refrigerating machine has been developed which extends to 12 °K the temperature region covered by Philips refrigerating machines operating on the Stirling-cycle principle. Round about 20 °K the machine has a high refrigeration output and a relatively high efficiency; it also takes up very little room. The main advantage, however, is the remarkable simplicity of the system, which makes maintenance almost superfluous and reduces attendance to switching on and off.

The new machine opens up a temperature region which has hitherto only really been accessible to workers in cryogenic laboratories.

Summary. The efficiency at low temperature of a refrigerating machine based on the Stirling cycle can be improved and the minimum temperature lowered by compensating the regeneration losses by means of a series of additional expansion spaces whose temperature lies between room temperature and that of the last expansion space. The working gas, after compression, flows through a regenerator into the first auxiliary expansion space, but part of it passes through a second regenerator to the next expansion space, and so on. A machine is described which has one extra expansion space. The most favourable operating temperature of this "three-space machine", as it is called, depends on whether the extra space is required to supply cold, e.g. for the liquefaction of gases, or not. The machine is designed to produce cold (100 W) at 20 °K with good efficiency (17 % of the efficiency of the Carnot cycle), while at the same time meeting reasonable requirements as a hydrogen liquefier (production 5 l/h). The machine has a cylinder with a narrow and a wide section, in which a displacer moves similarly consisting of two sections having different diameters; the intermediate expansion space is the part of the wide section of the cylinder in which the narrow section of the displacer moves. The lowest temperature that can be reached is about 12 °K.

Phase theory

I. Introduction to unary and binary systems

J. L. Meijering

541.12.01

The highlights, the outstanding achievements of engineering which so often impress us, would not be possible without a firm scientific foundation. We have therefore thought it useful to call attention once again to a branch of knowledge which, by enabling us to control the properties of materials, may be regarded as one of the principal scientific aids to technology — phase theory.

Right from the very beginning the Dutch have left their mark on this branch of knowledge; names such as Bakhuis Roozeboom, Van der Waals and Van Laar constantly recur in every textbook on the subject. In the person of Prof. Meijering, who was awarded the Bakhuis Roozeboom medal in 1960, this tradition continues. We are greatly indebted to Prof. Meijering for being willing to write an article for us on phase theory.

The article will appear in three parts, the first of which follows below. This article does not adopt the usual practice of placing Gibbs' phase rule first and foremost. The author felt it was better from a didactic point of view to take as his starting point the striving towards minimum free energy — which in any case brings in the phase rule automatically. Although the phase rule constitutes a very important part of Gibbs' work, it is certainly not the quintessence of it.

Those who have thought of phase theory as a branch of science built up with almost mathematical rigor will find their view confirmed in this article. Even so the author demonstrates that the phase theory is capable of a lighter approach.

The whole gamut of industrial activities, from fundamental scientific research right up to the manufacture of products, is based to some extent, and to a greater extent than is generally realized, on the theory of heterogeneous systems.

The theory of heterogeneous systems, or phase theory as it is termed, is in general not very familiar to the layman, who regards it as the closed domain of a few specialists. The aim of this article is to make phase theory more comprehensible to the layman, and to show him how he might apply it to practical cases in his own field. To this end we shall explain, in a perhaps somewhat motley succession, some simple relationships, work out rules of thumb and apply them, and present some brief calculations. Because of the special aim of this article, its plan is and must obviously be different from that of most text books and articles on this subject.

We still have to define the concepts "heterogeneous" and "phase", which we have already used. A heterogeneous system is one consisting of more than one distinct phase, a phase — in the equilibrium state — being that part of the system whose chemical composition, structure and macroscopic properties are everywhere identical.

In this article we shall be mainly concerned with heterogeneous *equilibria*. This implies that where necessary we shall resort to thermodynamics, which is after all a theory of equilibria. For the present it will suffice to recall — and to use as our starting point — that every system tends towards a state of minimum free energy:

$$G = U - TS + pV,$$

where U is the energy, S the entropy, and T, p and V are the absolute temperature, the pressure and the volume.

A few simple examples follow with which we shall try to substantiate our statement that heterogeneous systems are of such great practical importance.

Many engineering materials consist of solid agglomerates of various phases, and their properties are determined to a marked extent by their heterogeneous

Prof. Dr. J. L. Meijering, formerly with Philips Research Laboratories, Eindhoven, is now Professor of Inorganic Chemistry and Metallurgy at the Technical University, Delft, Netherlands.

structure. It is known, for example, that copper wire becomes ferromagnetic when it contains iron as an impurity, whereas if the impurity is nickel — which likewise belongs to the ferromagnetic metals — it does not exhibit this effect. The explanation is that iron is not readily soluble in copper, and forms small islands of a second phase: nickel, on the other hand, dissolves very easily in copper. Again, the precipitation of bismuth (especially at the grain boundaries) is the cause of brittleness in copper contaminated with bismuth, while arsenic, which has a much better solubility, has scarcely any effect on the ductility of copper. On the other hand, if the conductivity of copper is the important consideration, then the arsenic content is more critical than that of bismuth, since impurities dissolved in the crystal lattice (As) reduce the conductivity much more than islands of a second phase (Bi). Such an important property as resistance to oxidation can be conferred upon a metal by dissolving in it a large quantity of chromium (about 20 at.%). This can be done successfully in nickel (nichrome), iron (chromium iron) and steel (austenitic chromium-nickel steel), but it cannot be done with copper, in which it is not even possible to dissolve as little as 0.2 at. % chromium.

Knowledge of the conditions governing "immiscibility" is an important aspect of phase theory. It is indeed indispensable to the application of chemical methods of separation and purification, such as for example fractional crystallization, fractional distillation, zone refining, etc. Other instances are processes where materials have to be brought into close contact without one dissolving into the other to any significant extent. Solid silver and iron in contact with each other can readily be heated to 950 °C, but at the same temperature gold and iron show considerable mutual solubility (3 at.% Au in Fe and roughly 60 at.% Fe in Au). Copper is melted in a graphite crucible, because it is known that only 0.0005 at.% C dissolves in molten Cu at 1100 °C, as against more than 1% C in molten iron and nickel. Very many other examples might be given of the practical importance of phase theory.

Although we shall confine ourselves mainly to (heterogeneous) equilibria, this does not of course imply that systems which are not in equilibrium are seldom encountered in practice. At relatively low temperatures the diffusion coefficients may be so small or reaction rates so slow that equilibrium cannot be established in a short time. We shall see later that many desirable properties of e.g. technical alloys in fact depend upon special non-equilibrium states. Moreover, the problem of controlling the properties of such materials — e.g. by "freezing in" a state of equilibrium followed by a suitable ageing process — can often be solved with the aid of phase theory.

The counterpart of this problem is the question of deducing the previous history of a non-equilibrium state encountered in practice. This question arises in industrial material testing and also in geology. By way of example we shall discuss later in this article a recent investigation relating to meteorites found on the earth.

Unary systems

We begin with a cursory treatment of unary systems, by which we mean a system which has the same composition in all its phases. This section may be considered more or less introductory to the treatment, in the next section, of binary systems which, being of greater importance, will receive more attention.

What happens when a piece of ice, at a pressure of 1 atm, is slowly heated? First, at 0 °C it begins to melt: below 0 °C the Gibbs' free energy of ice is smaller than that of water, and above 0 °C it is greater (*fig. 1*). Below 0 °C ice is *stabler* than water, the reverse being the case above 0 °C.

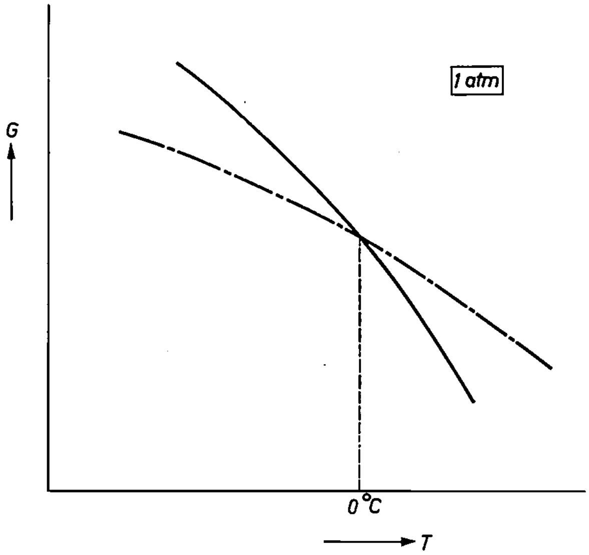


Fig. 1. Gibbs free energy G of water (solid line) and of ice (dot-dash line) as a function of the temperature at 1 atm pressure.

Secondly, the transition is *sharp*. Since this is regarded as more or less self-evident in the case of a unary system, it is perhaps as well to realize what this property depends on. *Fig. 2* illustrates the hypothetical case of an *un*sharp transition. There then exists a series of temperature values at which mixtures of the phase a and b are stabler than a alone or b alone. A certain temperature interval, from T_a to T_b, would have to be traversed in order to convert a completely into b.

In our example of a unary system, ice-water, where the temperature can only rise when the ice is entirely

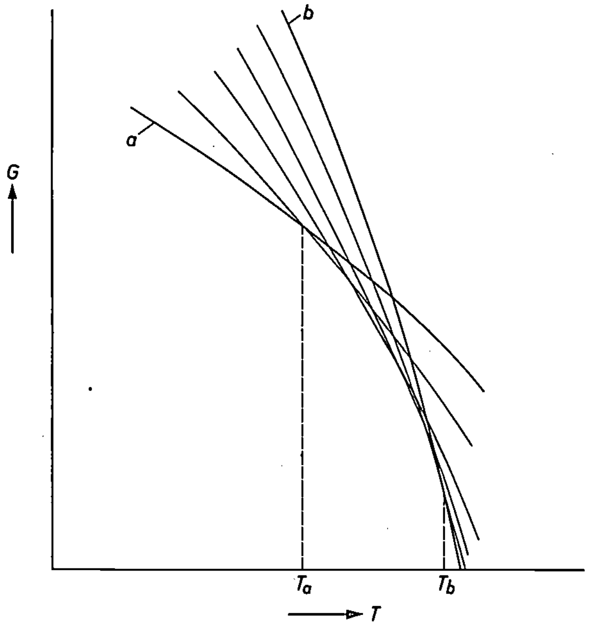


Fig. 2. Hypothetical *G-T* diagram of a unary system with phases a and b between which there is no sharp transition. The thick lines relate to the pure phases a and b, the thin lines to mixtures of the phases. Contrary to the ice-water system, a certain temperature interval from T_a to T_b has to be traversed in order to change a completely into b.

melted, the essential point is that the Gibbs' free energy *G* is *additive*. To understand this, let us consider a mixture of *x* moles of ice and (1 — *x*) moles of water. If the *G* of the mixture is equal to $xG_{\text{ice}} + (1-x)G_{\text{water}}$, in the equilibrium condition ($G_{\text{ice}} = G_{\text{water}}$) it will be independent of *x*. The difference from the hypothetical case considered in fig. 2 would appear in fig. 1 as follows: the *G-T* curves for the phases ice and water in different mixing ratios would all intersect at the same point.

A non-additive *G* is characteristic, for example, of systems where one or more of the phases occur in a very fine distribution, as in colloidal systems. In such systems the particle size influences the stability, a fact which is bound up with the surface free energy of the particles. We shall return to this point in Part III of the article. For the rest we shall confine ourselves to systems with additive *G*.

All extensive thermodynamic properties, such as *G*, *V* and *S*, will be given throughout this article per mole.

The p-T diagram

The fact that, when the temperature rises, ice is converted into water and not vice versa, follows from the greater entropy [1] of water. Since we can derive thermodynamically that

$$\left(\frac{\partial G_{\text{water}}}{\partial T}\right)_p - \left(\frac{\partial G_{\text{ice}}}{\partial T}\right)_p = -(S_{\text{water}} - S_{\text{ice}}) ,$$

it follows that G_{water} with increasing *T* must ultimately become smaller than G_{ice}; see also fig. 1. For simplicity we write the above expression in the form:

$$\frac{\partial \Delta G}{\partial T} = -\Delta S, \quad \ldots\ldots \quad (1)$$

where ΔG and ΔS are the differences in the Gibb's free energy and in the entropy between both phases. Likewise, we come to the conclusion that an increase of pressure promotes the formation of the phase with the lowest volume. This appears from the relation, analogous to equation (1):

$$\frac{\partial \Delta G}{\partial p} = \Delta V. \quad \ldots\ldots\ldots \quad (2)$$

More generally we can write:

$$\mathrm{d}(\Delta G) = \frac{\partial \Delta G}{\partial T}\,\mathrm{d}T + \frac{\partial \Delta G}{\partial p}\,\mathrm{d}p = -\Delta S\,\mathrm{d}T + \Delta V\mathrm{d}p,$$

which, in the case of equilibrium, i.e. in the temperature and pressure combinations at which there is no difference in the Gibbs' free energy between the phases, yields the relation:

$$\frac{\mathrm{d}p}{\mathrm{d}T} = \frac{\Delta S}{\Delta V}. \quad \ldots\ldots \quad (3)$$

The slopes given by these expressions for the vaporization curve, the sublimation curve and the fusion curve in a *p-T* diagram are, as a rule, all three *positive*, because in the sequence vapour-liquid-solid both the entropy and the volume generally decrease (*fig. 3*).

Exceptions are, for example, ice and gallium (*fig. 4*). The solid phase here is not very densely packed; because of the greater volume of the solid phase the slope of the fusion curve is negative. Upon a pressure increase the liquid phase evolves from the solid phase,

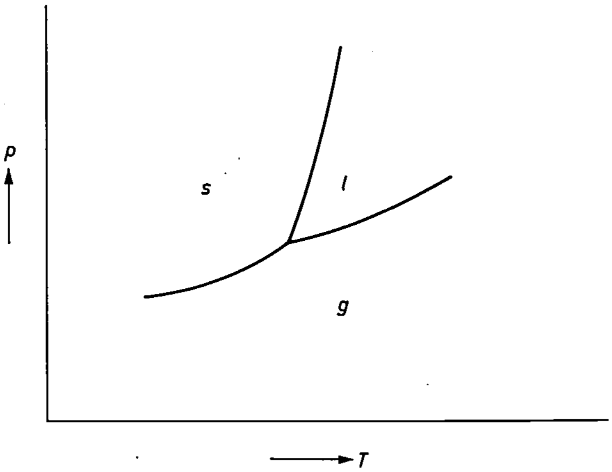


Fig. 3. Schematic *p-T* diagram in its most common form, with positive slopes for the equilibrium curves. g gas phase, l liquid phase and s solid phase. The three equilibrium curves meet at what is termed a *triple point*.

and not the other way round. There is also a chance that at a higher pressure another — more densely packed — solid phase will become stable (Ga_{II} in fig. 4). For ice at high pressure no fewer than five modifications have been found.

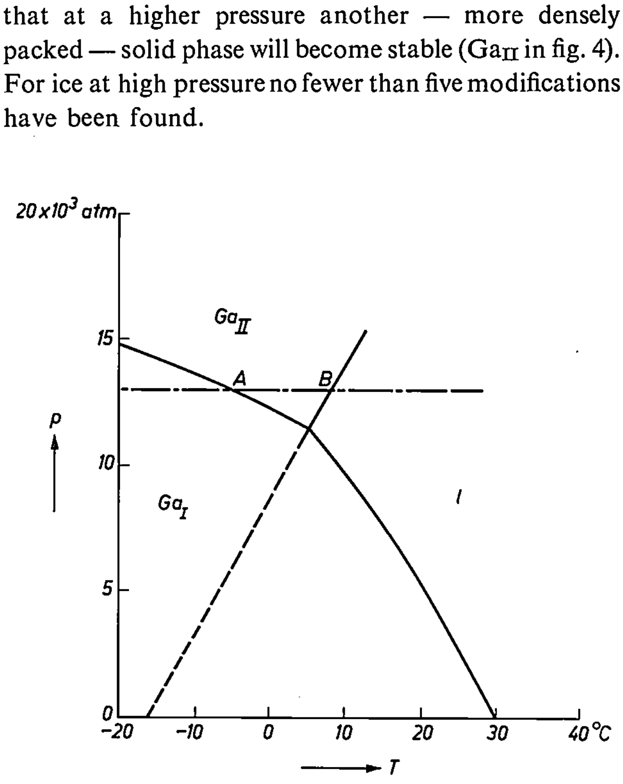


Fig. 4. *p-T* diagram of gallium, with triple point of the liquid phase l and the two solid modifications Ga_I and Ga_{II}. The dashed line indicates metastable equilibria which can occur if Ga_I, upon cooling, does not crystallize from the liquid phase. The dot-dash curve and the letters *A* and *B* relate to fig. 5. The triple point Ga_I + l + gas lies at very low pressure. It is nearly at the point where the fusion line of Ga_I cuts the *T* axis. If the vaporization and the sublimation curves were drawn in the figure, they would nearly coincide with the *T* axis.

Metastability

Fig. 5 gives the *G-T* diagram of gallium for a pressure higher than the triple-point pressure (see dot-dash line in fig. 4). In fig. 4 and fig. 5, *A* and *B* represent the same points: *A* is the transition point $Ga_I \rightarrow Ga_{II}$, *B* the melting point of Ga_{II} at the given pressure. State *C* in fig. 5, which would correspond to the melting point of Ga_I, and all states represented by dashed lines, are non-stable. It is usual to refer to such states as metastable. The significance of metastable states will be dealt with at length later.

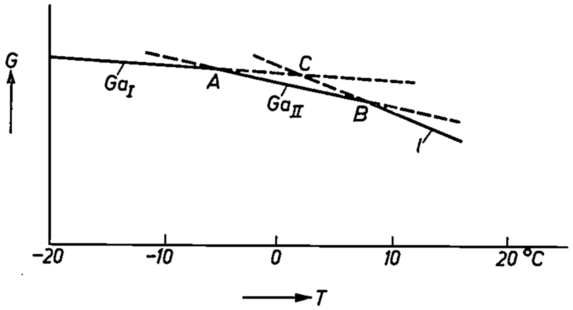


Fig. 5. Schematic *G-T* diagram of gallium (l, Ga_I and Ga_{II}) at the pressure denoted by the dashed line in fig. 4. *A* transition point $Ga_I \rightleftarrows Ga_{II}$, *B* melting point Ga_{II} and *C* melting point Ga_I which is metastable at the pressure considered. The solid lines indicate the *G* values of the phases in the stable state, the dashed lines those of the phases in the metastable state.

At the triple-point pressure the three *G-T* curves intersect at one point; *A*, *B* and *C* then coincide. In *fig. 6* we see that below the triple point pressure the points of intersection *A* and *B* are metastable, while *C* is now stable.

When liquid Ga is cooled to below the solidification point, it can happen that the solidification fails to take place, owing to difficulties in nucleation (undercooling). This means that in fig. 6 we follow the metastable line *CB* instead of the extension of *AC*. By undercooling liquid Ga to more than 45 °C at normal pressure, Defrain was able to solidify the metastable phase of Ga_{II} [2] (see also fig. 4).

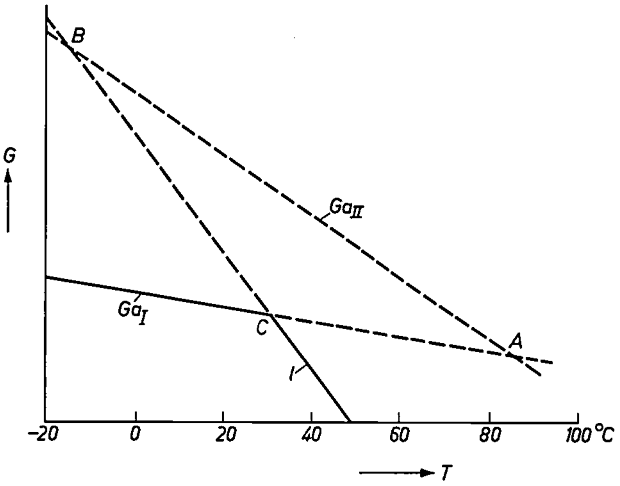


Fig. 6. Schematic *G-T* diagram of gallium at a pressure of 1 atm (i.e. below the triple point pressure, see fig. 4). By undercooling it has proved possible to cause solidification of the metastable phase Ga_{II} (at *B*).

Fig. 7 shows the *p-T* diagram of sulphur. There are here three stable triple points. The dashed lines represent the metastable two-phase equilibria in which the monoclinic phase is absent. They intersect at a *metastable triple point* (orthorhombic phase + liquid + gas), which can be realized experimentally. Each of the two-phase lines can of course be drawn beyond this metastable triple *point.* This has been done in the figure for the line liquid$\rightleftarrows$gas (the vaporization curve of liquid sulphur). Left of the metastable triple point this line becomes, as it were, doubly metastable. The system can

[1] For a detailed discussion of the entropy concept see: J. D. Fast, Philips tech. Rev. **16**, 258, 298 and 321, 1954/55.

[2] A. Defrain, Métaux, Corrosion, Industries **35**, 175, 245 and 300, 1960, and C. R. Acad. Sci. Paris **250**, 483, 1960.

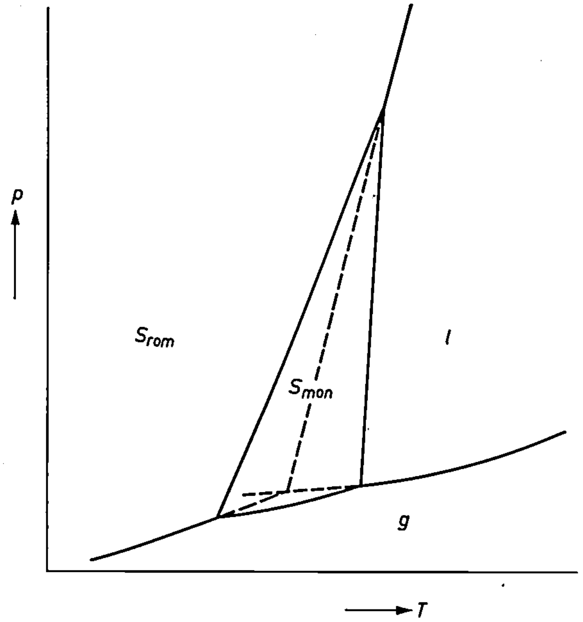


Fig. 7. Schematic p-T diagram of sulphur. S_{rom} is orthorhombic sulphur, S_{mon} is monoclinic sulphur. The dashed line; again indicate metastable equilibria. To the left, behind the metastable triple point, crystallization of both orthorhombic and monoclinic sulphur will bring the system into a more stable state.

now be brought into a stabler state not only by crystallization of the monoclinic phase but also of the orthorhombic phase.

Saturated vapour pressure as a measure of stability

The saturated vapour pressure of a substance can be made to serve as a useful measure of G. Assuming that the behaviour of the vapour is approximately that of an ideal gas, we may write for the G of the vapour, and hence for the G of the phase in equilibrium with it, the expression:

$$\frac{\partial G}{\partial p} = V = \frac{RT}{p},$$

from which

$$G = C + RT \ln p . \quad \ldots\ldots \quad (4)$$

To make this relation between vapour pressure and Gibbs' free energy clearer, it may be pointed out that in an ideal gas energetic interactions are insignificant and the value of G is solely governed by the entropy.

At —1 °C the saturated vapour pressure of undercooled water is higher than that of ice. This can be shown experimentally by placing an inverted U-tube with one limb in the undercooled water and the other limb in the ice; we then see that, via evaporation and condensation on the ice, the metastable water changes entirely into the stable ice — that is to say without the ice bringing about solidification by direct contact. Here we have a simple demonstration of the fact that the free energy strives towards a minimum.

Binary systems

The G-x and the T-x diagram

A system is called binary when, in addition to the variables already introduced, p and T, a concentration variable is needed to define the system. We begin by considering mixtures of silver and gold. At a temperature of 1000 °C — i.e. above the melting point of silver and below that of gold — these two metals can form both liquid and solid mixtures (mixed crystals), depending on their mixing ratio. At this temperature silver with only a little gold is liquid, gold with a little silver is solid. In *fig. 8* the mixing ratio is expressed in mole fractions of gold, x. On the ordinate, as a function of x, we have plotted the G of the solid phase (curve s) and the G of the liquid phase (curve l). For all x values on the left of the point where the curves intersect the liquid phase is stabler than the solid, and on the right on that point the converse applies. One might therefore be inclined to regard the given temperature of 1000 °C as the melting point of the alloy having the concentration given by the point of intersection. But we have not taken account of the possibility that the solid and the liquid phases may "coexist", in other words that they can form a heterogeneous mixture.

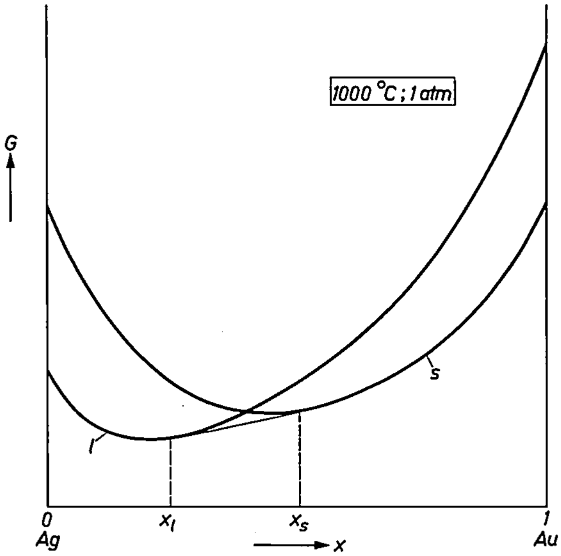


Fig. 8. Schematic curves of the Gibbs' free energy G of the binary Ag-Au system at 1000 °C and at normal pressure, as a function of x, the mole fraction Au. x_s and x_l denote respectively the composition of the solid and liquid phases in mutual equilibrium.

If the two phases of a heterogeneous mixture have free energy values G_1 and G_2, and if α is the fraction of the phase with the value G_1, then the G per mole mixture is given by:

$$G = \alpha G_1 + (1—\alpha) G_2 . \quad \ldots \quad (5)$$

Let x_1 and x_2 be the mole fractions of the two respective phases, then we can write for the Au mole fraction of the mixture:

$$x = \alpha x_1 + (1 - \alpha)x_2 \,. \quad . \quad . \quad . \quad . \quad (6)$$

It is not difficult to derive from (5) and (6) that

$$\frac{G - G_1}{G_1 - G_2} = \frac{x - x_1}{x_1 - x_2} \,. \quad . \quad . \quad . \quad . \quad . \quad (7)$$

This means that the G of the heterogeneous mixture can be represented as a function of x by the *straight line* joining the points denoting the G and the x of the constituent phases.

It can be seen that the "melting point" corresponding to the point of intersection of the curves in fig. 8 is not stable. For of course one can draw numerous lines between points of the two curves which, at the concentration in question, are lower than the point at which the curves intersect. The lowest one can reach is the *double tangent line* shown in fig. 8, which thus represents the most stable states. The corresponding concentration region comprises all (stable) alloys consisting of heterogeneous mixtures, of *the same* liquid, having the composition x_l, and *the same* mixed crystal having the composition x_s; a change in x (gross composition of the mixture) corresponds merely to a change in the relative amounts of the phases.

Outside this concentration region, both to the left and right of it, no heterogeneous combination is to be found that yields a lower G than the liquid or the solid phase in itself.

The G-x diagram changes with T of course, and so also do the concentrations x_l and x_s corresponding to the tangent points on the double tangent line. At the melting point of Ag the G-x curves intersect at $x = 0$, and the whole curve of the solid phase lies below that of the liquid phase. There is no double tangent line: both x_l and x_s are equal to zero. The higher the value of T the higher the value of x at which the G-x curves intersect, and the more the double tangent line shifts to the right, until at the melting point of Au, they intersect at $x = 1$; the G-x curve of the liquid phase then lies entirely below that of the solid phase.

A plot of each temperature T versus the appertaining values of x_l and x_s produces the T-x diagram of the system. In *fig. 9* this is shown schematically for the Ag-Au system. In a diagram of this kind the respective curves for x_l and x_s form the upper and lower limits of the two-phase region l + s; they are referred to as the *liquidus* and *solidus* curves respectively. Unlike the situation with unary systems, the relative amount of the solid phase (or liquid phase) in an equilibrium mixture (at a given x) depends on the temperature: while at point E in the figure the first crystals are about to appear, at point A the amounts of solid and liquid phase are proportional to the lengths of the lines AB and AC, and at point D only the solid phase is present. In short, in order with the given composition to convert the solid phase entirely into the liquid phase (or vice versa) it is necessary to traverse the interval DE; we call this the melting range.

In the previous example there was no sharp melting point at any single concentration — leaving out of account the two pure components. There is, however, another possible type of T-x diagram. Imagine a succession of G-x diagrams belonging to a series of ascending temperature values where the first diagram in the series is one in which the G-x curves of the various phases are *tangent* to one another instead of intersecting each other, and not at one of the end points but somewhere in the middle. The T-x diagram then appears as illustrated in *fig. 10*. There is then one composition at which a sharp melting point does appear.

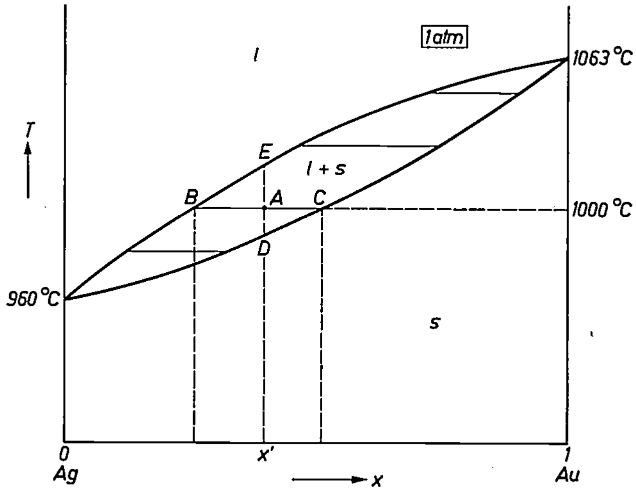


Fig. 9. Schematic T-x diagram of Ag-Au at normal pressure. There are three distinct zones: that of the mixed crystal phase s, that of the liquid phase l and the zone l + s where both phases "coexist". DE is the melting range of a mixed crystal with concentration x'. The concentrations of the coexisting phases can be found at any temperature within this range by drawing horizontal lines and determining the points where they intersect the upper and lower boundaries of the two-phase domain. In each case the ratio of the amount of the liquid phase to that of the solid phase is given by the horizontal section cut off by the lower boundary to that cut off by the upper boundary. For example, at 1000 °C the ratio is AC/AB. The upper and lower boundaries of the two-phase domain are referred to as the liquidus and solidus curves, respectively.

Fig. 11 shows a G-x diagram of a binary system with a liquid phase l and *two* solid phases s_I and s_{II}. At the given temperature *both* components in the pure state are solid (G_{s_I} and $G_{s_{II}} < G_l$). In the concentration region between the two inner dashed lines, however, the liquid phase is stable. In addition we have the two-phase equilibria l + s_I and l + s_{II}, represented by the two double tangent lines.

If we lower the temperature, then curve l occupies a higher position with respect to curves s_I and s_{II}, and the

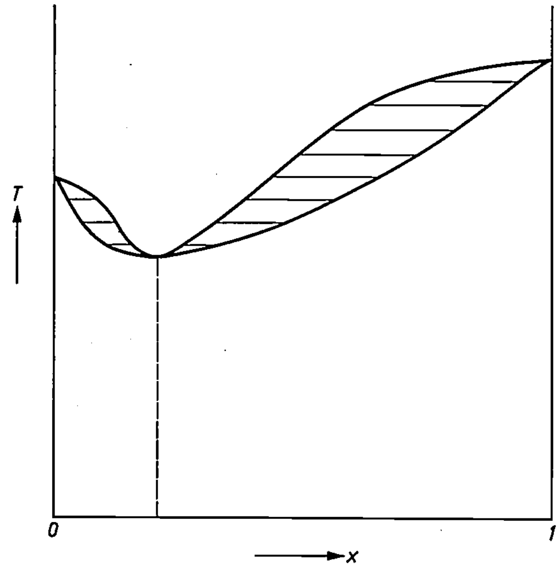


Fig. 10. T-x diagram of a binary system showing a melting-point minimum.

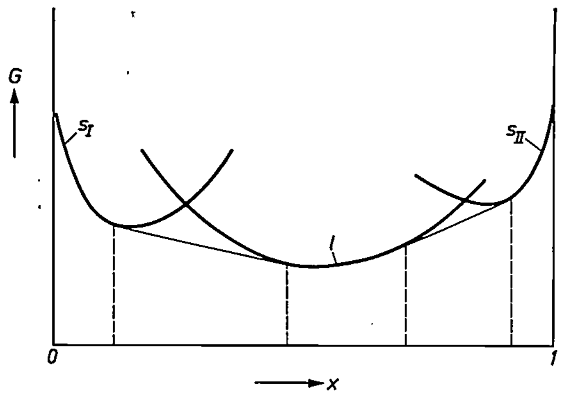


Fig. 11. G-x diagram of a binary system with a liquid phase l and two solid phases s_I and s_{II}.

tangent points on curve l approach each other. The temperature at which these tangent points coincide, and one straight line is tangent to all three curves, is called the *eutectic* temperature. This is the temperature at which *three phases* can be in equilibrium with each other. Let us now look at the T-x diagram in *fig. 12.* Here E is the eutectic point. It is interesting to consider what happens when the temperature of the liquid phase l is lowered. For this purpose we can follow the dashed line in the figure. Where this line intersects the liquidus curve the first crystals s_A begin to form. The lower the temperature drops the more crystals precipitate, and this is accompanied by a change in the composition both of the crystals s_A and of the liquid phase (as can be read from the solidus curve and the liquidus curve respectively). In this process the composition of l approaches the "eutectic" composition. When the eutectic temperature is reached, the remaining l fraction separates, while solidifying, into the solid phases s_A and s_B.

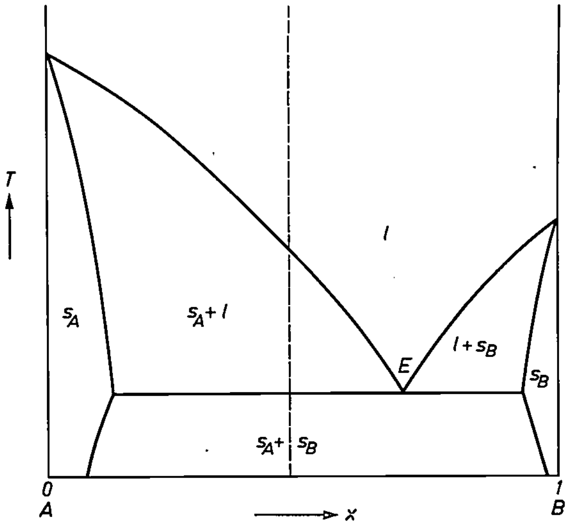


Fig. 12. T-x diagram with eutectic E. The diagram shows the liquid phase l, two homogeneous mixed-crystal phases s_A and s_B, and the three two-phase equilibria s_A + l, s_B + l and s_A+ s_B. The dashed line is mentioned in the text.

The traces of the latter event are to be found in the solidified mass as domains with a characteristic eutectic structure, that is to say domains consisting of alternate layers of s_A and s_B. The closer the composition of the liquid phase lies to the eutectic composition the more marked this structural pattern appears.

In systems in which a "*eutectoid*" occurs, i.e. a three-phase equilibrium that consists of three solid phases instead of one liquid and two solid, the same structural pattern appears. A good example of such a eutectoid structure can be seen in *fig. 13.*

Mention should be made of various other characteristics of the binary diagram. *Fig. 14* shows a G-x dia-

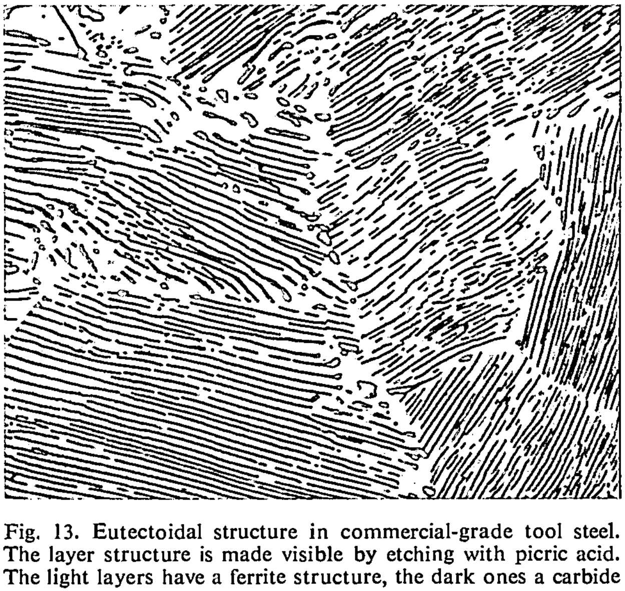


Fig. 13. Eutectoidal structure in commercial-grade tool steel. The layer structure is made visible by etching with picric acid. The light layers have a ferrite structure, the dark ones a carbide structure. Magnification 1400 ×.

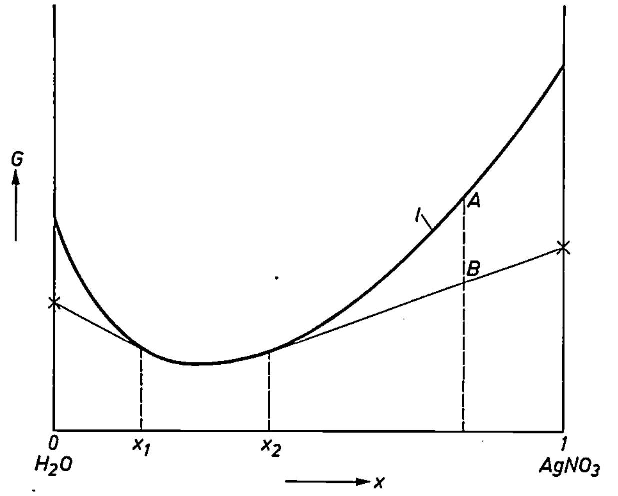


Fig. 14. Schematic G-x diagram of the $AgNO_3$-H_2O system at a temperature *above* the eutectic temperature. l liquid phase. The crosses denote the G values of ice and solid $AgNO_3$: the two components do not form mixed crystals. For $x < x_1$, a two-phase mixture is stable which consists of ice and liquid, with composition x_1. For $x_1 < x < x_2$ the liquid phase is stable. For $x > x_2$ the stable phase is a two-phase mixture of solid $AgNO_3$ and liquid, with a composition x_2. A and B are states which have the same composition, the first being supersaturated with $AgNO_3$, the second consisting of a saturated solution + $AgNO_3$ crystals. See also page 21.

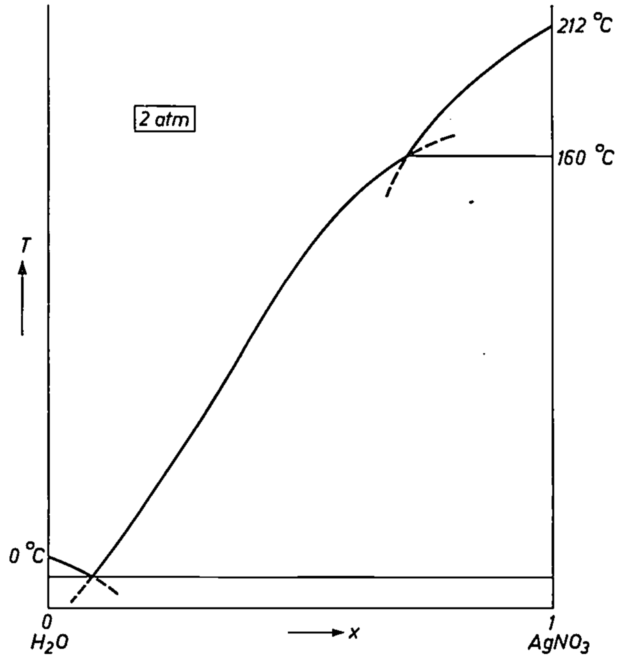


Fig. 16. T-x diagram of $AgNO_3$-H_2O at a pressure of 2 atm. In addition to the eutectic there is a three-phase equilibrium at 160 °C. At this temperature $AgNO_3$ changes to another modification. The dashed lines denote metastable two-phase equilibria.

gram of the system $AgNO_3$-H_2O. The difference compared with fig. 11 resides in the fact that in the crystal lattices of both ice and of solid $AgNO_3$ the other component is not incorporated to any significant extent (there is hardly any mixed crystal formation). The crosses indicate the G values of ice and of solid $AgNO_3$. In fact we again have double tangent lines, since the mutual miscibility in the solid state is not rigorously zero. To understand this we must imagine that in fig. 11 the curves on the extreme left and right are much more strongly curved, so that the tangent points on them coincide more or less with the initial points.

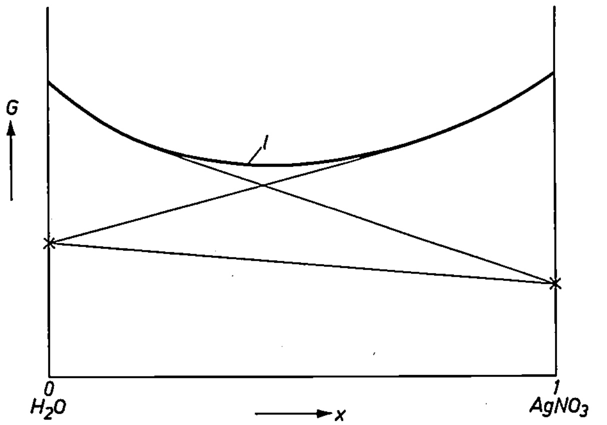


Fig. 15. Schematic G-x diagram of the $AgNO_3$-H_2O system at a temperature *below* the eutectic point. Only heterogeneous mixtures of ice and solid $AgNO_3$ are stable. The lines tangent to l indicate metastable two-phase equilibria, which can be produced if either H_2O or $AgNO_3$ does not crystallize.

Fig. 15 presents a G-x diagram of $AgNO_3$-H_2O at a temperature below the eutectic. The liquid phase now is not stable at any single concentration. Stability is shown solely by heterogeneous mixtures of ice and solid $AgNO_3$, as represented by the straight line between the crosses. Even so, we are justified in drawing the lines tangent to the curve l, as done in the figure, because these give the *metastable* equilibria between on the one hand the silver nitrate solution + ice, and on the other hand the silver nitrate solution + solid $AgNO_3$. The first equilibrium is realized when, upon cooling, no $AgNO_3$ crystallizes, and the second equilibrium when there is no ice formation. *Fig. 16* shows the T-x diagram of $AgNO_3$-H_2O, in which the metastable equilibria are represented by dashed lines.

The same figure also shows that $AgNO_3$ has a transition point at 160 °C. At this temperature, then, as in the case of the eutectic, three phases are in equilibrium with each other: the liquid with the low and high temperature modifications. Compared with the eutectic, however, this three-phase equilibrium is rather trivial: two of the three phases have virtually the same composition.

Attention should also be drawn to the kink in the liquidus curve in the neighbourhood of the transition point. With a kink of this form the metastable extensions lie in the two-phase domains (and not in that of the liquid phase). This is in fact a *general* characteristic of binary phase diagrams, which is also noticed for example in the case of the eutectic. This characteristic can be deduced from the G-x curves.

It would be going too far to attempt to derive all existing types of phase diagrams from the *G*-*x* curves at different values of *T* and *p* [3]. It will be enough to remember that the stable equilibria can always be thought of as derived by stretching, as it were, a string along all *G* curves of the phases concerned, as illustrated in fig. 11. We then have alternate homogeneous and two-phase fields, the first of which in particular may be very narrow.

Immiscibility

We have already mentioned in the introduction how important the phenomenon of immiscibility is in practice. Immiscibility in its simplest form — and we are inclined to use the term only for this form — is the occurrence of twó phases of similar structure, e.g. two liquid phases or two solid phases with the same (e.g. body-centred cubic) crystal structure. In the *G*-*x* diagram a two-phase equilibrium thus formed is given by a double tangent line on the same curve (see the second curve from the bottom in fig. 17) which might apply, for example, to the system Ag-Cu below the eutectic temperature.

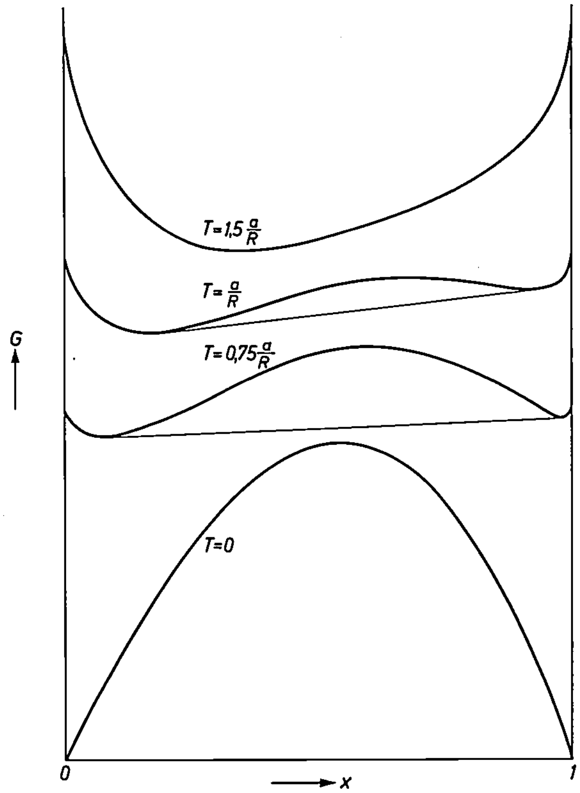


Fig. 17. *G*-*x* curves calculated from the formula $G ='a(2x\text{-}x^2\text{-}x^3) + RT\{(x \ln x + (1-x) \ln (1-x)\}$. This formula is used instead of eq. (10) in the text, which would only have produced curves symmetrical with respect to $x = \frac{1}{2}$. The four curves were computed for the same *positive* value of *a* (unfavourable energetic interaction between the components) and for various values of *T* ($T = 0$, $T = 0.75\ a/R$, $T = a/R$ and $T = 1.5\ a/R$; for clarity, the zero of *G* is drawn successively higher). From the points of contact of the double tangent lines the miscibility gap given in fig. 18 can be constructed.

Immisicibility will in general be the consequence of an unfavourable *energetic* interaction between the components. Silver atoms prefer to have their own kind around them in the crystal lattice rather than copper atoms. The simplest expression for the enthalpy of mixing ($U + pV$ for a homogeneous mixture) is:

$$H = ax(1-x)\,, \quad \dots\dots\dots \quad (8)$$

and for the entropy of mixing:

$$S = -R\{x \ln x + (1-x) \ln (1-x)\}\,. \quad . \quad (9)$$

When *a* is positive, equation (8) expresses that owing to the mixing the enthalpy increases in proportion to the number of dissimilar pairs of neighbours, it being assumed that the atoms are randomly distributed. Equation (9) is Gibbs' familiar formula for the entropy of mixing. Substituting (8) and (9) in $G = H - TS$ we obtain for the *G*-*x* curve the expression:

$$G = ax(1-x) + RT\{x \ln x + (1-x) \ln (1-x)\}\,. \quad . \quad (10)$$

It is evident that the shape of the curve depends on the value of *a* in relation to that of *T*, as illustrated in *fig. 17*. We again obtain the *T*-*x* diagram by plotting for each *T* value the concentrations that correspond to the tangent points on the double tangent lines. This gives us the typical miscibility gap that can be seen in *fig. 18*.

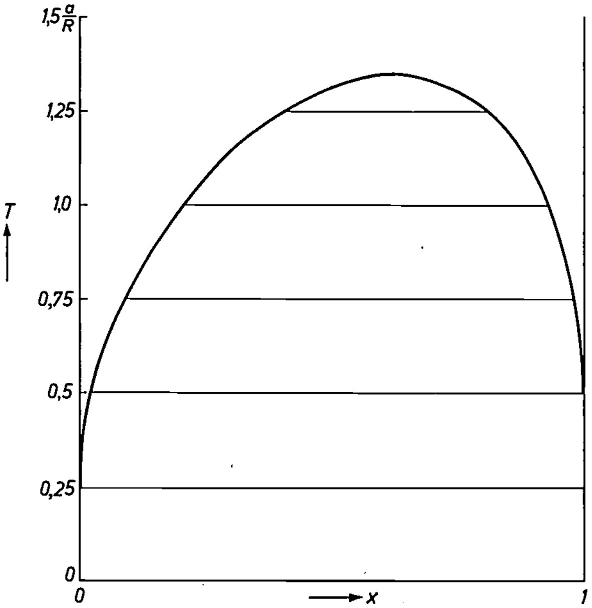


Fig. 18. A miscibility gap in the binary *T*-*x* diagram, calculated from the formula given in the caption to fig. 17. For each temperature the concentrations at the tangent points on the double tangent lines in the *G*-*x* diagram are plotted.

It will be noted, incidentally, that the curves shown in figs. 17 and 18 relate to a more general case than that defined by equation (10); in particular the curves are not drawn symmetrically with respect to $x = \frac{1}{2}$. We shall deal with this subject in more detail in part II.

The equilibrium criterion; differential or partial quantities

The question as to the equilibrium criterion for binary systems is in fact answered in figs. 8, 11 and 17: it is the coincidence of two or more tangent lines in the G-x diagram. A tangent line can be denoted by the ordinate values of its points of intersection with the axes $x = 0$ and $x = 1$ (*fig. 19*). We call these ordinate values μ_1 and μ_2 the "chemical potentials" of the system, and the equilibrium criterion is then that the chemical potentials should be equal for each of the coexisting phases.

The chemical potentials are termed differential or partial quantities. From fig. 19 it is easily seen that

$$\mu_1 = G - x\frac{dG}{dx}, \quad \dots\dots \quad (11)$$

$$\mu_2 = G + (1-x)\frac{\mathrm{d}G}{\mathrm{d}x}, \quad \dots \quad (12)$$

from which:

$$G = (1-x)\,\mu_1 + x\mu_2 . \quad \dots \quad (13)$$

Several kinds of thermodynamic quantities can be expressed as differential or partial quantities. For example we can write:

$$V = (1-x)\,V_1{}^* + xV_2{}^*, \quad \dots \quad (14)$$

and

$$S = (1-x)S_1{}^* + xS_1{}^*, \quad \dots \quad (15)$$

where $V_1{}^*$ and $V_2{}^*$, $S_1{}^*$ and $S_2{}^*$ stand in a similar relation to V and S as μ_1 and μ_2 to G. Their significance will become self-evident in the course of this article.

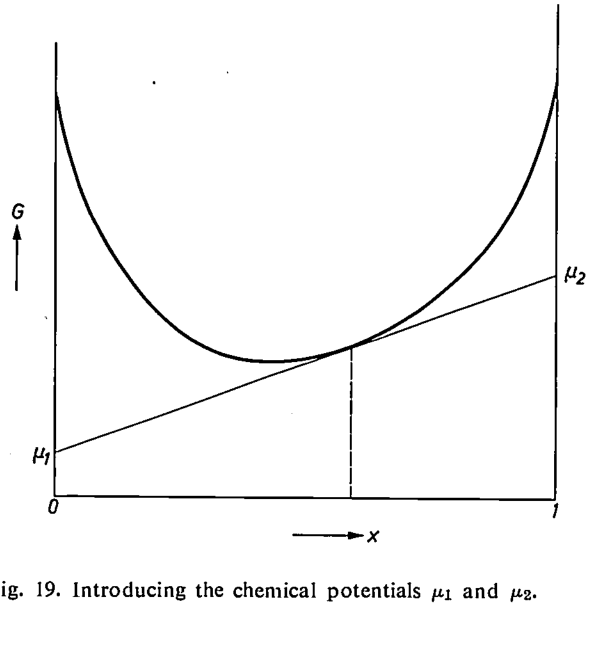


Fig. 19. Introducing the chemical potentials μ_1 and μ_2.

Chemical potential and partial vapour pressure

Just as in unary systems a relation can be established between the G and the saturated vapour pressure of a phase, in binary systems we can do the same between the chemical potentials μ and the partial vapour pressures [4].

Suppose that fig. 19 relates to the liquid phase of the system water-acetone and let x be the mole fraction, then μ_1 gives the partial pressure of the water vapour which is in equilibrium with the water-acetone mixture having the concentration of the tangent point, and μ_2 likewise gives the partial pressure of the acetone vapour.

At this point it is interesting to return to the experiment with the inverted U-tube described at the end of the section dealing with unary systems. As can be seen in fig. 14, an aqueous solution which is supersaturated in $AgNO_3$ (e.g. state A) has a lower μ_{H_2O} (partial vapour pressure of water) than a saturated solution + $AgNO_3$ crystals of the same composition (state B). The opposite applies to the partial vapour pressure of $AgNO_3$, but this is negligibly small. If we perform the inverted U-tube experiment with this system, we must then expect that the vaporization will take place from the *stable* system (accompanied by the further crystallization of $AgNO_3$), while the *metastable* system will "grow" as a result of the condensation of water. With unary systems this would be out of the question; the fact that it is possible with binary systems is connected with the changes that occur in the composition of the separate systems. (This can be inferred directly from fig. 14; as the x of the saturated system B increases at the expense of the supersaturated system A, the total G decreases owing to the greater slope of the G of the supersaturated system.) The transfer of vapour from the stable to the metastable system ceases as soon as a sufficient quantity of water vapour has condensed in the supersaturated solution for the latter to become saturated.

Solubility under changing pressure

Concerning the question of how the solubility can change on changing pressure (at constant T) it is instructive to glance at *fig. 20*. Curve 1 gives the volume of a liquid mixture of A and B as a function of x at a particular pressure. The composition, denoted by a dashed line, is the liquid mixture just saturated with A. The cross on the axis $x = 0$ is the volume of pure solid A. Since the liquid phase at any value of x has a greater volume than the solid phase of A, one might at

[3] See A. H. Cottrell, Theoretical structural metallurgy, Arnold, London 1955. See also R. Vogel, Die heterogenen Gleichgewichte, Akad. Verlagsges., Leipzig 1959.

[4] See also K. Denbigh, The principles of chemical equilibrium, Cambridge Univ. Press, 1957.

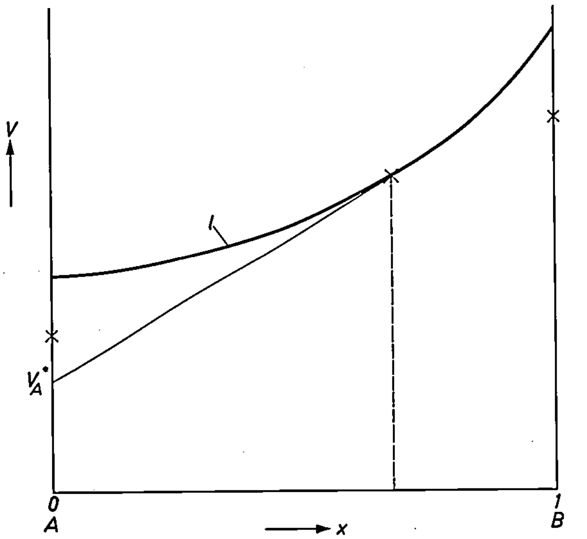


Fig. 20. V-x diagram. The curve l indicates the volume of the liquid phase, which is miscible in all proportions; the crosses on the axes $x = 0$ and $x = 1$ denote the volume of the components in the solid state, which do not form mixed crystals. At the concentration represented by the dashed line the liquid phase is just saturated with A. These data permit the conclusion that, at the concentrations in question, raising the pressure will increase the solubility of A. The conclusion follows in particular from the smaller partial volume of A in the liquid (V_A^*) compared with the volume of A in the solid state (see the cross on the $x = 0$ axis).

first sight expect that an increase of pressure would promote the crystallization of A. After closer examination, however, we arrive at a different conclusion. The crystallization of A implies that a little A has been extracted from the liquid mixture; the accompanying reduction of volume per mole of precipitated A is equal to the partial volume V_A^* of A in the liquid mixture. Since the partial volume of A is *smaller* than the volume of solid A, a pressure increase will not lead to the crystallization of A: a rise of pressure causes an increase in the solubility of A in the mixture.

There is another way of looking at this. Let us consider that the cross indicating the volume of the saturated solution in question is connected by a straight line to the cross denoting the volume of the solid A. Since this line passes to the left of the saturated concentration *above* the tangent line, it lies *above* the V-x curve of the liquid mixture. In other words, the volume of the saturated solution *containing a little undissolved solid A* is greater than the volume of a solution — not stable at the given pressure — having the same composition *in which all A would be dissolved.* For this reason, the solubility of A must be expected to increase with rising pressure.

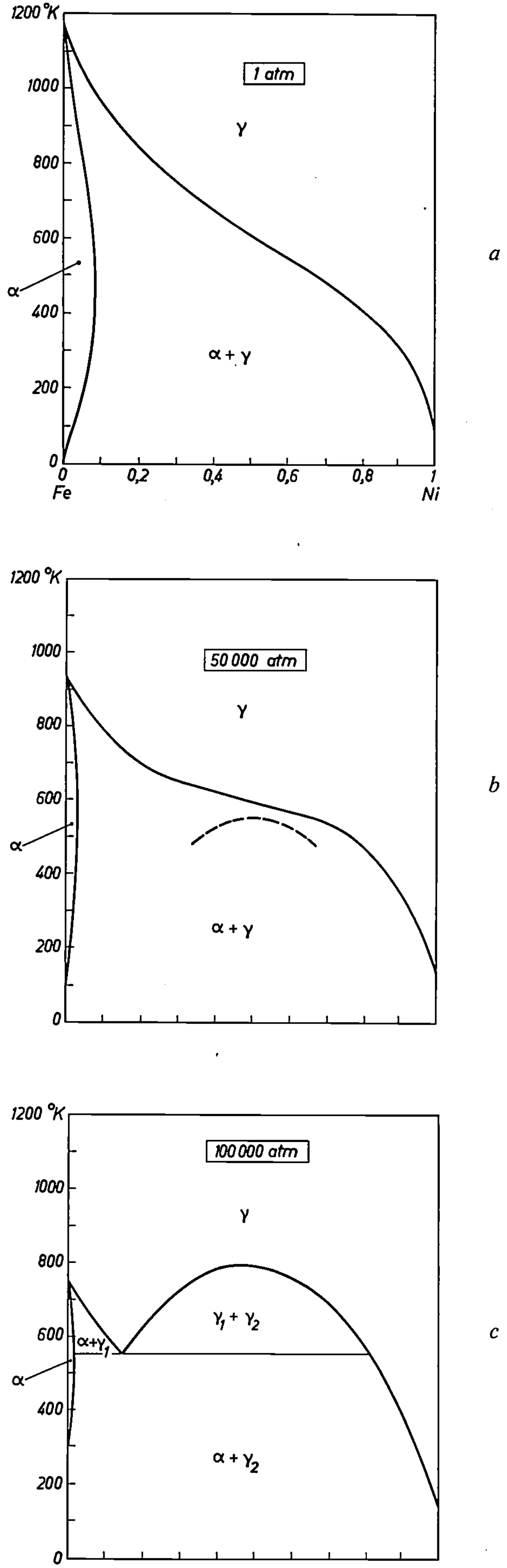


Fig. 21. Theoretical T-x diagrams of the iron-nickel system at pressures of a) 1 atm, b) 50 000 atm, c) 100 000 atm [5]. α body-centred cubic, γ face-centred cubic. In c, owing to the demixing of γ, foreshadowed in b in the form of the metastable miscibility gap marked by dashed curve, a eutectoidal three-phase equilibrium has formed ($\alpha + \gamma_1 + \gamma_2$). This might explain why certain meteorites, consisting of iron and nickel, have a eutectoidal structure. The necessary high pressure might have been produced at the moment of the meteorite's impact on the earth.

Theory concerning the origin of a eutectoidal structure in iron-nickel meteorites

In an attempt to explain the eutectoidal structure found in certain metallic meteorites, the possible influence of the pressure has also been investigated. The meteorites in question consist of iron to roughly 90 at.% and nickel to about 10 at.%. The *T-x* diagram of the Fe-Ni system, however, does not show a eutectoid point under ordinary pressure (*fig. 21a*). It has been suggested that the meteorites, in a certain period of their history, were exposed to high pressures (and temperatures) — e.g. at the moment of impact on earth — and that the eutectoidal structure was then formed. Kaufman and Ringwood have worked out diagrams which lend support to this theory [5]. According to their calculations, a eutectoid would exist at extremely high pressure (above about 60 000 atm).

The main reason for this is thought to lie in the relative expansion of the γ phase upon mixing Fe and Ni. (Precisely the opposite of the "contraction" of the liquid phase in fig. 20.) Hence the expectation that the higher the pressure the greater the tendency will be towards immiscibility of the γ phase. According to the calculations mentioned, the *T-x* diagram at 50 000 atm would be as shown in fig. 21*b*. The miscibility gap (dashed line) under these circumstances is not yet stable. It should also be noted that, compared with diagram *a*, the transition point from the α to the γ-phase has dropped from roughly 900 to 650 °C. This is explained by the fact that the volume of γFe is smaller than that of αFe. This is of the utmost importance to the theory we are now considering. For of course it means that an increase of pressure cuts both ways: not only does the (metastable) miscibility gap shift to higher temperatures, but also the upper limit of the two-phase domain $\alpha + \gamma$ drops. For example, at 100 000 atm a diagram as in fig. 21*c* can be worked out. This diagram indeed offers a possible explanation for the eutectoidal structure of the meteorites concerned. One should qualify this statement, however, by saying that, owing to the approximations that Kaufman *et al.* had to use, the diagram cannot be expected to be quantitatively correct.

The fact that meteorites have at some time been under extremely high pressure (and temperature) — whether or not upon collision with the earth — appears likely, completely independent of the previous argument, from the discovery of small diamonds in meteorites.

Solubility under varying temperature

To consider how the solubility varies with *temperature* we can apply the same procedure as for pressure, which resulted in fig. 20, with the difference of course that we must now take the entropies into account. The problem here too is made more general by considering equilibria between liquid mixtures and *mixed crystals* of *A* and *B* (i.e. instead of the pure component *A*). Since now both the solid and the liquid phases can occur at different mixing ratios of *A* and *B*, we must know the *S-x* curves of *both* phases. We shall take as our starting point the schematic diagram in *fig. 22.* For the entropy of mixing of both the solid and liquid phases we have taken Gibbs' equation (eq. 9). The change in entropy between the two phases of the pure components (entropy of fusion) is assumed to be identical, and roughly 1.1 R. This value is a good average for normal metals (Richards' Rule).

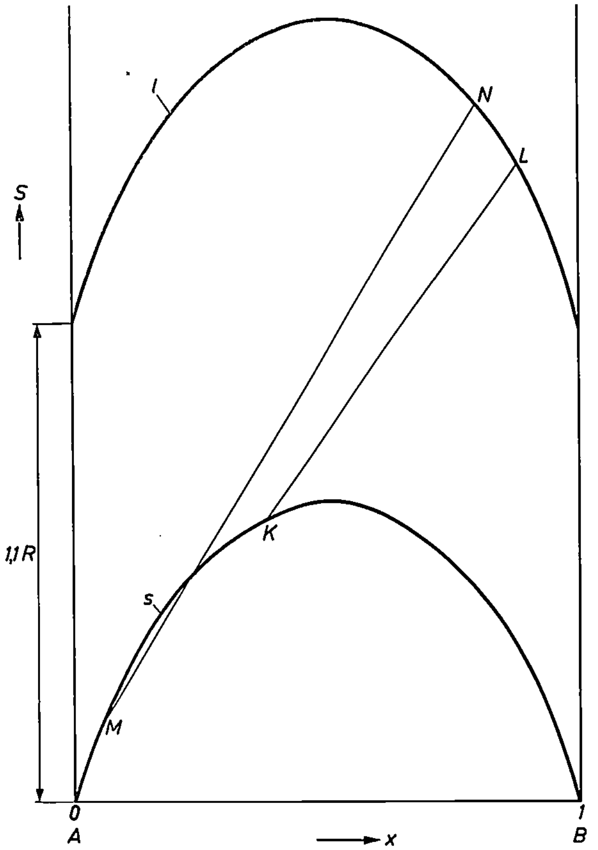


Fig. 22. Idealized *S-x* diagram for binary alloys of normal metals. l refers to the liquid phase, s to the solid phase. Two examples of solid/liquid equilibrium mixtures are considered: *K* in equilibrium with *L*, and *M* in equilibrium with *N*. In the text it is shown that at the eutectic temperature the first kind of equilibrium corresponds to a *T-x* diagram as in fig. 23*a* and the second to a *T-x* diagram as in fig. 23*b*.

In the diagram we consider first the (heterogeneous) equilibrium represented by the line *KL*, that is mixed crystals having the same composition and entropy denoted by point *K*, and a liquid mixture with the composition and entropy pertaining to *L*.

The choice of these points for our example was rather arbitrary. This does *not* imply that they are indefinite

[5] L. Kaufman and A. E. Ringwood, Acta metall. 9, 941, 1961.

for a given system or that it would not be possible to determine them by experiment. We shall return to this point on page 25.

For the same concentrations the *KL* line referred to lies higher than the curve of the homogeneous solid phase. Now we know that increasing the temperature promotes the formation of the situation with the greatest entropy and so when the temperature of the saturated mixed crystal (*K*) is raised the liquid phase forms, and, since the liquid phase contains less *A* than the mixed crystal, the mixed crystal phase is thereby enriched with *A*. We have thus described the "normal" behaviour to which a *T-x* diagram as in *fig. 23a* conforms: the relevant solidus curve bends continuously to the left.

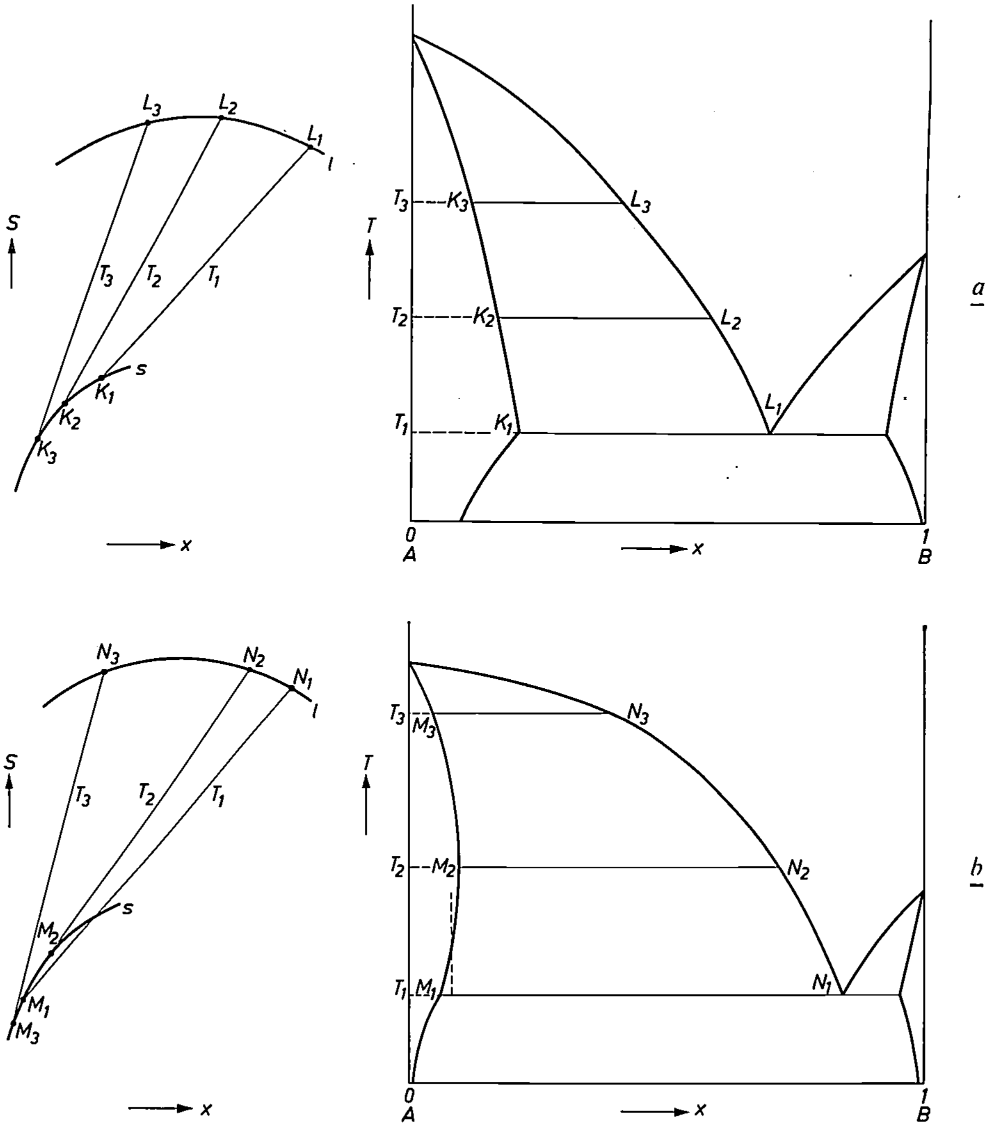


Fig. 23. Two eutectic diagrams differing essentially in regard to the shape of the solidus curve on the left; *a*) a "normal" solidus curve, *b*) a retrograde solidus curve. The situation of the equilibrium lines *KL* and *MN* in the *S-x* diagram of fig. 22 is schematically indicated on the left for three temperatures T_1, T_2 and T_3. In *a* both *K* and *L* move monotonically to the left. In *b*, however, *M* moves initially to the right, until the line *MN* is tangent to the lower curve, after which its movement resembles that in *a*.

To make matters clearer, particularly in connection with the considerations to follow, we have drawn to the left of the *T-x* diagram a fragment of the *S-x* diagram, showing how the equilibrium line *KL* in the *S-x* diagram shifts upon a temperature increase from T_1 (the eutectic temperature) to T_3.

We shall now examine the equilibrium represented by the line *MN*. Moving on this line from *M* to the right we now arrive *below* the *S-x* curve of the homogeneous mixed-crystal phase. This means that an *increase* in the temperature of the saturated mixed crystal does *not* lead to the formation of the liquid phase (at least not directly). This now happens when the temperature is *reduced*. The solidus curve in this case, like the left-hand solidus curve in fig. 23*b*, first moves to the *right* with increasing temperature, then bends over at the temperature T_2, after which it proceeds in the "normal" way to the melting point. This is known as the "*retrograde solidus curve*".

Here again, the situation is represented by the equilibrium line in the S-x diagram. T_1 is again the eutectic temperature, and at the temperature T_2 the line MN is just *tangent* with the lower curve; from this temperature onwards the behaviour is "normal" once more.

Let us now consider another aspect. Suppose we had determined the position of MN in some way or other by experiment, but that it was done at a temperature *higher* than T_2. If we had used this as a basis for predicting whether or not the solidus curve would be retrograde, we should have arrived at the wrong conclusion. The lower the temperature the greater the chance of being below the — unknown — temperature T_2. In principle, then, we can best take as our basis the position of MN at the *eutectic temperature*. That we can bring about the unusual phenomenon of liquid separation by *cooling* in a system with a retrograde solidus curve, is at once apparent from the dashed line in the figure. The existence of systems having a retrograde solidus curve was predicted by van Laar [6] a quarter of a century before the first experimental example was greeted with astonishment and disbelief. In 1948 ten to twenty such systems were known. After the invention of the transistor in that same year, the number went up very considerably.

The advent of the transistor attracted attention to Ge and Si crystals in which an extremely small quantity of a foreign substance was dissolved. From fig. 22 it is evident that the occurrence of the retrograde solidus curves is promoted by a low solubility of B in solid A, while B is readily soluble in the liquid. This becomes even plainer when we remember that $\mathrm{d}S/\mathrm{d}x$ is infinite at $x = 0$. On these grounds we predicted that the phenomenon of the retrograde solidus curve would be observed more and more frequently as the methods of determining small solid solubilities became more refined [7]. This prediction was confirmed in particular by research on germanium and silicon.

We have not yet discussed the method of determining the location of the equilibrium line in the S-x diagram, on the basis of which it is possible, as described, to predict whether or not the solidus curve will be retrograde. To do this we start from the liquidus curves and the equilibrium curves below the eutectic temperature in the T-x diagram. These can readily be determined by experiment, unlike the solidus curves, about which there is often considerable uncertainty (for reasons which will be dealt with in part II). With the aid of the first mentioned curves we can easily determine the composition of the various phases of a eutectic equilibrium. (This can be seen by considering the T-x diagram of fig. 23, and imagining that it does not contain the solidus curves.) Using the compositions obtained in this way we can roughly ascertain in the schematic S-x diagram in fig. 22 the corresponding point on the S curve of the solid and that on the S curve of the liquid phase, and thus by connecting these points find the equilibrium line of the relevant system at the eutectic temperature. It is fortunate that it is precisely at the eutectic temperature that we can establish the location of this equilibrium line.

For completeness we should add that the necessary knowledge of the composition of the phases of the relevant eutectic contains implicit data on the enthalpy of the system. That enthalpy and not only entropy data are needed can be understood from the fact that the solidus curve represents equilibrium states, for which the free enthalpy is essential. Enthalpy data vary considerably from one system to another and cannot therefore with advantage be schematized like entropy data. In the manner described however, it is relatively easy to arrive at the data needed.

We have just seen that liquid can separate from a mixed crystal upon cooling, and the same holds for *gas*. Since the change in entropy between gas and solid (entropy of sublimation) is greater than between liquid and solid (entropy of fusion) the implication is that the solubility of the gas should then be extremely small. If we consider the solubility of *hydrogen* in certain solid metals in equilibrium with hydrogen of 1 atm, we see that the solubility shows a marked tendency to decrease with decreasing temperature if it is roughly smaller than 0.05 at. %, and to increase if it is greater than 0.05 at. %. Here too, then, we find that, in accordance with our theory, the smaller the solubility the greater is the likelihood of "retrograde behaviour".

Having thus given some insight into the manner in which qualitative predictions can be made in respect of binary systems, we shall present in part II of this article, appearing in the next number, a more quantitative approach to binary systems.

[6] J. J. van Laar, Z. phys. Chemie **63**, 216, 1908; **64**, 257, 1908.
[7] J. L. Meijering, Philips Res. Repts. **3**, 281, 1948.

Appendix

One should be beware of considering metastability as a state of minor significance. After all, most of the very numerous organic compounds are metastable: through conversions, giving rise to inorganic substances such as H_2O, CO_2 etc., the free energy decreases. For example, CS_2, a widely used organic solvent, is metastable with respect to graphite + sulphur. At room temperature, however, the decomposition rate of CS_2 is too low for this to be troublesome.

One can go a step further and say that in fact all elements are metastable with respect, perhaps, to iron. True, it is a big step from chemical transformations to nuclear transmutations, but the difference is not a fundamental one from the thermodynamic viewpoint. For the rest, however, we are not concerned in this article with nuclear transformation.

If the two components of binary systems cannot convert one to the other, as for instance in the Ag-Au system and in the $AgNO_3$-H_2O system, the G value of each of the two components (which is fixed except for a constant), can be chosen independently of each other. The *minima* of the G-curves in these figures cannot therefore have any physical significance — though the tangent points on these curves have.

The systems in which the components have the same chemical composition (isomers or polymers) are in a class on their own. Since in principle they are mutually convertible, it is necessary to reckon with a certain difference in the G values of the components. The system acetaldehyde-paraldehyde normally behaves like a binary system. Paraldehyde is a trimer of acetaldehyde. The addition of a drop of sulphuric acid catalyses the formation of an equilibrium between these two substances ($3C_2H_4O \leftrightarrows (C_2H_4O)_3$) and the mixtures behave like a one-component system with a sharp melting point and a sharp boiling point. In this case the minimum of the G-x curve of a particular phase is indeed significant: at the appertaining composition the phase is most stable, and it is here that the equilibrium has to be established. A system of this kind is called a pseudobinary system. There is no essential difference between a binary system of isomers which, in certain conditions, can behave as a unary system, and a "one-component" system which, in certain circumstances, behaves like a binary system. An example of the latter is hydrogen; at low temperatures the transformation orthohydrogen $\leftrightarrows$ parahydrogen is considerably slowed down. Another example is water. It has long been known that it can be regarded as consisting of at least two kinds of molecules, but because they change one to the other very quickly in normal conditions, water behaves nevertheless as a unary system.

Where transformations in certain conditions are too slow for unary behaviour and too fast for binary, it is no longer readily possible to apply thermodynamic principles and phase theory in the normal way.

An interesting example of a pseudobinary system

In a transition between two modifications one is used to thinking in terms of modifications with a different crystal structure. Remarkably enough, this does not always turn out to be correct. In the case of cerium a transition point between two face-centred cubic modifications has been found at 150 °K. The modifications differ fairly substantially (about 6%) in their lattice constant. In addition to this intriguing fact there is a second one. Investigations by Ponyatovskii and by Beecroft and Swenson make it seem very likely that the transition line between both modifications ends at a critical point in the region of 300 °C and 12 000 atm [8]. This is also unexpected because by a critical point we are accustomed to think purely of liquid and vapour.

What seems puzzling at first sight becomes comprehensible if we consider cerium not as a unary but as a pseudobinary system, consisting of two kinds of Ce atoms. The transition from one kind to the other corresponds to the transition of an electron from the 4f shell to the 5d shell, making the atom smaller. It has to be assumed that the atomic transition is extremely fast, and that a state of equilibrium is always quickly established. At a certain p and T the G-x curve can have the shape of e.g. the second curve from the bottom in fig. 17. But we cannot put the beginning and end point of the curve at an equal height, because the atoms are transformed one into the other. Since this takes place rapidly, the double tangent line plays no part at all, and the *lowest* G minimum gives the concentration at which the equilibrium sets in. Let us assume that the component with the smallest atoms is on the right (*fig. 24*). When the pressure is raised the right end-poind of the curve will drop in relation to the left end-point, and the minimum (on the left) will shift to the right.

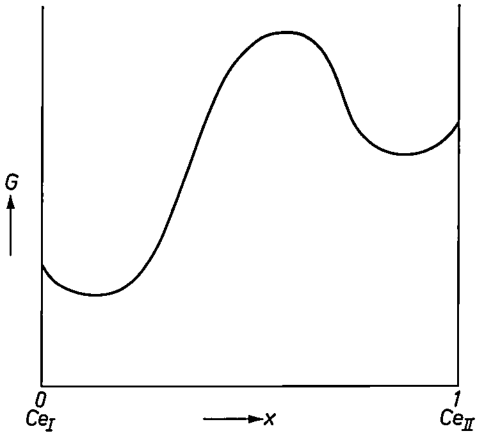


Fig. 24. Explaining the behaviour of cerium. Ce$_I$ consists of larger atoms than Ce$_{II}$.

This increase in the equilibrium concentration of small atoms under increasing pressure does not proceed continuously: at a given pressure there are two G minima which are equally low. This marks the transition line. At a somewhat higher pressure the minimum on the *right* denotes the most stable state and the equilibrium establishes itself at the appertaining concentration.

If, however, the temperature is so high that there are no bends in the G-x curve (cf. upper curve in fig. 17) the change can then only take place continuously; we are then above the critical temperature.

[8] E. G. Ponyatovskii, Doklady Akad. Nauk S.S.S.R. **120**, 1021, 1958; R. I. Beecroft and C. A. Swenson, Phys. Chem. Solids **15**, 234, 1960.

Summary. This treatment of unary and binary systems is intended primarily as an introduction to the subject. Special attention is devoted to the derivation of phase diagrams from the change of the Gibbs' free energy. Other subjects dealt with are metastability, saturated vapour pressure as a measure of stability, immiscibility, the equilibrium criterion for binary systems, the relation between chemical potentials and partial pressures, solubility under changing pressure and under changing temperature, a theory on the origin of eutectoidal structures in meteorites, and finally van Laar's prediction in 1908 of the *retrograde solidus curve*, which has since been convincingly established by experiments, especially on transistor materials.

Cryogenic production of ultra-pure hydrogen

S. Shaievitz

661.96:621.593

In recent years the industrial demand for ultra-pure hydrogen has shown a considerable increase. The following article describes a unit for economically producing hydrogen in an ultra-pure state. This unit enables the relatively small consumer to reap benefits derived from conventional large scale processing.

Introduction

With the growth of the electronics industry and the requirements of certain metal fabricators, the demand for hydrogen in an ultra-pure form (99.995%) has shown substantial expansion in recent years. The application generally involved is the brazing of components in an atmosphere that must exclude oxygen, water vapor, hydrocarbons, and nitrogen. For instance, electrical and magnetic properties of materials associated with the manufacture of several electronic devices are quite susceptible to impurities of the order of 1 : 10^6 or less. As a result, there is a constant effort to upgrade all materials connected with the manufacture of these components. Other applications of ultra-pure hydrogen as a protecting atmosphere in the electronics industry include the production of silicon and other types of crystals, and the metallizing of ceramics.

Also, some specialized applications may be mentioned here. In the refractory metals field, where powder metallurgy techniques are employed, an ultra-pure hydrogen atmosphere appears to have a beneficial effect. Molybdenum rolling, a difficult operation, is said to be improved in an atmosphere of ultra-pure hydrogen. Certain metals such as titanium, columbium, molybdenum, tungsten, and their alloys form undesirable products with nitrogen and oxygen at elevated temperatures. Consequently, the hydrogen used in the heat treatment of these materials must be nearly free of impurities. Ultra-high-purity hydrogen atmospheres have been applied to the bright annealing of stainless steel and also appear to be useful in the production of tin plate [1].

For large-scale users, on-site generation using the steam reforming and shift converter process [2] is the most economical approach, and in terms of capacity it outranks other methods. Raw materials for this process, natural gas and steam, are often readily available and low priced.

In general, however, the electronics and metal heat-treating industries may be considered as small scale users of hydrogen (under 150 m^3 per hour at normal temperature and pressure). The capital costs of the steam reforming process are then prohibitive, and unless the consumer happens to be located adjacent to an electrolytic operation, he is dependent upon purchases of cylinder or trailer supplies. This method of supply is not very economical however, because of the costs involved in transporting, handling, and distributing relatively small quantities, and because cylinder gas may require additional processing before ultra-high purity can be realized.

The steam reforming process consists of two steps. In the first step natural gas (consisting principally of methane) together with steam is passed over a solid catalyst at a pressure of 20 atm and a temperature of 650-950 °C to produce hydrogen and carbon monoxide. In the second step, the carbon monoxide is further treated with steam to form additional hydrogen and carbon dioxide. This second step is commonly referred to as the "shift reaction". Usually the shift reaction is applied in three stages to reduce the carbon monoxide content to about 1 : 10^5 (10 p.p.m.).

An economical means of transporting and handling hydrogen for application on a relatively small scale is in the form of liquid anhydrous ammonia. Since large quantities of ammonia are produced the small user is, in effect, obtaining the advantages of large scale hydrogen production and the convenient low cost of shipping liquid anhydrous ammonia. It remains to dissociate the ammonia into a 75%H_2-25%N_2 mixture and to process this mixture to obtain the desired purity level. The first step presents no difficulty: ammonia dissociators are manufactured by several firms and are readily available. For the purification of the

S. Shaievitz, B.Chem.Eng., is a research worker at Philips laboratories, Briarcliff Manor, N.Y., U.S.A.

[1] A survey of techniques in which an atmosphere of ultra-pure hydrogen is used can be found e.g. in Chemical Week, May 19, 1962, pp. 104-121, particularly on p. 120 and 121.

[2] A description of the steam reforming process can be found in the article cited under [1]. Also, see Chem. Engng., August 7, 1961, pp. 62-64, or Brit. chem. Engng. **8**, 466-470, 1963.

gas obtained, a cryogenic process has been developed at Philips Laboratories, Briarcliff Manor, U.S.A., that lends itself to nearly automatic operation and, as a consequence, requires little operating labour. This process will now be described.

The cryogenic purification process

The new cryogenic purification process involves four steps, viz:
1) cooling of the gaseous N_2-H_2 mixture, resulting from the dissociation of ammonia, to approximately — 193 °C accompanied by the condensation of most of the nitrogen present in the mixture;
2) the separation of the vapor from the condensed liquid nitrogen;
3) the removal of the remainder of the nitrogen by adsorption at low temperatures; and
4) the isothermal regeneration of the adsorbers, by means of purging with low pressure hydrogen. We shall now give a detailed description of the process, following the flow sheet *fig. 1.*

At point *A* (at the left in the figure) dissociated ammonia is combined with recycled purge hydrogen and fed to the compressor *B* at a suction pressure of about 0.3 atm. Feed gas (24 %N_2 - 76 %H_2) is compressed to 60 atm, and, after removal of the compression heat in an aftercooler *C*, passes through one of two ammonia adsorbers *D*. These adsorbers serve to remove any traces of undissociated ammonia and moisture that might be present in the gas. The gas is cooled to a few degrees above its dew point by the outgoing gas streams in the main heat exchanger *E*. It is then brought to approximately — 193 °C by flowing through a condensing coil *F* which is immersed in the liquid nitrogen tank *G*. At the outlet of this coil the nitrogen concentration in the vapor has been reduced from 24 % to 7.5 % while the condensed liquid nitrogen contains approximately 12.5 % dissolved hydrogen. The vapor is separated from the liquid in the separator-surge vessel *H*. This vessel is immersed in the liquid nitrogen bath in order to prevent re-evaporation of the liquid nitrogen condensate which would result from heat leak. The vapor leaving *H* flows through one of two automatically operated reversing adsorbers *Ads* (in the situation sketched, adsorber *Ads 1*). In so doing, nearly all the nitrogen impurities are removed.

At the point *I* (top right) the purified hydrogen is split into two streams. The major portion passes through one of a set of dual adsorbers *K* which removes the trace quantities of nitrogen remaining in this stream. The effluent from these adsorbers may contain less than 1 : 10^5 (10 p.p.m.) of impurities. These final nitrogen adsorbers *K* are manually regenerated in the conventional way. Since the quantity of nitrogen to be adsorbed is rather small, the units may be conveniently sized for a daily cycle. Flowing through the heat exchanger *E*, the hydrogen product is rewarmed to nearly ambient temperature. A back pressure valve *V4* on the product line *L* maintains pressure in the plant.

The smaller fraction of the split hydrogen stream (30 % or less) is throttled to slightly above atmospheric pressure by valve *V1*. This stream is used to regenerate one of the two reversing adsorbers (adsorber *Ads* 2 in the situation sketched). Regeneration is accomplished isothermally: that is, regeneration takes place at the same temperature as adsorption and is effected by purging the adsorber bed with pure hydrogen at a reduced pressure. This method of regeneration is chosen because it offers the possibility of automatic adsorber operation and consequently of greatly reducing the cost-price of the hydrogen produced. Of course, automatic continuous operation imposes a certain requirement to the ratio between the flow rates of the two streams (see Appendix).

Immersing the adsorbers in the liquid nitrogen bath serves two purposes. One function of the bath is to remove the heat of adsorption. Allowing the operating adsorber (*Ads 1* in the situation sketched) to warm up would result in a decrease in its adsorptive capacity. The other function of the nitrogen bath is to remove the refrigeration ("heat" of desorption) produced in the adsorber being purged. Cooling of this adsorber would result in a decrease in its ability to be purged by low pressure gas. Purge gas leaving adsorber *Ads 2*, containing hydrogen and the nitrogen swept off the bed, is warmed in the main heat exchanger *E* and recycled to the compressor.

Liquid nitrogen is continuously removed from the separator-surge vessel *H* and is automatically throttled (valve *V-2*) into the liquid nitrogen tank *G*. A portion of the liquid nitrogen and all of the dissolved hydrogen flash to a vapor. These vapors, plus the nitrogen boil-off, leave the tank. A gas refrigerating machine *M* recondenses a portion of the nitrogen vapor and returns it as a liquid to the tank. The remaining vapors are warmed in the main heat exchanger and are discharged from the plant. One of the objects of this process is to realize a high hydrogen recovery. The only hydrogen lost is that dissolved in the liquid nitrogen during the condensation step at elevated pressure (60 atm). Hydrogen recovery is in the vicinity of 95 %. *Fig. 2* shows a pilot plant built during the development.

Choice of pressure

The most critical process parameter to be chosen is perhaps the compressor head pressure. Several factors — besides the increased operating and fabrication costs at higher pressures — must be considered. *Fig. 3*

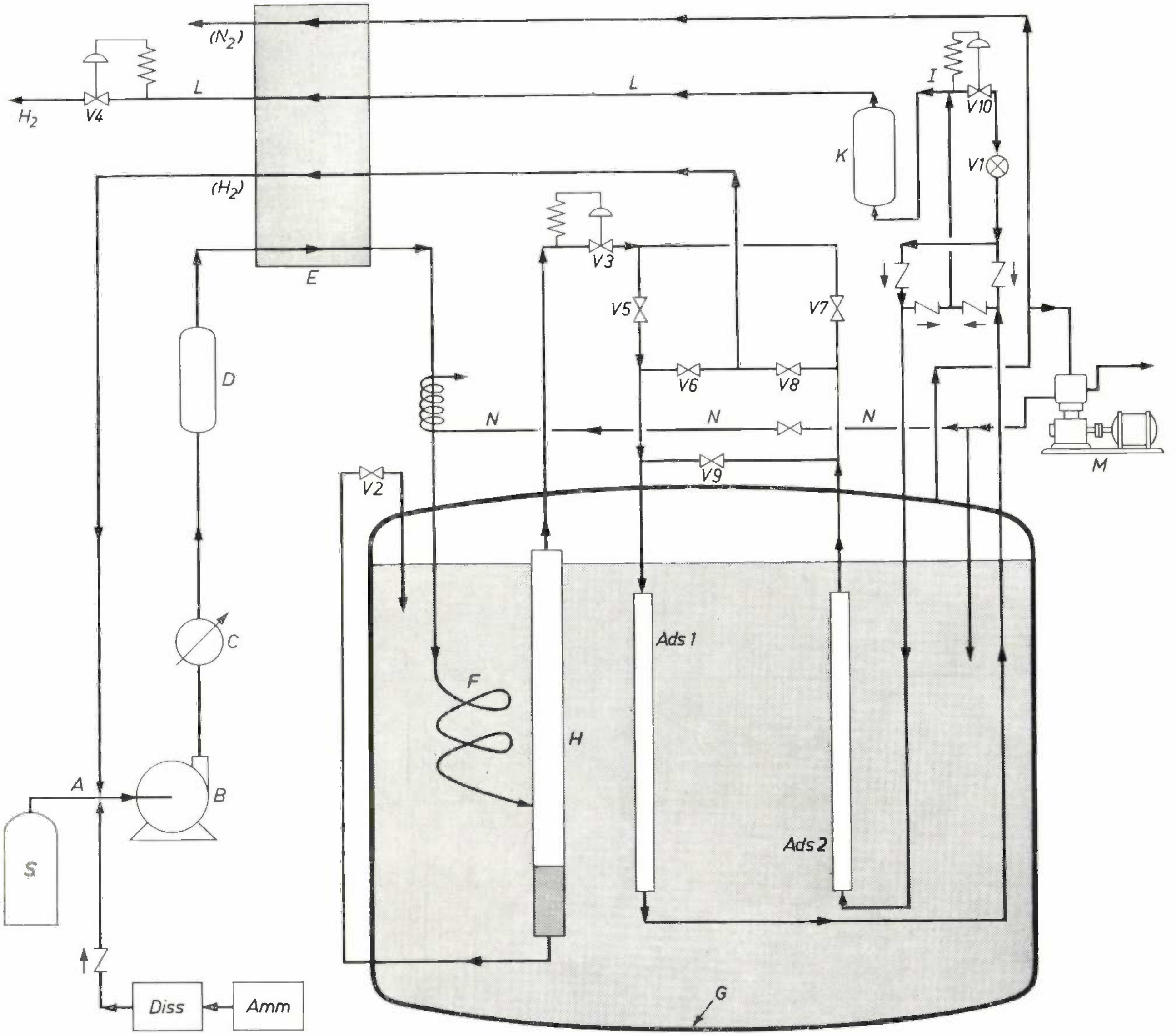


Fig. 1. Flow sheet of generator for ultra-pure hydrogen. *Amm* ammonia storage. *Diss* ammonia dissociator. *B* compressor; compresses the gas mixture to 60 atmospheres. *C* aftercooler. *D* pair of NH_3 adsorbers, operating in turn. *E* main heat exchanger. *F* condensing coil. *G* liquid nitrogen tank. *H* separator and surge vessel. *Ads 1* and *Ads 2* reversing main nitrogen adsorbers, operating in turn. *I* branching point of gas stream. *K* pair of small final nitrogen adsorbers. *L* product line. *M* gas refrigerating machine. The greater part of the liquid nitrogen produced by *M* is fed back to the liquid nitrogen tank, the remainder flows through line *N*, cools the gas mixture before entering the coil *F*, and then, in gaseous form, passes through the heat insulating container (the "cold box") in which the cryogenic equipment is housed. In this way condensables present in atmospheric air that could decrease the heat insulation are excluded from the interior of the cold box. *S* surge tank. *V* valves. — *V1* expansion valve; *V2, 3, 4* and *10* back pressure valves; *V5-9* solenoid-operated switch-valves, electrically controlled by cycle timer (not drawn) for automatic switching of the main nitrogen adsorbers.

Not drawn are two liquid level sensors; one of them controls the liquid level in the tank by starting and stopping *M*, the other opens *V2* when the liquid level in *H* has reached a certain height.

relates vapor and liquid equilibrium compositions of the hydrogen-nitrogen system at 80 °K as a function of pressure. As shown, the hydrogen concentration in the vapor phase increases rapidly with pressure at low pressures until it reaches about 90 % at 20 atm. It remains between 90 % and 93 % until 110 atm, and falls off steadily with increased pressure until, at 182 atm, the critical pressure is reached. The hydrogen concentration is then 55 %.

If the removal of nitrogen were the only factor governing the choice of pressure, a value of about 25 atm would be satisfactory since operation at higher pressures has only a slightly beneficial result. However, as is discussed in the Appendix, the performance of the revers-

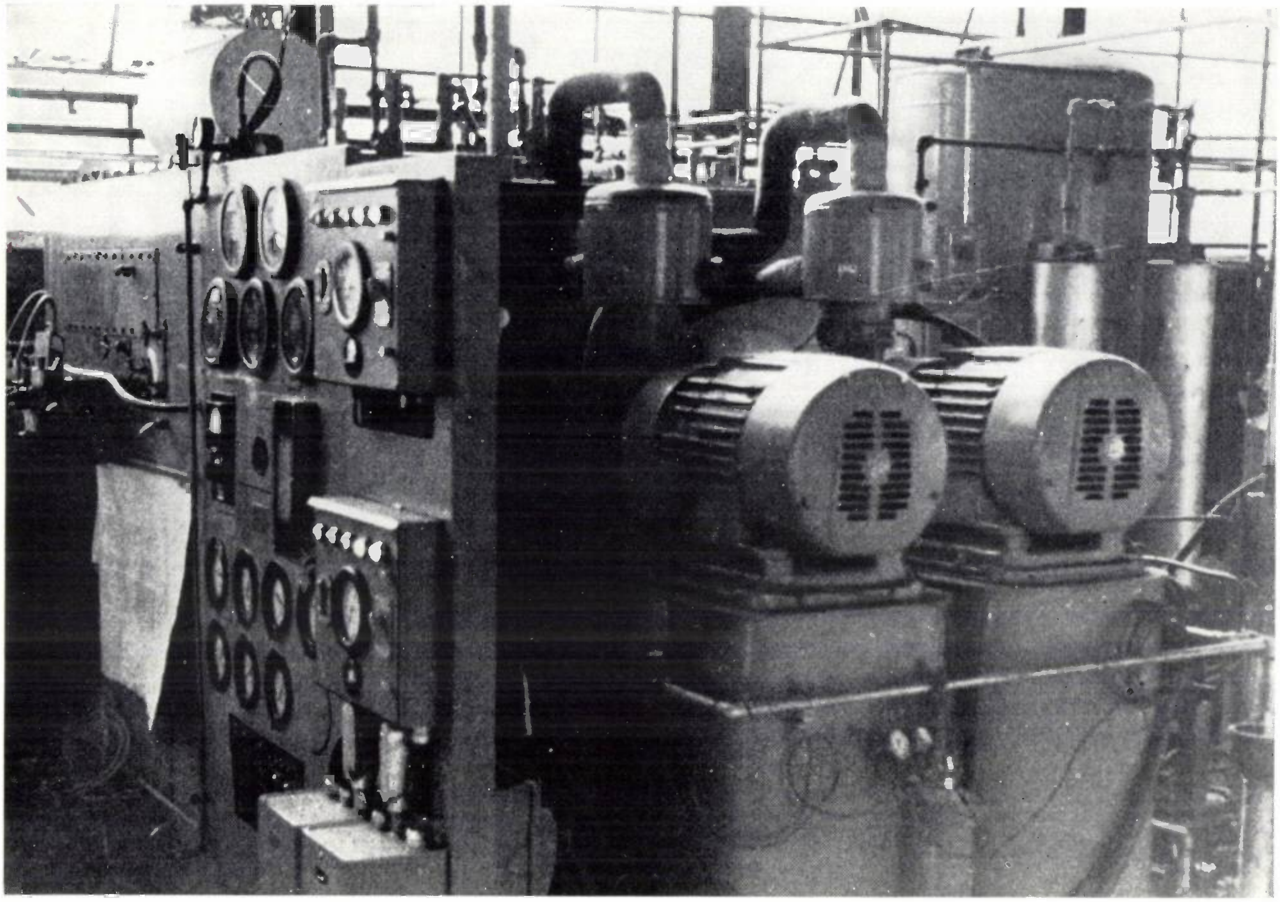


Fig. 2. Photograph of pilot unit [3]. For *M* in fig. 1 two gas refrigerating machines (Norelco "Cryogenerator" [*]) were used which are visible in the foreground.

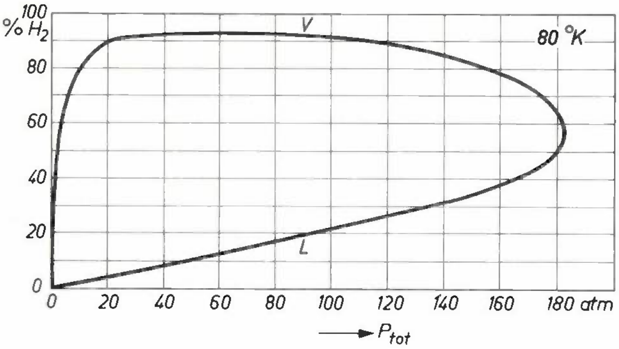


Fig. 3. Equilibrium composition vs. pressure diagram for a mixture of H_2 and N_2 at 80 °K. *V* vapor part, *L* liquid part of the curve.

ing adsorbers is dependent upon the ratio of adsorbing and desorbing pressures: a minimum ratio is necessary for steady operation and higher ratios enhance product purity. On the other hand there are several factors that set a limit to the increase of this pressure ratio. As fig. 3 indicates, the hydrogen content of the vapor phase falls off rather rapidly above 100 atm. Furthermore, the concentration of hydrogen dissolved in liquid nitrogen increases steadily with pressure. The latter effect represents a loss of product and a diminution of recovery. Increased compression power is another factor that must be considered when determining the adsorption-desorption pressure ratio. In view of the above factors, an operating pressure of 60 atm has been selected.

The switching of the adsorbers

The method by which the adsorbers are automatically switched merits discussion. Assume that adsorber *Ads 1* is at compressor head pressure (60 atm) and that adsorber *Ads 2* has been purged at low pressure (1.2 atm). Referring to the flow sheet, the switch valves are in the following position: *V5* (open), *V6* (closed), *V7* (closed), *V8* (open), and *V9* (closed). Switching of the adsorbers occurs in two steps: 1) equalization of pressures, and 2) pressure build-up and blow-down. During the first step, valves *V5*, *V6*, *V7* and *V8* are closed, and *V9* is opened. This permits adsorbers *Ads 1* and *Ads 2* to equalize in pressure. During the equalization step, pressure is being built up in the vessel *H* since the compressor continues to operate. As can be derived from fig. 3, it is desirable to minimize this build-up in pressure since the equilibrium nitrogen concentration in the vapor increases with pressure above 60 atm. Thus, the vessel *H* is also designed to act as a

[*] Norelco "Cryogenerator" is a registered trademark of North American Philips Co. Inc.

surge vessel and thereby prevents excessive pressure build-up.

The second step in the switching operation is placing adsorber *Ads 2* on the adsorption part of the cycle and adsorber *Ads 1* on the desorption part of the cycle. To accomplish this, switch valves *V6* and *V7* are opened, and *V5*, *V8* and *V9* are closed. At this instant, adsorber *Ads 1* and adsorber *Ads 2* are both at the same pressure, viz, about one-half of the head pressure. Adsorber *Ads 1* then proceeds to blow down while the pressure in adsorber *Ads 2* builds up to operating pressure. A surge tank installed on the compressor suction (*S* in fig. 1) prevents excessive pressure build-up while adsorber *Ads 1* is blowing down. A check valve prevents back flow into the ammonia dissociator.

Between the separator-surge vessel *H* and the switch valves is a back pressure valve (*V3*). The function of this valve is to maintain the separator at or above normal operating pressure. Were this valve not present, the reversing of the switch valves would cause the separator-surge vessel to rapidly blow down from its elevated pressure to the equalized pressure of the adsorbers. As seen from fig. 3, this would result in some flashing of the liquid nitrogen within the separator and a higher nitrogen vapor concentration. In addition, foaming may result from the sudden effervescence of the liquid nitrogen. This in turn would lead to liquid carry-over to the adsorbers and cause fouling of the adsorbers.

The back pressure valve *V10* has two functions. During adsorber switch-over it prevents the on-stream adsorber (*Ads 2*) from blowing down while it is in the process of pressure build-up. It also prevents the flow of gas to the other adsorber (*Ads 1*) which is being blown down, prior to being purged. The level in the liquid nitrogen tank is automatically regulated by a liquid level sensor (not drawn in fig. 1) which controls on-off operation of the refrigerator.

The pilot plant built during the development of the hydrogen generator and illustrated in fig. 2, was dissimilar to the ultimate design in that a somewhat different method was employed for purging the reversing adsorbers, and the final nitrogen adsorbers were not included. Nevertheless a product purity greater than 99.92% was realized in this pilot plant.

Appendix: Calculation of the required purge flow rate

As mentioned, for economical reasons the removal of adsorbed nitrogen from the adsorbers *Ads 1* and *Ads 2* is effected by the decompression method — rather than by the more conventional heating of the adsorber bed — to enable automatic adsorber operation. We shall now examine the purge flow rate necessary to clean an adsorber within the time interval during which the other can continuously be in operation.

The starting point of our consideration is the formula:

$$y = \varepsilon \frac{P_0}{P_{\text{tot}}}, \quad \dots\dots\dots \quad (1)$$

in which y is the mole fraction of nitrogen in vapor phase, P_0 the vapor pressure of pure nitrogen at the temperature in question, P_{tot} the total pressure of the gas mixture and ε a proportionality factor. A formula of this form can be used for the description of a binary liquid-vapor system as well as for the adsorption of one component of a binary gas mixture [4]. In the former case the factor ε is the ratio of the mole fraction of condensable component (here N_2) actually present in the vapour, to that predicted for ideal behavior, so it then describes to what extent the behavior of the mixture deviates from that of an ideal mixture; for the latter $\varepsilon = 1$. When used for the description of an adsorption process, the factor ε will be provided with an index a.

Assuming that the nitrogen concentration in the product is negligible compared to the inlet concentration, the condition for steady state operation of the reversing adsorbers — quantity purged = quantity adsorbed in the same period of time — can be expressed by the formula:

$$y_{\text{I}} W_{\text{I}} = y_{\text{II}} W_{\text{II}}. \quad \dots\dots\dots \quad (2)$$

In this formula:

W_{I} = mass flow during adsorption (moles of mixture per hour),
W_{II} = mass flow during purging (moles of mixture per hour),
y_{I} = inlet concentration of nitrogen during adsorption (mole fraction),
y_{II} = outlet concentration of nitrogen during purging (mole fraction); y_{II} is assumed to be constant.

With the help of (1), when related to the adsorption process, it follows from formula (2) that:

$$\frac{W_{\text{II}}}{W_{\text{I}}} = \frac{(\varepsilon_{\text{a}}/P_{\text{tot}})_{\text{I}}}{(\varepsilon_{\text{a}}/P_{\text{tot}})_{\text{II}}}. \quad \dots\dots\dots \quad (3)$$

The value of the ratio $W_{\text{II}}/W_{\text{I}}$ can be calculated if we assume that the ratio of the ε_{a} values for the two pressures in question is the same as that of the corresponding ε values for a liquid-vapor system. As adsorption can be considered to be a type of condensation process, this assumption seems reasonable for making a rough estimate. Calculation of the ε values in question gives $\varepsilon(1 \text{ atm}) = 1$ and $\varepsilon(60 \text{ atm}) = 4.2$, leading to $W_{\text{II}}/W_{\text{I}} = 0.091$. Therefore a purge flow rate as low as about 9% of the flow during the adsorption part of the cycle is theoretically sufficient for steady state operation. Although this figure is only a rough estimate, its low value clearly illustrates the technical feasibility of the isothermal adsorption-desorption process.

[3] This unit, built by Cryogenerators Division of North American Philips Co. Inc., for Hamler Industries Inc., was in operation for some time at the Commercial Steel Treating Co., Detroit, Mich. (U.S.A.).

[4] See M. J. Hiza and A. J. Kidnay, Advances in cryogenic engineering **8**, 174-182, 1963.

Summary. An economically operating unit has been developed which is capable of producing ultra-pure hydrogen (99.995%) at a rate up to 300 m^3 (NTP) per hour. The design of this hydrogen generator evolved from experimental work undertaken by Philips Laboratories, Briarcliff Manor, U.S.A., and from operating experience gained from a pilot installation. The process consists of two main steps, viz, 1) the dissociation of ammonia into hydrogen and nitrogen, and 2) the separation of nitrogen from the gas mixture. This separation is mainly performed by cooling the mixture to 80 °K at a pressure of 60 atm. A liquid consisting mainly of nitrogen and a vapor consisting mainly of hydrogen are then obtained. For the final purification adsorbers are used, viz, two reversing adsorbers contained in the liquid nitrogen tank and an additional pair in the product line. When one of the reversing adsorbers is in operation (at 60 atm), the other is isothermally purged at low pressure and vice versa. The other adsorbers are regenerated in the conventional manner.

Recent scientific publications by the staff of the Philips laboratories and factories

Reprints of those papers not marked with an asterisk * can be obtained free of charge from the Philips Research Laboratories, Eindhoven, Netherlands, where a limited number of reprints are available for distribution. Requests need only quote the reprint number given in italics at the end of each entry.

W. Albers, C. Haas and H. J. Vink: Quenching effects and the determination of the existence region of semiconducting compounds.
Philips Res. Repts **18**, 372-376, 1963 (No. 4). *R 486*

E. Andrich: Anwendung von PTC-Widerständen. (Application of PTC thermistors; in German.)
Elektron. Rdsch. **17**, 132, 135, 136, 138, 1963 (No. 3). *A 73*

V. Belevitch: Factorization of scattering matrices with applications to passive-network synthesis.
Philips Res. Repts **18**, 275-317, 1963 (No. 4). *R 481*

G. Blasse: New type of superexchange in the spinel structure; some magnetic properties of oxides Me^{2+}Co$_2$O$_4$ and Me^{2+}Rh$_2$O$_4$ with spinel structure.
Philips Res. Repts **18**, 383-392, 1963 (No. 5). *R 488*

G. Blasse: Magnetic properties of oxidic spinels containing tetravalent manganese.
Philips Res. Repts **18**, 400-404, 1963 (No. 5). *R 490*

G. Blasse: Magnetic-garnet phases containing pentavalent antimony.
Philips Res. Repts **19**, 68-72, 1964 (No. 1). *R 500*

G. Blasse and **J.F. Fast:** Néel temperatures of some antiferromagnetic oxides with spinel structure.
Philips Res. Repts **18**, 393-399, 1963 (No. 5). *R 489*

G. Brouwer: Graphical analysis of stationary surface injection phenomena in semiconductors.
Philips Res. Repts **18**, 432-446, 1963 (No. 5). *R 493*

G. P. Brouwer: Brightness variations on tungsten ribbon.
Philips Res. Repts **18**, 361-371, 1963 (No. 4). *R 485*

S. Duinker and **J. A. Geurst:** Long-wavelength response of magnetic reproducing heads with rounded outer edges.
Philips Res. Repts **19**, 1-28, 1964 (No. 1). *R 496*

R. A. Ford, E. Kauer, A. Rabenau and **D. A. Brown:** The electronic states of octahedral and tetrahedral Mn^{++} in α, β and γ manganous sulphide.
Ber. Bunsenges. phys. Chemie **67**, 460-465, 1963 (No. 5). *A 76*

A. H. Gomes de Mesquita, C. Langereis and **J. I. Leenhouts:** The structure of NbSn$_2$.
Philips Res. Repts **18**, 377-382, 1963 (No. 5). *R 487*

N. Hansen and **W. Littmann:** Automatisches Gerät zur Bestimmung der Oberflächengröße feinteiliger Substanzen. (Automatic device for determining the surface area of powders; in German.)
Z. Instrumentenk. **71**, 153-159, 1963 (No. 6). *A 77*

W. F. Knippenberg: Growth phenomena in silicon carbide. (Thesis Leiden, June 1963.)
Philips Res. Repts **18**, 161-274, 1963 (No. 3). *R 480*

J. L. Meijering: Miscibility gaps in ferromagnetic alloy systems.
Philips Res. Repts **18**, 318-330, 1963 (No. 4). *R 482*

B. Okkerse: Consecutive Laue and Bragg reflexions in the same perfect crystal.
Philips Res. Repts **18**, 413-431, 1963 (No. 5). *R 492*

H. G. Reik: Derivation of Holstein's expression for the high temperature mobility of small polarons from Kubo's formula.
Physics Letters **5**, 236-237, 1963 (No. 4). *A 72*

E. Roeder and **M. Klerk:** Untersuchungen mit dem Elektronenstrahl-Mikroanalysator an druckgesintertem Tantalkarbid mit geringem Mangan- und Nickelzusatz. (Investigations on hot-pressed tantalum carbide containing small additives of manganese and nickel, by means of electron-probe X-ray emission microanalysis; in German.)
Z. Metallk. **54**, 462-470, 1963 (No. 8). *A 75*

K. J. Schmidt-Tiedemann: Tensor theory of the conductivity of warm electrons in cubic semiconductors.
Philips Res. Repts **18**, 338-360, 1963 (No. 4). *R 484*

P. Schnabel: Four-point method for measuring the anisotropy of resistivity.
Philips Res. Repts **19**, 43-52, 1964 (No. 1). *R 498*

J. Schröder: Apparatus for determining the thermal conductivity of solids in the temperature range from 20 to 200 °C.
Rev. sci. Instr. **34**, 615-621, 1963 (No. 6). *A 74*

F. Sellberg: Plasma resonance in a germanium rod.
Philips Res. Repts **19**, 53-67, 1964 (No. 1). *R 499*

N. C. de Troye: Some properties of symmetric switching functions.
Philips Res. Repts **18**, 331-337, 1963 (No. 4). *R 483*

J. L. Verster: On the use of gauzes in electron optics. (Thesis Delft, Oct. 1963.)
Philips Res. Repts **18**, 465-605, 1963 (No. 6). *R 495*

J. Verweel: Permeability of dense nickel-zinc ferrites in polarizing fields.
Philips Res. Repts. **19**, 29-42, 1964 (No. 1). *R 497*

K. J. de Vos: Refined replicating technique for studying structural changes in alnico-type permanent-magnet alloys.
Philips Res. Repts **18**, 405-412, 1963 (No. 5). *R 491*

W. J. Witteman and **Th. Werkman:** Cylindrical high-pressure apparatus.
Philips Res. Repts **18**, 447-463, 1963 (No. 5). *R 494*

Volume 26, 1965, No. 1 pages 1-32 Published 14th January 1965

PHILIPS TECHNICAL REVIEW

VOLUME 26, 1965, No. 2

Generating light with selective thermal radiators

E. Kauer

621.32:535-1

The close link between the making of new materials and the development of modern electrical engineering has frequently been emphasized. The following article brings out this relationship clearly, and shows that the systematic search for certain combinations of material properties — or demonstrating that they are fundamentally impossible — entails profound research based on solid state theory.

Introduction

The thermal radiator as a light source was already known to prehistoric man, in the form of the open fire. The first new light sources of our technical age — gaslight and the electric lamp — were also thermal radiators. Both the latter are still used on a very wide scale, even though "cold" light sources have meanwhile been developed which have a considerably higher efficiency than thermal radiators.

But have the potentialities of the thermal radiator already been completely exhausted?

In the history of lighting engineering there has been no lack of attempts to improve the efficiency of thermal light-sources. It was very soon realized that the energy losses of an incandescent lamp mainly consisted of infra-red radiation, and efforts have been made to minimize these losses. This amounts to trying to make incandescent filaments that radiate selectively in the visible spectrum. With this end in view extensive research was carried out in the years from 1920 to 1934 into the emission properties of oxides having a high melting point, especially on ceramics, but in spite of the wealth of experimental data obtained, these investigations failed to produce a complete picture, and gave no indication of the direction in which one could hope to achieve essential improvements.

In dealing once again with the generation of light by selective thermal radiators, we do so because it seems justified for two reasons: firstly, advances in the field of solid state physics promise to provide a deeper insight into the problem, and, secondly, investigations of single crystals nowadays offer better defined data than could previously be obtained from investigations on polycrystalline materials.

Physical principles

The spectral energy distribution of a black body is given by Planck's radiation law:

$$E_0(\lambda,T) = \frac{c_1\lambda^{-5}}{\exp\left(\frac{c_2}{\lambda T}\right) - 1}, \quad . \ . \ . \ . \ (1)$$

with the two constants

$$c_1 = 2hc^2 = 1.19 \times 10^{-16} \text{ Wm}^2,$$

$$c_2 = \frac{hc}{k} = 1.438 \times 10^{-2} \text{ m degrees}$$

(h is Planck's constant, c the velocity of light, and k Boltzmann's constant).

The energy distribution of a radiating body can be written in the form:

$$E = \varepsilon E_0. \quad . \ . \ . \ . \ . \ (2)$$

The factor ε, the emissivity, depends in general on the wavelength λ, the temperature T and the angle Θ at which the energy is radiated. Kirchhoff's radiation law relates the emissivity of the radiating body to its optical properties. This law may simply be written:

$$\varepsilon(\lambda,T,\Theta) = A(\lambda,T,\Theta), \quad . \ . \ . \ (3)$$

where A is the absorption factor of the radiating body, i.e. the fraction of the incident radiation which is absorb-

Dr. E. Kauer, assistant director of the laboratory Aachen (Germany) of Philips Zentrallaboratorium GmbH.

ed by the body. It follows from Kirchhoff's law that the emissivity can never be greater than unity, and therefore the spectral emission of a radiating body is always below that of a black body at the same temperature. A radiating body is called non-selective when ε is independent of λ. Bodies whose emissivity have this wavelength independence are often referred to as "grey bodies".

The luminous efficiency of a radiator is defined as the ratio of the luminous flux radiated to the total power consumed. The luminous efficiency of a thermal radiator can thus be found from the formula:

$$\eta(T) = \frac{\int_0^\infty V(\lambda)\, \varepsilon(\lambda,T)\, E_0(\lambda,T)\, \mathrm{d}\lambda}{\int_0^\infty \varepsilon(\lambda,T)\, E_0(\lambda,T)\, \mathrm{d}\lambda\ (+E_\mathrm{v})} \quad . \quad . \quad (4)$$

Here $V(\lambda)$ is the relative luminous efficiency in lm/W, which reaches its maximum at 555 nm, i.e. about 682 lm/W. The term E_v in the denominator lumps together the losses due to heat conduction, convection, ballast, etc. As a rule the emissivity depends very little on the angle, and this dependence is therefore left out of the formula.

The efficiency of a thermal radiator that has a constant (or nearly constant) value of ε can be increased, as seen from equation (4), only by raising the temperature. If we calculate the efficiency as a function of temperature for a grey body we obtain a curve as shown in *fig. 1.* The curve has a maximum of 95 lm/W at a temperature of about 6000 °K, which is roughly the temperature of the radiating surface of the sun. Present-day gas-filled tungsten lamps operate at temperatures in the region of 2800 °K and deliver about 12 lm/W (a value applicable to a 100 W lamp); about 97% of the consumed power is radiated in the infra-red. The filament temperature cannot be raised any higher owing to the marked increase of evaporation, although the use of transport reactions, as for example the tungsten-iodine regenerative cycle, has opened up new prospects in this connection [1]. By means of this cyclical process the evaporated tungsten is returned to the filament, which can therefore be operated at a higher temperature than hitherto. But even using such cyclical processes a natural limit is ultimately reached, for there are no substances with a melting point higher than about 4000 °K. If one were able to make an incandescent lamp with substances possessing the highest melting points, such as TaC and ZrC ($T_\mathrm{m} \approx$ 4100 °K), and if these substances were to be operated at incandescence just below their melting point, e.g. at 3700 °K, the luminous efficiency obtained would be in the region of 50 lm/W. This looks like the upper limit for non-selective solid-state radiators.

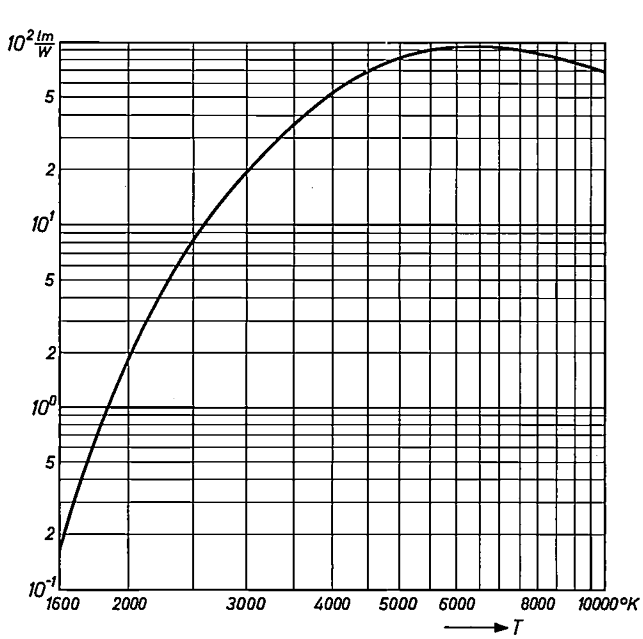


Fig. 1. Luminous efficiency of a grey (or black) body as a function of temperature T.

A considerable improvement in efficiency compared with that of the incandescent lamps nowadays manufactured might be achieved, according to equation (4), if it were possible to make $\varepsilon\ (\lambda,T)$ zero for all wavelengths that make no photometric contribution. The requirements are thus:

$$\varepsilon \approx 1 \text{ for } \lambda < 700 \text{ nm}, \quad . \ . \ . \ . \quad (5a)$$

$$\varepsilon \approx 0 \text{ for } \lambda > 700 \text{ nm}. \quad . \ . \ . \ . \quad (5b)$$

No separate demands need to be made for the ultraviolet region ($\lambda <$ 400 nm) since only a small fraction of the total energy is radiated in this region. If one could meet the conditions of (5) it would be possible at the operating temperatures of normal incandescent lamps to achieve luminous efficiencies in the neighbourhood of 200 lm/W, which is considerably higher than that of the best (non-thermal) light sources at the present time (sodium lamps 150 lm/W, fluorescent lamps 75 lm/W).

In the practical design of a selective radiator it is of the utmost importance to keep the emissivity in the infra-red as low as possible, particularly in the *near* infra-red. This is illustrated in *fig. 2,* which gives the luminous efficiency of radiators at T = 2000, 2500 and 3000 °K calculated on the assumption that the emissivity at wavelengths above 700 nm is not zero but has a constant value $\bar{\varepsilon}$, here plotted on the abscissa. It can be seen that for quite high values of luminous efficiency the emissivity in the infra-red should be less than 10^{-2}, particularly if the radiator is to be operated at a relatively low temperature.

In equation (5) we have taken 700 nm as the upper wavelength limit because in that case the resultant light

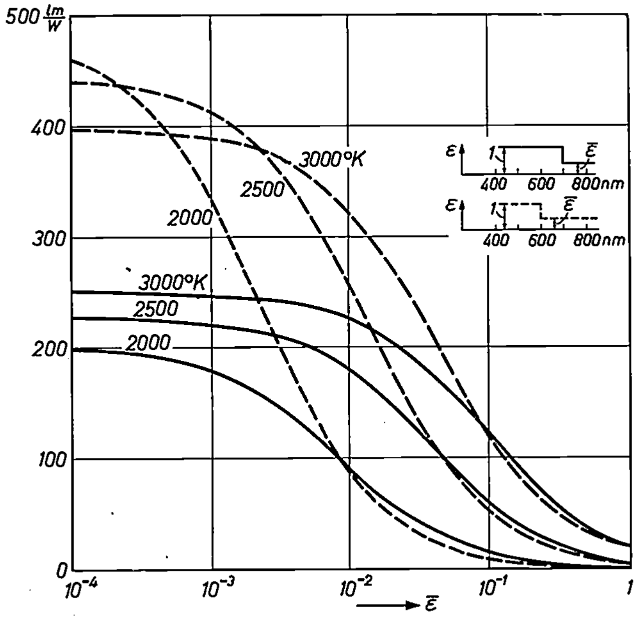


Fig. 2. Luminous efficiency of a selective radiator, with discontinuous emissivity, at three different temperatures, as a function of emissivity $\bar{\varepsilon}$ in the infra-red. For the solid curves, $\varepsilon = 1$ at $\lambda \leqq 700$ nm, for the dashed curves $\varepsilon = 1$ at $\lambda \leqq 600$ nm.

is very nearly white and gives good colour rendering. If, for the sake of a high efficiency, we are prepared to accept a poorer colour rendering, we can take a lower wavelength limit, e.g. 600 nm; we then obtain the upper set of curves in fig. 2. This results in luminous efficiencies higher than 400 lm/W, but again only on condition that $\bar{\varepsilon}$ is extremely small at wavelengths above the limit now chosen.

There are in principle two methods of meeting the requirements of equation (5); either to use a radiator which is "transparent" in the infra-red, or one which is highly reflective in the infra-red (metallic radiation). Given a plate of thickness d, the emissivity perpendicular to its surface, taking account of multiple internal reflections, is given by the general expression:

$$\varepsilon = \frac{(1-R)\,[1-\exp(-ad)]}{1-R\exp(-ad)}, \quad . \; . \; . \quad (6)$$

where the reflectivity R of a boundary surface is given by the relation:

$$R = \frac{(n-1)^2 + K^2}{(n+1)^2 + K^2}. \quad . \; . \; . \; . \quad (7)$$

Here a is the extinction coefficient (i.e. the absorption per unit of layer thickness) and $a = 4\pi K/\lambda$; n is the refractive index.

For a transparent plate we have $ad \ll 1$. It therefore follows from equation (6), if we expand the exponential function in a series and disregard the terms Rad and R, that

$$\varepsilon \approx ad. \quad . \; . \; . \; . \; . \; . \; . \quad (8)$$

This relation thus exists in the region $\lambda > 700$ nm for a selective radiator of the first kind, which is transparent in the infra-red or at least absorbs very little.

On the other hand, for a highly reflective (metallic) plate, where $R \approx 1$ and $ad \gg 1$ we find from equation (6):

$$\varepsilon = 1 - R. \quad . \; . \; . \; . \; . \; . \quad (9)$$

This relation, then, holds in the region $\lambda > 700$ nm for a selective radiator of the second kind, which reflects strongly in the infra-red.

Finally, there is the possibility, hitherto little noted, of influencing the energy distribution of a radiator by means of selective filters, in the sense that the filter passes the visible light unattenuated, but reflects the infra-red radiation back to the incandescent body where it is again absorbed. The power to be supplied to the incandescent body to make it radiate at a given temperature can then be proportionally lowered, so that the luminous efficiency is improved. This device is not of course limited to grey or black bodies, but can also be applied to the forms of selective radiator mentioned above, resulting in an additional improvement.

In the three sections following, the first will deal with the transparent selective radiator, the physical principles of which were recently reviewed at some length [2]. Here we shall be concerned only with a few fundamental considerations. We shall see that there is little prospect of developing a transparent selective radiator into a practical lamp. As far as the metallic selective radiator is concerned, however, the prospects in this connection are promising, as will be shown in the relevant section (page 42) and the same applies to the application of infra-red reflecting filters, dealt with in the last section (page 44).

The transparent selective radiator

The substances suitable as the basic material for a selective radiator which is transparent in the infra-red are characterized, as we have seen, by equations (5) and (8). These substances are insulators and semiconductors capable of withstanding temperatures in the region of 2500 °K. At sufficiently short waves, and hence for high-energy photons, substances of this kind always have a region of strong absorption. This absorption is due to electrons in the highest occupied energy band (valence band) transferring by optical excitation to the lowest energy band (conduction band); see *fig. 3*.

[1] See J. W. van Tijen and J. J. Balder, Iodine incandescent lamps, Philips tech. Rev. 23, 237-244, 1961/62.

[2] R. Groth and E. Kauer, Lichterzeugung mittels thermischer Selektivstrahler, Z. angew. Physik 16, 130-143, 1963.

If the substance is to be transparent in the infra-red, then the energy gap ΔE at the temperature of incandescence must be greater than 1.8 eV, a value which corresponds to a wavelength of 700 nm. In view of the fact that chemical bonds become looser with increasing temperature, the energy gap generally decreases with increasing temperature; the average temperature coefficient is -5×10^{-4} eV/degree. Taking this into account, we see that the energy gap at room temperature should be $\Delta E \geqq 2.8$ eV. At such a value of ΔE a semiconductor in the pure state is transparent to wavelengths in the entire visible spectrum, i.e. it is colourless (or white if a powder).

From fig. 2 we have already seen that high luminous efficiencies depend on the emissivity in the infra-red being lower than 10^{-2}, i.e. according to equation (8): $\alpha d \leqq 10^{-2}$. On the other hand, the substance is required to be a good radiator in the visible region, that is to say — apart from reflection losses — it should resemble a black body as closely as possible, the condition for which is $\alpha d \geqq 2$. By combining these two requirements, we see that the substance, irrespective of its thickness, must satisfy the relation:

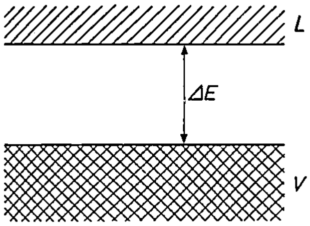


Fig. 3. Energy level diagram of a semiconductor. V valence band, in which at low temperatures all levels are occupied by electrons. L conduction band, where all levels are empty at low temperatures. ΔE energy gap, to which the absorption edge in the absorption spectrum corresponds.

$$\frac{\alpha_{\text{visible}}}{\alpha_{\text{infra-red}}} \geqq 2\times10^2. \quad . \quad . \quad . \quad . \quad (10)$$

It will be shown below that this condition is difficult to fulfil. The reason is that in the infra-red, at high temperatures, some absorption is hardly to be avoided, e.g. $\alpha = 10$ cm^{-1}, and according to equation (10) α should then be at least 2×10^3 cm^{-1} in the visible region. Absorption as strong as this is generally found only in the neighbourhood of the absorption edge ($h\nu \geqq \Delta E$); it is hardly to be brought about by doping a semiconductor, that is to say by introducing controlled amounts of absorbent foreign atoms in its lattice.

The various absorption processes governing the efficiency of a selective radiator will now be discussed in more detail with reference to the schematic diagram in *fig. 4*. The main processes are:

1) absorption due to lattice vibrations;
2) absorption due to free charge-carriers;
3) absorption due to lattice imperfections, including defects introduced by doping;
4) absorption near the absorption edge.

The first kind of absorption involves the interaction of electromagnetic waves with dipoles of the crystal lattice; the other kinds involve interaction with free or bound electrons in the solid.

The absorption processes (1) and (2) are fundamentally disadvantageous to the transparent selective radiator, whereas processes (3) and (4) can have a favourable effect due to "blackening" in the visible region.

1) *Absorption due to lattice vibrations*

The strong absorption at long waves found in all ionic crystals is due to lattice vibrations, that is to say vibrations of the sublattice of the positive ions in relation to those of the negative ions. Together with this absorption a strong metallic reflection occurs; certain ionic crystals are therefore often used for the purpose of isolating by reflection the long-wave region (the "residue" or "rest") of the spectrum. For this reason the absorption edge in this region is frequently referred to as the Reststrahlen (residual radiation) band.

According to the classical theory [3] the lattice vibrations of cubic crystals can be expected to have only one resonant frequency, viz:

$$\nu_0 = \frac{1}{2\pi}\sqrt{\frac{2r(M_1+M_2)}{M_1M_2}}, \quad . \quad . \quad . \quad (11)$$

where r is the binding force, and M_1 and M_2 the respective masses of the positive and negative ions. This frequency lies as a rule in the far infra-red. The short-wave "tails" of these Reststrahlen bands may extend into the near infra-red, however, and give

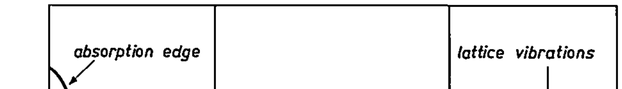


Fig. 4. Schematic representation of the spectral regions in which the various absorption mechanisms of a selective radiator operate. The absorption edge (extreme left) and the absorption due to lattice defects contribute to the absorption in the visible region and are therefore, in general, useful. The other absorption mechanisms are not.

[3] M. Born, Atomtheorie des festen Zustandes, Teubner, Leipzig 1923.

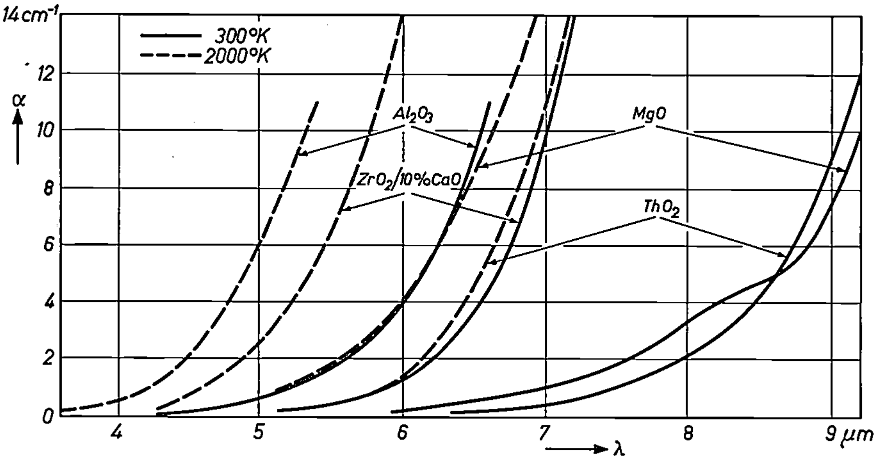


Fig 5 Short-wave "tail" of the absorption due to lattice vibrations in various high-melting oxides, at room temperature (solid curves) and at 2000 °K (extrapolated broken curves).

rise there to radiation losses which can no longer be disregarded in the energy balance. The adverse effect of these "tails" can be reduced by ensuring that the absorption bands are concentrated at the longest possible waves. According to equation (11) this is more readily achieved with substances composed of heavy ions and having a low binding force, i.e. a low melting point. This already brings us into conflict with our original aims, for of course the second condition is not compatible with giving the selective radiator a high operating temperature.

If $\varepsilon(\lambda,T)$ is known, that is to say the absorptivity as a function of wavelength and temperature, the influence of the lattice vibrations on luminous efficiency can be calculated from equation (4).

Theoretical statements regarding the wavelength dependence of the absorption due to lattice vibrations in the region of the short-wave tail are not easy to make. To do so, one would have to know exactly the complete spectrum of the lattice vibrations, and also have to do a great deal of computing work. Our only course is therefore to determine the absorption in all relevant cases by measurement, preferably up to the required operating temperatures.

Fig. 5 shows the results of such absorption measurements on various crystals at room temperature and at 2000 °K. The substances investigated, in order of apparent usefulness, were ThO_2, MgO, ZrO_2 with 10% CaO, and Al_2O_3. In all cases the absorptivities increased linearly with temperature. This enables us to indicate for the efficiency a limiting value governed by the lattice vibrations. For this purpose we assume that the emissivity in the visible region has been increased by suitable doping to $\varepsilon \approx 1$, without the doping giving rise to additional emission in the near infra-red. We therefore provisionally consider the near infra-red as free from absorption, and calculate now from equation (4) the efficiency of crystal wafers of three thicknesses, viz 0.2, 0.5 and 1 mm. It would obviously be advantageous to use even thinner wafers, but they are ruled out by being mechanically too weak. The result of the calculation for magnesium oxide (MgO) is represented by the solid curves in *fig. 6.* The luminous efficiency increases steeply with temperature. For the thinnest wafer it is only 33 lm/W at 2000 °K but as large as

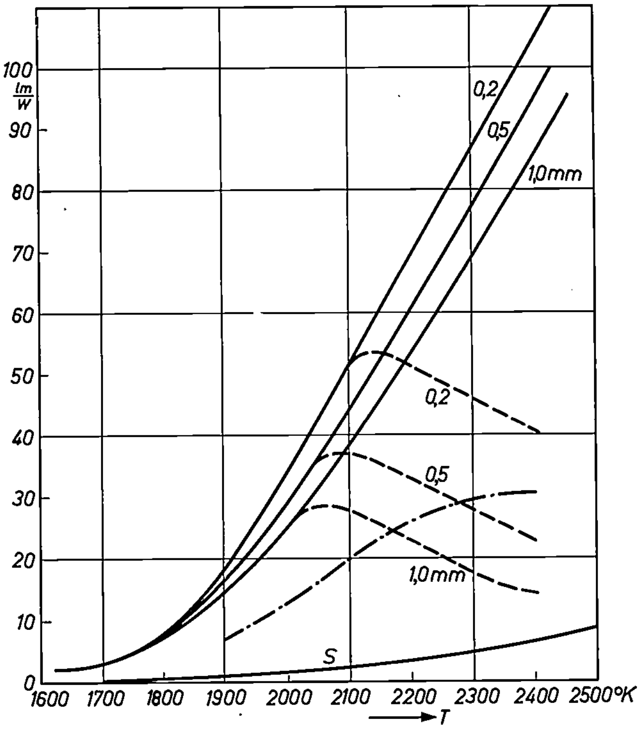


Fig. 6. Luminous efficiency of selective radiators of magnesium oxide MgO of differing thicknesses, as a function of temperature:
a) taking account of lattice vibrations (solid curves);
b) taking account of lattice vibrations + absorption by free charge-carriers (dashed curves);
c) taking account of lattice vibrations + absorption by free charge-carriers + convection losses (dot-dash curve).
The black-body curve (*S*) is included for comparison.

100 lm/W at 2400 °K. This substantial rise with temperature is due to the fact that the fraction of the visible radiation (with $\varepsilon \approx 1$) increases rapidly, whereas the emission due to lattice vibrations rises only linearly with temperature.

If we decide that the selective radiator should have a luminous efficiency at least equal to that of an incandescent lamp (roughly 15 lm/W), we can draw from fig. 6 another conclusion, of importance to our subsequent considerations, that operating temperatures below 1900 °K are ruled out because of the inadequate emission of visible light. We assume that the tail of the absorption due to lattice vibrations is no more favourable with other substances than with MgO, but this can be regarded as quite certain according to fig. 5 and equation (11). This conclusion sets a severe limit to the number of eligible substances. On the whole there remain only the oxides with a high melting point (ThO_2, HfO_2, ZrO_2, MgO, BeO) together with a few nitrides (BN, AlN) and carbides (SiC, Al_4C_3).

Although an efficiency of 100 lm/W is certainly attractive, it is clear that, as a result of the lattice vibrations alone, the values we have calculated for ideal conditions (see fig. 2) are reduced quite substantially. For MgO the situation is still relatively favourable, and the only better substance might perhaps be ThO_2. As regards SiC and BN the residual radiation bands lie at about 12 μm, that is to say at even shorter wavelengths than for the oxides. Little is known about the other high-melting compounds. There is not much hope, however, of finding among them any substances that will behave better as far as lattice vibrations are concerned.

2) *Absorption due to free charge-carriers*

We have not yet said anything about the manner in which the selective radiator has to be heated, although of course we have in mind electric heating, for even materials that are excellent insulators when cold become conductive at high temperatures. This conduction at high temperatures can have several causes. In substances with a medium energy gap (3 to 5 eV) considerable intrinsic conduction is already obtained at 2000 °K, since electrons are thermally excited from the valence band to the conduction band. In substances having a larger energy gap this intrinsic conduction is less significant than the conduction which is based on lattice imperfections, caused by thermal agitation. This conduction may be due to electrons alone as well as to both electrons and ions.

For our purposes it is now important that the free charge-carriers responsible for this conduction should also give rise to absorption of electromagnetic waves. This was first calculated by Drude [4] entirely on the basis of classical theory. In the region of wavelengths and temperatures in which we are interested ($kT/h\nu \geqq 0.2$; $\nu = \omega/2\pi$ = the frequency of light) the results of the classical theory agree very well with those of modern quantum-mechanical theory, provided that the mobility μ of the charge-carriers is governed by scattering from the thermal lattice vibrations [5]. At elevated temperatures this is always the case. For the extinction coefficient α the theory yields the equation:

$$\alpha = \frac{\gamma N e^2}{c m^* \varepsilon_0 n\,(\omega^2 + \gamma^2)}, \quad . \quad . \quad . \quad . \quad (12)$$

where $\gamma = e/\mu m^*$ is the damping constant. Here e is the elementary charge, N the charge-carrier density, ε_0 the dielectric constant of the vacuum and m^* the effective mass of the charge carriers [6]. For $\omega^2 \gg \gamma^2$, and introducing the DC conductivity $\sigma_0 = e\mu N$, we find:

$$\frac{\alpha}{\sigma_0} = \frac{e^2}{\varepsilon_0 c n (\mu m^*)^2 \omega^2}. \quad . \quad . \quad . \quad . \quad (13)$$

In order to minimize the ratio α/σ_0 (and we shall see presently why this is desirable) we need to look for substances with the highest possible value of the product μm^*. We have seen, however, that the choice is already limited by the melting temperature. For most eligible substances the condition $\omega^2 \gg \gamma^2$ is not fulfilled. Owing to the generally very low mobility, especially at elevated temperatures, the contrary is in fact found, $\gamma^2 \gg \omega^2$, and we have:

$$\frac{\alpha}{\sigma_0} = \frac{1}{\varepsilon_0 c n}. \quad . \quad . \quad . \quad . \quad . \quad (14)$$

This ratio is therefore independent of the value of the product μm^* and of the wavelength.

Let us again consider our model material, magnesium oxide. *Fig. 7* shows the infra-red absorption of non-

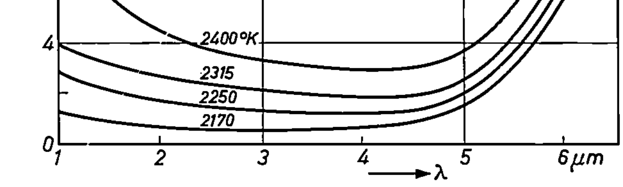


Fig. 7. Absorption α in the infra-red by free charge-carriers in undoped MgO crystals, as a function of wavelength, at various temperatures.

doped crystals of this material at high temperatures. In the wavelength range from 1 to 5 μm the absorption is fairly constant. Accordingly, the absorption is found to be proportional to the conductivity; the ratio α/σ_0, amounting to 240 Ω, is also in good agreement with the theoretically predicted value. The increase of the absorption at longer waves is due to the "tail" of the lattice vibrations, while at shorter waves and higher temperatures an absorption is found which is due to the lattice imperfections (absorption process 3).

The influence of free charge-carriers certainly makes itself felt. Taking it into account in eq. (4) results in the dashed curves in fig. 6, having a maximum of 52 lm/W at an operating temperature of about 2100 °K for the thinnest plate under consideration (0.2 mm). As a result of the exponential increase of the absorption by free charge-carriers, the efficiency drops again at higher temperatures.

As regards absorption by free carriers, silicon carbide is especially interesting. Unlike the other materials, in SiC the effective mass and mobilities of the carriers are sufficiently well-known. Since we are concerned with intrinsic conduction, an extrapolation of the values measured at 1300 °K is certainly permissible. We can therefore compute exactly the luminous efficiency as a function of the operating temperature and the layer thickness of the radiator. Since, however, an upper limit of 2000 °K is set to the operating temperature of the radiator, we can turn the problem round and ask what layer thickness is needed to obtain a given luminous efficiency. Calculation shows that at 2000 °K the layer thickness should be no more than 2 μm for a luminous efficiency of only 20 lm/W. A layer a few μm thick is ruled out on the grounds of mechanical strength alone, quite apart from the consequences of material losses due to evaporation.

Because of the disadvantageous effect of the free charge-carriers, it has frequently been suggested that substances might be used in which *ionic conduction* prevails. In view of the large effective mass of the ions, which in this case may be regarded as the free charge-carriers, one might expect a more favourable situation according to equation (12). In the case of ionic conduction, however, the advantage of the large effective mass is to a great extent cancelled by the disadvantage of the substantially lower mobility. Moreover, it is doubtful whether equation (12) still holds in the case of ionic conduction, which depends on thermally activated exchange processes among the lattice sites.

As long as we are concerned with intrinsic conduction, the absorption due to free charge-carriers cannot possibly be favourably influenced by doping. In the case of intrinsic conduction due to lattice imperfections, i.e. where there are deviations from the stoichiometric composition, then it is possible to vary the conductivity within certain limits, for instance by varying the partial oxygen pressure where oxides are concerned. There is not much room to manoeuvre, however, since the use of very high pressures is limited by practical considerations, and the use of very low pressures by the increasing dissociation.

3) and 4) *Absorption due to lattice defects and in the region of the absorption edge*

In all the foregoing calculations we have assumed that the condition $\varepsilon \approx 1$ can be fulfilled in the visible region without extra absorption taking place at the same time in the infra-red. This is only very rarely more or less in agreement with the facts. A case in point is α-SiC, which has an energy gap of 2.86 eV at room temperature, and is therefore virtually colourless in the undoped condition. When the temperature is raised, the absorption edge is shifted -5×10^{-4} eV/degree, i.e. the temperature coefficient mentioned on page 36, so that the crystal acquires colour. *Fig. 8* shows how the absorption edge shifts with temperature; at an operating temperature in the region of 2000 °K the edge lies at exactly the right place, i.e. at the boundary between the visible and the infra-red. Given sufficiently colourless SiC crystals, the change in the selectivity of the radiation associated with the shift of the absorption edge is plainly observable. Such crystals begin at about 1100 °K to radiate bluish light, which

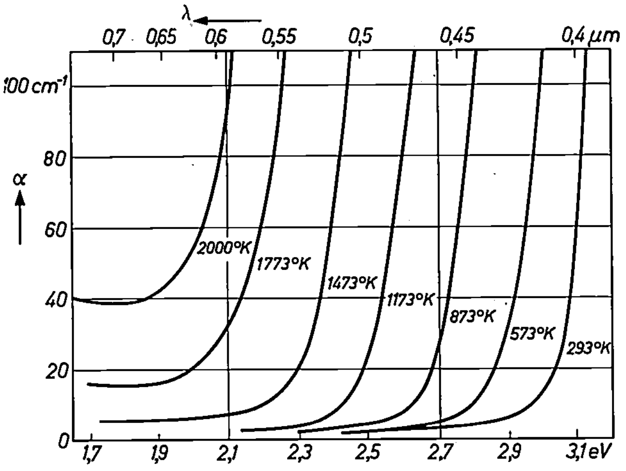


Fig. 8. Absorption edge of hexagonal silicon carbide crystals at various temperatures (the crystals were irradiated along the *c* axis). (From R. Groth and E. Kauer, Physica status solidi **1**, 445, 1961.)

[4] P. Drude, Phys. Z. **1**, 161, 1900.

[5] K. J. Planker and E. Kauer, Z. angew. Phys. **12**, 425, 1960.

[6] The influence of the periodic potential of the lattice on the movement of the free charge-carriers can be taken into account by assigning to them an effective mass which is in general lower than that of the free electron. For the exact definition of effective mass, see e.g. R. A. Smith, Semiconductors, Cambridge University Press 1961.

at higher temperatures changes through greenish-yellow to yellow (black body radiation). Similar phenomena were observed by Mollwo on ZnO crystals [7].

Fig. 9 shows a photograph of an incandescent SiC crystal whose thickness in the viewing direction is roughly 0.5 mm. The current passed through the wafer gives it a temperature of roughly 1400 °K, and the absorption edge (ΔE) as in fig. 8 is thereby brought roughly into the middle of the visible region. At photon energies $h\nu > \Delta E$, $\alpha d > 1$, whereas at $h\nu < \Delta E$, $\alpha d \ll 1$. Consequently, the crystal emits mainly in the green-blue part of the spectrum of a black body at 1400 °K. The crystal contains some impurities, however, which are already visible in the cold state as light-absorbent spots in the otherwise colourless crystal. Inside these spots there is also strong absorption where $h\nu < \Delta E$, i.e. $\alpha d \geqq 1$. The radiation from these spots therefore resembles very closely that of a black body at the temperature of the crystal, in this case 1400 °K, so that in the photograph in fig. 9 the spots are seen to be yellowish. If we now look at the narrow edge of the crystal, we see that the thickness in the viewing direction is much greater, i.e. 3 mm. Even with a crystal containing no impurities it is then no longer true that $\alpha d \ll 1$ for photon energies $h\nu < \Delta E$. The crystal thus radiates in that direction roughly like a black body. This effect, too, is visible in fig. 9, even though the photograph was taken looking at the broad edge of the crystal. This is due to the fact that the black radiation from the long path of 3 mm on the crystal surface is partly deflected into the viewing direction. Consequently the crystal edges are seen to be yellowish, like the defects in the interior of the crystal.

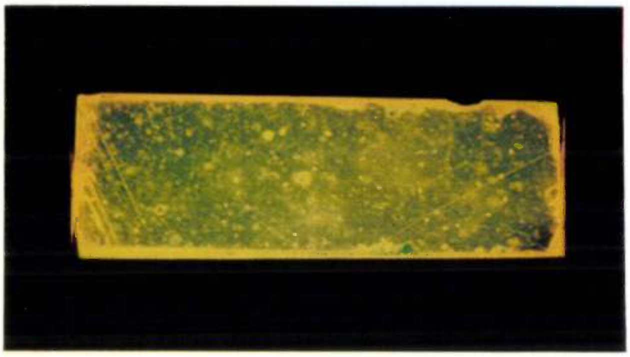


Fig. 9. Selective radiator of SiC at $T \approx 1400$ °K. Owing to the situation of the absorption edge (see fig. 8) the undisturbed parts of the crystal wafer emit greenish-blue radiation. The impure parts, however, radiate approximately like a black body and thus appear yellowish. Because of the diffraction of radiation from a much greater layer-thickness, the crystal edges are also yellowish in hue.

The fact that the absorption edge in SiC is so favourably situated is purely fortuitous. In most other cases it is necessary to bring about absorption in the visible region by doping with transition metals or rare earths. The absorption curve towards the long-wave side does not then show nearly such an abrupt discontinuity as in the case of an absorption edge. A classic example of this is the Auer mantel used for gaslight. It consists of a ThO_2 skeleton, which serves as heat reservoir, and contains cerium atoms which are situated at thorium lattice sites and which are responsible for the absorption in the visible region. *Fig. 10* gives a plot of the emissivity of the incandescent gas mantel as a function of wavelength for various cerium contents. The undoped mantel shows relatively low emission and scarcely any selectivity. Doping with cerium, which is optimum at roughly 0.5 %, causes the appearance of an absorption edge in the blue-green. At higher Ce percentages there is a marked increase of emissivity in the infra-red, almost completely destroying the favourable emissive properties of the gas mantel.

Attempts were made to produce incandescent bodies with the material used for the Auer gas mantel that could be electrically heated and thus employed for electric lamps. For various reasons these efforts failed to produce the desired result. In the first place, it proved impracticable to give the incandescent body a layer thickness small enough to obtain the intrinsically favourable properties of the Auer mantel. But even had it been possible to make the layer thin enough, the incandescent body at its normal burning temperature of 1800 °K would have had a luminous efficiency far below that of the incandescent electric lamps of today. With electric heating higher operating temperatures would be possible, but the situation in this case would not be significantly improved. Because the absorption by free charge-carriers increases with rising temperature, the initially low emissivity in the infra-red would rapidly increase. At the normal operating temperatures for electric lamps the Auer mantel would behave roughly like a black body.

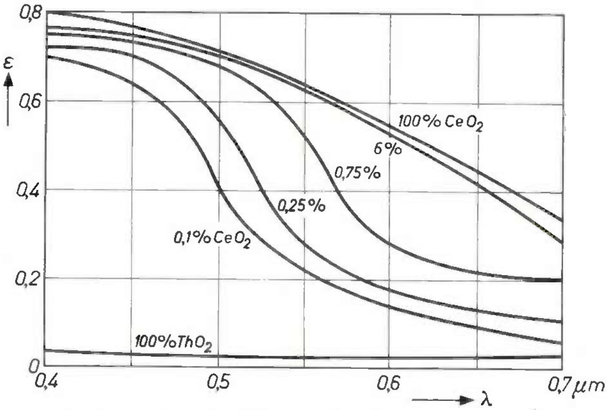


Fig. 10. Spectral emissivity ε of polycrystalline ThO_2-CeO_2 skeletons of varying composition.
(After H. E. Ives, E. T. Kingsburry and E. Karrer, J. Franklin Inst. **186**, 401, 585, 1918.)

[7] E. Mollwo, Z. angew. Phys. **6**, 257, 1954.

Vapour pressure and convection losses

So far we have solely been concerned with the *energy balance* of a selective radiator. When we turn to practical applications, then of course other factors are involved, such as the vapour pressure of the radiator and, in the same connection, its useful life. The considerations here are no different from those applicable to normal incandescent lamps. Since the output of visible light rises steeply with the temperature T, the latter should be as high as evaporation permits. The situation can best be understood by considering a material heated to incandescence in a vacuum. Between the evaporated quantity of material w (in grams) and the vapour pressure p (in torr) the following gas-kinetic relation holds:

$$p = \frac{17.14\, w}{At}\sqrt{\frac{T}{M}}, \quad \ldots . \quad (15)$$

where A is the surface area (in cm^2), t the burning time (in seconds) and M the molecular weight. With this equation we can easily estimate the maximum permissible vapour pressure, provided we introduce certain assumptions concerning the geometry of the radiator, the useful life required and the permissible loss of material by evaporation. Taking again a crystal wafer 0.2 mm thick, we specify a life of 1000 hours, which is normal for tungsten lamps, and we specify during this life-time a material loss of 10%; from eq. (15) we then find a vapour pressure of the order of 10^{-7} torr. If we know the vapour-pressure curve of the material, we can now use this value to read off the maximum permissible operating temperature. Unfortunately, our knowledge of the vapour pressures of high-melting oxides is still inadequate, and the results obtained by the various investigators show some marked discrepancies. *Fig. 11* gives the data relating to the vapour pressures of BeO, ZrO_2 and ThO_2; these substances have probably the lowest vapour pressures of all oxides. If we base our considerations on the vapour pressure of ThO_2, which is generally regarded as the most stable oxide in vacuo, we certainly cannot be accused of wishing to exaggerate the influence of the vapour pressure. But even with this oxide, and using the relatively low vapour-pressure values measured by Shapiro, we arrive with the above-mentioned assumptions at a maximum operating temperature of no more than 2000 °K. The luminous efficiency that can thus be obtained, taking into account the absorption due to lattice vibrations (see page 36) is far too low. For a selective radiator, then, burning in vacuo is entirely out of the question.

The evaporation can of course be counteracted by means of a gas atmosphere, making higher operating temperatures permissible. This, however, brings us up against the problem of convection losses. In the gas-filled tungsten lamp these losses are small compared with the radiation losses in the infra-red. With the selective radiator, on the other hand, the latter losses are largely suppressed and the operating temperature is lower, and therefore the convection losses assume decisive significance. Turning again to the case of magnesium oxide, to which the curves in fig. 6 relate, and taking into account the convection losses of an incandescent body in the shape of a strip 3 mm wide and 0.2 mm thick (the term E_v in eq. (4)), we obtain the result represented by the dot-dash curve in fig. 6. The maximum luminous efficiency now is no more than 30 lm/W at an operating temperature of roughly 2300 °K.

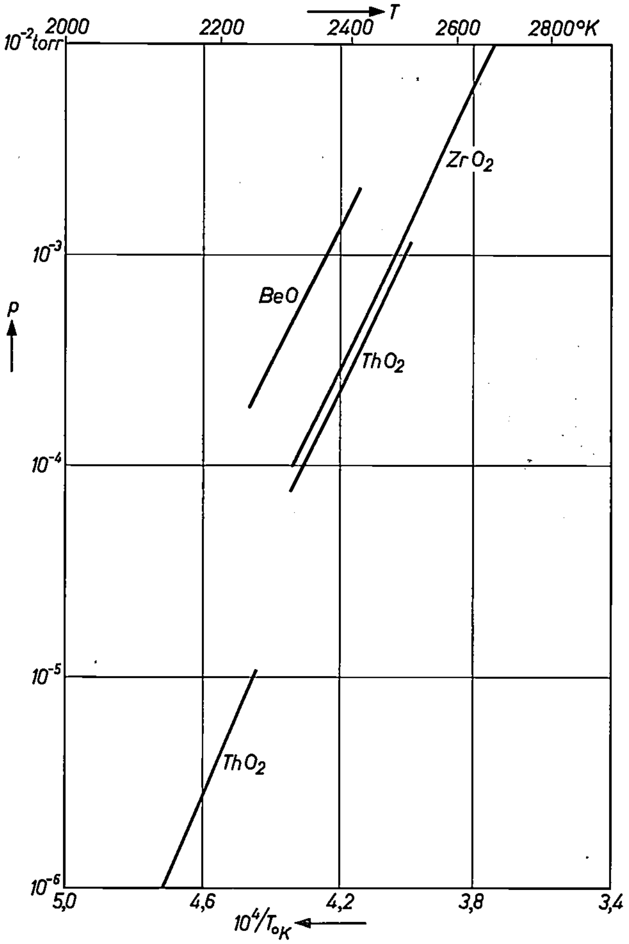


Fig. 11. Vapour pressure p of BeO, ZrO_2 and ThO_2 as a function of temperature.
BeO: after N. D. Erway and R. L. Seifert, J. Electrochem. Soc. **98**, 83, 1951.
ZrO_2: after M. M. Nakata, R. L. McKisson and B. D. Pollack, U. S. Atomic Energy Commission, NAA-SR-6095, 1961.
ThO_2, top curve: author's own measurements.
ThO_2, bottom curve: after E. Shapiro, J. Amer. Chem. Soc. **74**, 5233, 1952.

Summarizing, it can be said that using the materials examined here, which are transparent in the infra-red, and taking into account the technical difficulties involved, it is not possible to produce a selective radiator with a luminous efficiency higher than 30 lm/W. There are no grounds for hoping that better results might be obtained in this respect from a material not yet investigated or not yet known. The luminous efficiencies obtainable are just not high enough to justify the efforts and considerable technical difficulties involved in developing such a light source.

Metallic selective radiators

Among the metallic selective radiators we shall include, apart from metals proper, compounds which show metal-type conductivity and semiconductors which can be so highly doped as to produce a metallic reflection in the near infra-red. In general, owing to absorption by free-charge carriers, such substances will already be opaque in the entire visible and infra-red regions at layer thicknesses in the neighbourhood of 1 μm.

With selective radiators of this kind the reflectivity should ideally be $R = 0$ in the visible and $R = 1$ in the infra-red. The emissivity under the latter condition is, according to equation (9): $\varepsilon = 1 - R$. In the following we shall express the reflectivity R in terms of certain material constants, and we shall try to determine the circumstances in which the above-mentioned ideal case can best be approximated.

Absorbent media can be characterized by a complex refractive index n—jK, where K is related to the extinction coefficient α by the expression $\alpha = 4\pi K/\lambda$. As is mentioned above, for a medium containing free charge-carriers, Drude's theory [4] gives the following dispersion equations:

$$n^2 - K^2 = \varepsilon_g - \frac{Ne^2}{m^*\varepsilon_0(\omega^2+\gamma^2)}, \quad . \quad (16)$$

$$2nK\omega = \frac{\gamma Ne^2}{m^*\varepsilon_0(\omega^2+\gamma^2)} \quad . \quad . \quad . \quad . \quad (17)$$

Equation (17) can be reduced to equation (12). The symbols used in both equations have the same meaning; in particular γ is again the damping constant. The term ε_g in eq. (16) accounts for the polarizability of the bound charge carriers; in the absence of free charge-carriers, ε_g is the dielectric constant of the medium.

We can normalize equations (16) and (17) by introducing a characteristic frequency:

$$\omega_p = \sqrt{\frac{Ne^2}{\varepsilon_0\varepsilon_g m^*} - \gamma^2}, \quad . \quad . \quad . \quad (18)$$

which is defined by the condition $n^2 - K^2 = 0$ and is called the plasma frequency. (As a result of fluctuations in the distribution of the charge density, the plasma of free electrons can oscillate at a frequency given by equation (18); hence the name.) This results in more easily manageable mathematical expressions, with the dimensionless independent variable ω/ω_p and with γ/ω_p as the only parameter:

$$n^2 - K^2 = \varepsilon_g \left\{ 1 - \frac{1+\left(\frac{\gamma}{\omega_p}\right)^2}{\left(\frac{\omega}{\omega_p}\right)^2 + \left(\frac{\gamma}{\omega_p}\right)^2} \right\}, \quad . \quad (19)$$

$$2nK = \varepsilon_g \frac{\frac{\gamma}{\omega_p}\left\{1+\left(\frac{\gamma}{\omega_p}\right)^2\right\}}{\frac{\omega}{\omega_p}\left\{\left(\frac{\omega}{\omega_p}\right)^2+\left(\frac{\gamma}{\omega_p}\right)^2\right\}}. \quad . \quad . \quad (20)$$

Calculating n and K from these equations and then using these values to find the reflectivity from equation (7), we obtain the result presented in *fig. 12*. It can be seen that the slope of the curves, the steepness of which is a measure of the selectivity of the material depending on the free charge-carriers, is primarily governed by the ratio γ/ω_p. Plainly, this ratio should be as small as possible to obtain the most abrupt emissivity jump between the visible and the infra-red regions. The reflection edge, defined by $\lambda/\lambda_p = 1$, should lie at $\lambda_p = 0.7$ μm, which corresponds to a value $\omega_p = 2.7\times10^{15}$ s^{-1}. To estimate the charge carrier concentration required for this position, we work on the assumptions that $\gamma^2 \ll \omega^2$, $\varepsilon_g \approx 10$ and that the effective mass m^* is approximately equal to that of the free electron, m_0. With these data we find from equation (18): $N = 2.2\times10^{22}$ cm^{-3}. Since in general we can expect effective masses that are smaller than that of the free electron, this calculated concentration can be regarded as an upper limit. In nearly all metals, including tungsten, the charge-carrier concentration is higher than this value. Consequently their plasma frequencies are higher than 2.7×10^{15} s^{-1} and metallic reflection already exists in the visible region. Moreover, the damping constant γ of the high-melting metals, such as tungsten and molybdenum, is so high that the reflection curve is fairly flat. The curves corresponding to this case in fig. 12 are those with values of $\gamma/\omega_p > 1$. *Fig. 13* shows the emissivity of tungsten as a function of wavelength and temperature. In the infra-red especially, the curves are well defined by equations (16) and (17), provided the free parameters ε_g and m^* are suitably chosen. The temperature-dependence of the emissivity is primarily governed by the manner in which the damping constant or the mobility of the charge carriers varies with temperature. As can be seen from fig. 13, the selectivity of tungsten from the point of view of light generation is in fact poor. The emissivity in the visible region is roughly 0.45, whereas in the near infra-red, particularly at the

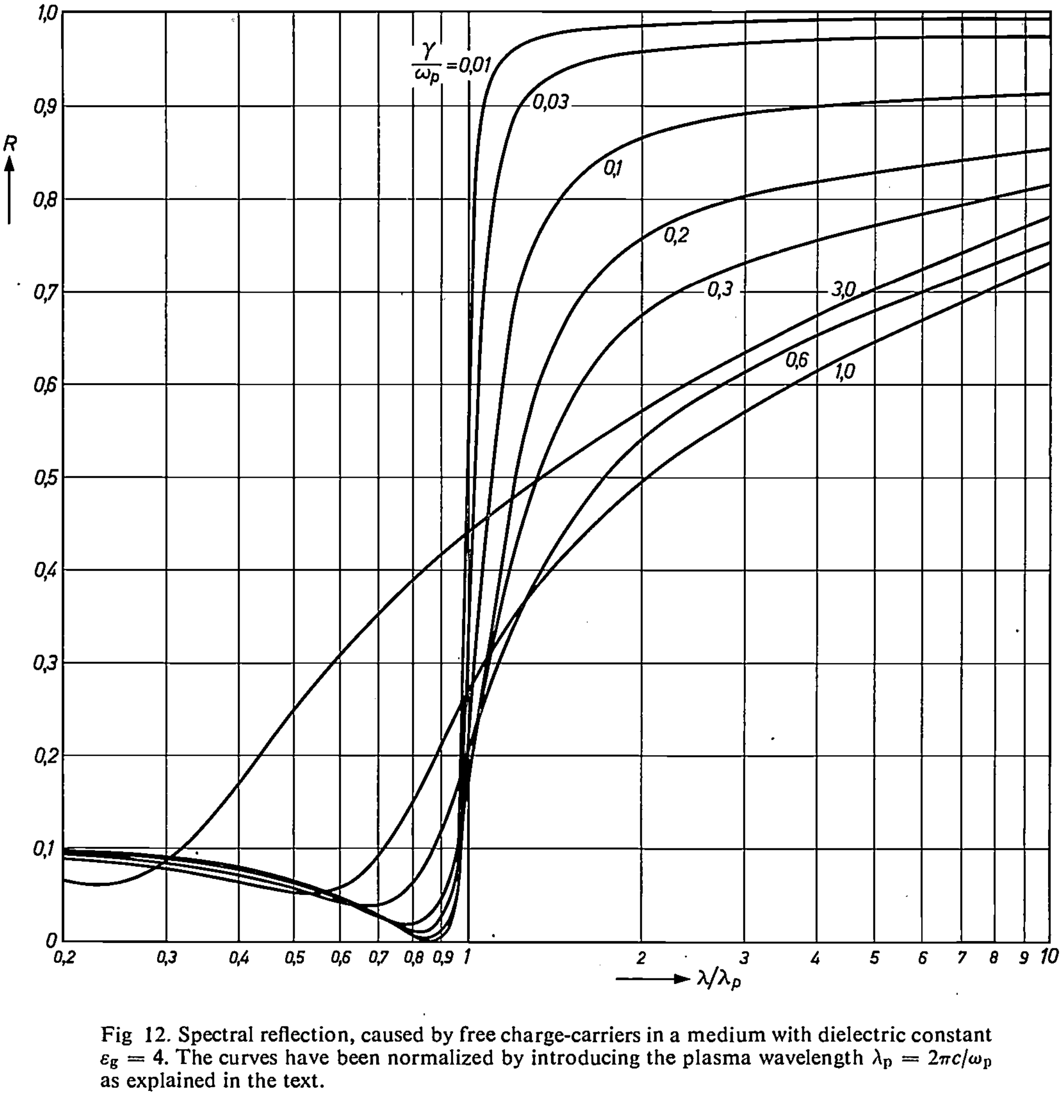


Fig 12. Spectral reflection, caused by free charge-carriers in a medium with dielectric constant $\varepsilon_g = 4$. The curves have been normalized by introducing the plasma wavelength $\lambda_p = 2\pi c/\omega_p$ as explained in the text.

wavelengths where the energy distribution of a radiator has a maximum at 2800 °K, the emissivity has dropped to only 0.35. The luminous efficiency of a tungsten radiator is therefore no more than about 30% higher than that of a black body at the same temperature.

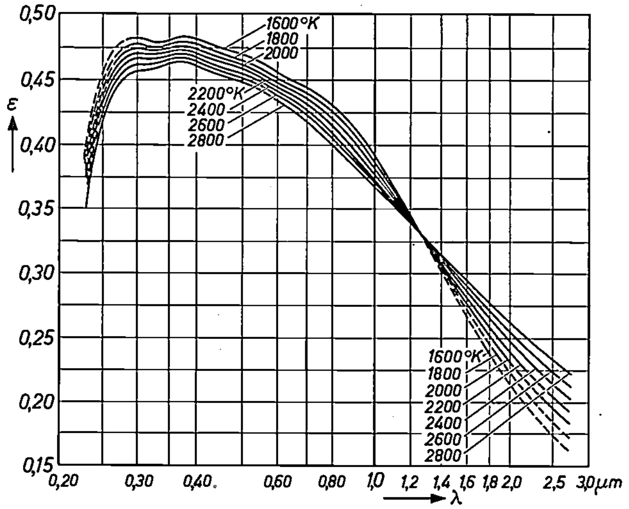


Fig. 13. Spectral emissivity ε of tungsten at various temperatures. (After J. C. de Vos, thesis Amsterdam, 1953.)

The use of tungsten for the filaments of incandescent lamps is therefore not based on the optical considerations discussed above, but rather on its very low rate of evaporation, which makes particularly high operating temperatures possible. Meanwhile, however, the regenerative processes referred to on page 34 are now being adopted in the manufacture of tungsten lamps, whereby the material that evaporates from the filament is returned to it. The first process of this type to find technical application is the tungsten/iodine process. It is not unlikely that the application of similar regenerative processes will make it possible to use filament materials that have a substantially higher vapour pressure than tungsten. Materials that have a higher selectivity than tungsten are therefore worth looking for.

By way of example, *fig. 14* shows the spectral reflectivity of lanthanum hexaboride (LaB_6). The

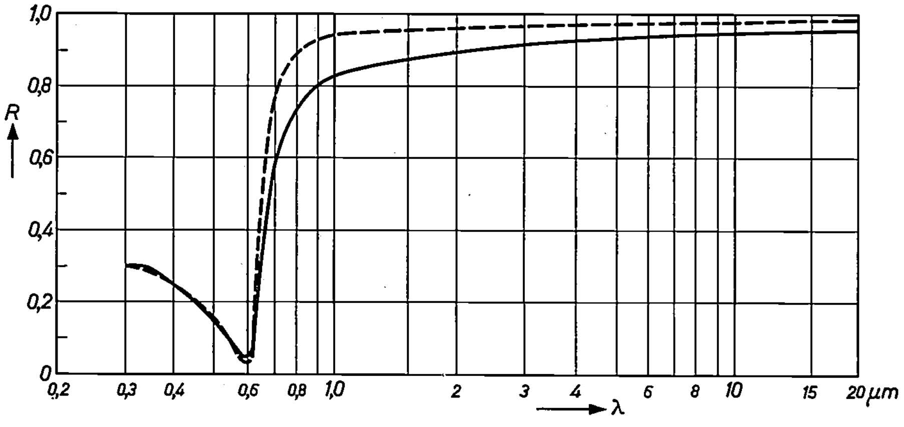


Fig. 14. Spectral reflectivity of LaB_6. Solid curve: measured values. Dashed curve: values calculated from equations (16) and (17).

reflection minimum at 0.6 μm and the steep rise of the curve towards longer wavelengths leave little doubt that we are concerned here with the reflection edge of the free electrons. Assuming in the first instance that the shape of the reflection curve remains roughly the same at higher temperatures, then the merits from the point of view of lighting engineering are obvious. In the visible region, especially within the region of maximum relative luminous efficiency for photopic vision, the emissivity is exceptionally high because of the low reflection and is almost equal to that of a black body. At the boundary of the infra-red the reflectivity increases steeply, and as a result the emissivity is already very low in the near infra-red.

To what material constants does LaB_6 owe these favourable properties? The conductivity of LaB_6, being $\sigma = 7.5 \times 10^4\ \Omega^{-1}\mathrm{cm}^{-1}$, is appreciably higher than that of most non-noble metals. This is not due, however, to a high charge-carrier concentration N of 1.45×10^{22} cm^{-3} that meets the condition formulated above with a view to the position of the reflection edge; the high conductivity is rather due to the high mobility of the carriers, which is $\mu = 32\ \mathrm{cm^2/Vs}$. Since each lanthanum atom contributes one electron to the conduction mechanism, LaB_6 constitutes what might be regarded as a metal diluted with boron atoms. Closer analysis shows that the dielectric constant ε_g in equation (16) has the value 15.6 and the effective mass $m^* = 0.32\, m_0$ [8]. With these data we find a damping constant $\gamma = 1.74 \times 10^{14}\ \mathrm{s}^{-1}$, a plasma frequency $\omega_p = 3.04 \times 10^{15}\ \mathrm{s}^{-1}$ and a ratio $\gamma/\omega_p \approx 0.06$. The favourable properties of LaB_6 are thus based on the fact that the plasma frequency is suitably situated, and that γ has a low value; in other words, according to equation (12), the product of mobility and effective mass has a high value.

Unfortunately, at higher temperatures LaB_6 loses to a large extent its favourable properties. As in the case of metals, the conductivity decreases as a result of the charge carriers being increasingly scattered by phonons. The temperature coefficient of the resistance in the temperature range between 300 °K and 1000 °K has a value 2.75×10^{-3} $°\mathrm{K}^{-1}$ and is mainly governed by the variation of mobility with temperature. If we extrapolate to operating temperatures of about 2300 °K we must take into account a mobility of only 5 $\mathrm{cm^2/Vs}$, i.e. a value $\gamma/\omega_p \approx 0.4$. As can be seen in fig. 12, the reflection curve will be much less steep, and this has a most adverse effect on the efficiency of the selective radiator. For this reason LaB_6 is not the ultimate answer to our problem. We shall rather have to look for the substance we want among those that have a substantially higher mobility at room temperature, e.g. more than 100 $\mathrm{cm^2/Vs}$, so that the inevitable drop in mobility when the temperature is raised will do less damage. As regards the possibility of finding substances of this kind there is good reason to be optimistic. There exists a series of hexaborides of the rare earths and of the alkaline-earth metals whose electrical properties tally with those outlined above; as yet, hardly any attention has been paid to their optical properties. Moreover there are numerous compounds having a high melting point whose conductivity almost rivals that of metals, but which have not yet been extensively investigated with a view to their suitability for use as selective radiators. If suitable substances can be found amongst these, it might be possible to use them in combination with appropriate regenerative cyclical processes, to develop a practical incandescent lamp with a luminous efficiency far higher than that of the tungsten incandescent lamp.

Application of selectively reflecting layers

Finally, we shall now examine the possibility of raising the luminous efficiency of a given radiator by controlling the spectral composition of the radiation with the aid of reflection filters. The object is to reflect the infra-red radiation back to the incandescent body while the visible light passes virtually unattenuated through the filter. In this way, depending on the quality of the filter, it should be possible to increase the efficiency of a light source quite considerably. We shall see, incidentally, that the application of this device is not confined to thermal radiators.

In the preceding sections it has been shown that the intrinsic difficulties encountered with the selective radiator concerns the requirements to be met by the materials at high temperatures. The use of selectively reflecting layers boils down to shifting the problems towards the properties of the materials at much lower temperatures.

We consider a spherically symmetrical or infinitely long cylindrically symmetrical arrangement in which a centrally situated radiating body is surrounded by a bulb coated with the filter mentioned above. We assume that the geometry is such that all rays emitted from the body, in so far as they are reflected from the filter, return to the body after a single reflection. We denote the transmission of the filter by $D(\lambda)$ and its reflectivity by $R(\lambda)$. For simplicity we disregard for the time being the angular dependence of these quantities, and assuming that the radiating body has the energy distribution of a black body, $E_0(\lambda,T)$, we can write for the luminous efficiency of the set-up:

$$\eta(T) = \frac{\int_0^\infty E_0(\lambda,T)D(\lambda)V(\lambda)\mathrm{d}\lambda}{\int_0^\infty [1-R(\lambda)]E_0(\lambda,T)\mathrm{d}\lambda}. \quad (21)$$

For maximum efficiency the filter should have zero absorption and should meet the following conditions:
$D = 1$ and hence $R = 0$ for $0.4 \leqq \lambda \leqq 0.7$ μm,
$R = 1$ and hence $D = 0$ for $0.7 \leqq \lambda \leqq \infty$.
These conditions are substituted for equations (5).

There are unfortunately no filters that can meet these conditions. The nearest approach to the ideal is to use interference filters on the basis of multiple dielectric layers. Such filters are rather expensive, however, and in lighting engineering can be used only in special cases, for example in certain types of projection lamp for the purpose of reducing the heating of irradiated objects. The use of these filters for raising the luminous efficiency of incandescent lamps has hitherto been confined to experiments [9] [10].

Fig. 15 shows the spectral characteristic for the reflection and transmission of a filter consisting of a number of dielectric layers, for angles of incidence of 0° and 45°. Their marked angular dependence is a serious drawback of all interference filters. This makes it necessary to keep the dimensions of the radiating body small compared with those of the bulb or the filters, and therefore the alignment of the radiating body has to meet very stringent requirements. Another drawback of the interference filter, also apparent in fig. 15, is that it gives the required strong reflection only in a narrow band of wavelengths. The remainder of the infra-red radiation is dissipated by absorption in the filter and therefore plays no part in raising the efficiency. The disadvantage of this narrow reflection band can be overcome to some extent by increasing the number of dielectric layers. This may be likened to the series connection of electrical resonant circuits which differ slightly in their tuning; the mathematical description of both systems is identical. We shall not go into this subject here, firstly because these filters offer nothing fundamentally new, and secondly because their expense rules out their application on a wide scale.

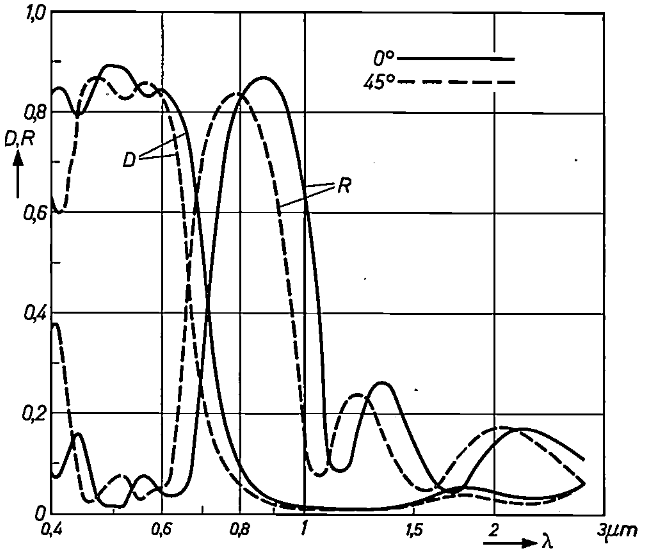


Fig. 15. Spectral transmission D and reflectivity R of an interference filter consisting of dielectric layers, for an angle of incidence $\Theta = 0°$ (solid curve) and $\Theta = 45°$ (dashed).

It might be asked, however, whether it would be possible to make with a single layer a reflection filter capable of approximately meeting the requirements mentioned above. It has long been known, for example, that thin gold and copper layers look green in transmitted light but reddish-yellow in reflected light. *Fig. 16* shows the transmission and reflectivity of a thin gold layer as a function of wavelength. It can be seen that the reflectivity already attains values of about 90% in the near infra-red, while the transmission of the layer in visible light is still round about 50%.

We have experimented with gold layers on incandescent lamps, but the results were negative. We attribute this to the evident imperfection of the filters and also to the fact that the coiled filament deviates considerably from the required spherical or cylindrical symmetry. In *sodium lamps*, on the other hand, the use of these gold filters did have the success we expected. This is bound up with the fact that the energy balance of a sodium lamp depends to a great extent on the heat losses of the discharge tube. To ensure optimum sodium-vapour

[8] E. Kauer, Physics Letters 7, 171, 1963.
[9] F. J. Studer and D. A. Cusano, J. Opt. Soc. Amer. 43, 522, 1953.
[10] E. M. Beesley, A. Makulec and H. H. Schroeder, Illum. Engng. 58, 380, 1963.

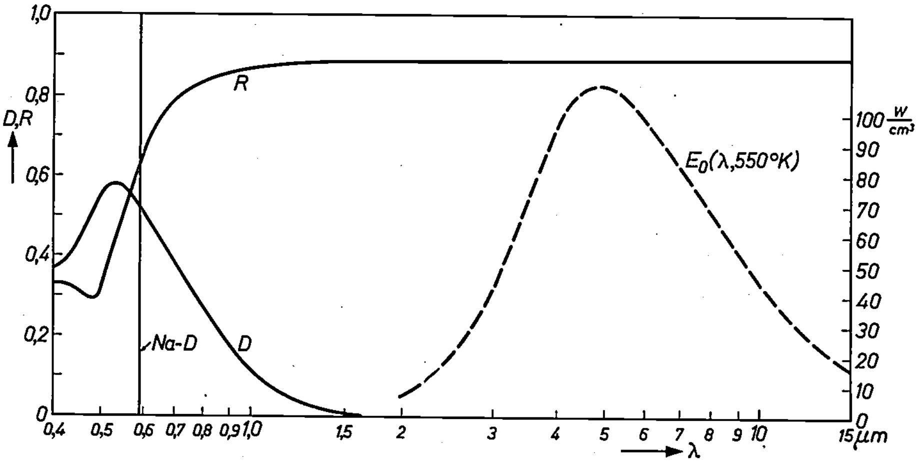


Fig. 16. Spectral transmission D and reflectivity R of a gold layer approx. 15 nm thick. The dashed curve represents the infra-red spectrum of the radiation from a sodium lamp operated at 550 °K.

pressure during operation, this tube has to be kept at a temperature of about 280 °C. In modern sodium lamps, with a sealed-in vacuum jacket, the heat losses of the tube consist largely of the thermal radiation from the discharge tube, which may be regarded as virtually a black body with a temperature of 550 °K. A simple calculation shows that a 140 W sodium lamp dissipates about 100 W as heat radiation. The maximum in the energy distribution of this radiation lies roughly at 5.5 µm, i.e. sufficiently far away from the Na-D lines, so that the transmission curve of the filter used need not be so steep as it is required to be in the case of the incandescent lamp. By coating the inside of the vacuum jacket of the sodium lamp with a layer of gold the absorbed power was reduced to such an extent that, taking ballast losses into account, the luminous efficiency rose from about 93 to 120 lm/W. In this way, for the first time in the history of lighting engineering, a practical lamp was made that had an efficiency of more than 100 lm/W [11].

The application of the gold layer, however, had the undesirable side-effect of reducing the total light output of the sodium lamp. This was partly due to the lamp having to operate at a lower current density at the same temperature, and also to the loss of light in the gold layer. The next thing was therefore to find out whether there were other materials than gold or copper which, in the form of a thin layer, combined strong infra-red reflection with lower light losses. Since the optical behaviour of copper and gold in the infra-red is governed by the free charge-carriers, it was obvious to look for substances possessing electrical properties similar to those of gold and copper, or which might be given them by doping. For a theoretical treatment we return to the equations (16, 17) on page 42. If the layer thickness is properly chosen, the boundary between the transmission and reflection of the filter is mainly determined by the plasma frequency ω_p. In order to separate the visible light from the near infra-red, it is necessary to fulfil practically the same electrical conditions as in the case of the ideal metallic selective radiators (see page 44). To give the reflection curve a steep slope and at the same time obtain strong reflection in the infra-red, we therefore require a high value for the product of mobility and effective mass — but now, as already mentioned, at relatively low temperatures.

If the wavelength dependence of the optical constants of a material in the visible and in the near infra-red regions are solely determined by free charge-carriers, that is to say if no other kinds of absorption mechanism are operative, the optical constants can be calculated from the theory, provided the material parameters in equations (16) and (17) are known. At a given layer thickness we can then calculate the transmission and reflection curves of a filter [12].

In this way we have calculated the transmission and reflection curves for a hypothetical filter with the following material parameters: $\varepsilon_g = 10$, $N = 1.13 \times 10^{22}$ cm^{-3}, $\mu = 100$ cm^2/Vs, $m^* = 0.5\ m_0$, and $d = 0.09$ µm. The curves are shown in *fig. 17.* As can be seen, given a suitable thickness, effective separation is possible between visible light and infra-red. The values we have assumed for the parameters are by no means exceptional if one compares them, for example, with those for

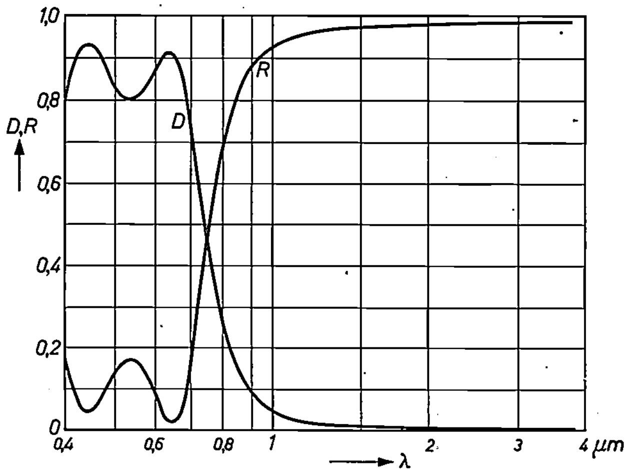


Fig. 17. Spectral transmission D and reflectivity R of a hypothetical filter, the material parameters being: $\varepsilon_g = 10$; $N = 1.13 \times 10^{22}$ cm^{-3}; $\mu = 100$ cm^2/Vs; $m^* = 0.5\ m_0$; $d = 0.09$ µm.

LaB$_6$. As regards the choice of substances, the same considerations apply as mentioned on pages 42-44.

In spite of this theoretical result, little success has yet been achieved in the making of filters that separate the visible light and the near infra-red. This is partly because there are not many substances that have the appropriate plasma frequency, and partly because some of these substances have a very high melting point (e.g. LaB$_6$) and are therefore difficult to make in the thin layers required. If the demands are less rigorous, however, as in the case of the sodium lamp, then practical solutions of the problem are certainly possible, offering better results than the above-mentioned gold filters. It has long been known that stannic oxide layers (SnO$_2$) produced by thermal decomposition of SnCl$_4$, possess good electrical conductivity and are also completely transparent. Extensive experiments designed to increase the mobility and concentration of the charge carriers in these layers by suitable manufacturing and doping conditions, have led to filters which have a reflectivity of 80% at 5 μm. As an example, *fig. 18* shows the transmission and reflection of an SnO$_2$ layer, 0.32 μm thick, coated on glass. For the Na-D lines the transmission of this filter is 90%, which is very nearly up to the value of 92% that might be expected from a glass plate without such a coating. For sodium light, then, this filter may be regarded as virtually free from losses. Used in a sodium lamp, this filter results in the same luminous efficiency as obtained with the gold filter. As a result of the better transmission in the visible region, however, the light output (luminous flux) is twice as high as that of the lamp with the gold layer.

Further investigations of the electrical and optical properties of substances which are not yet so well-known may well result in a simple reflection filter possessing a favourable reflection curve for radiators at high temperatures. Apart from the possibility, mentioned in this article, of raising luminous efficiency by using more or less exotic substances instead of tungsten for the filaments of incandescent lamps, there are real prospects that the efficiency of these traditional light sources may be substantially improved by a much less radical modification, that is by coating the bulb with a suitable filter.

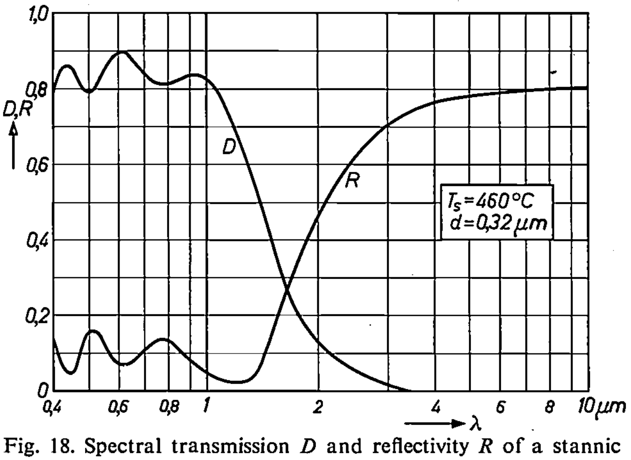


Fig. 18. Spectral transmission D and reflectivity R of a stannic oxide filter on glass. The layer is characterized by the following material parameters, determined by experiment: $\varepsilon_g = 3.8$; $N = 6\times10^{20}$ cm^{-3}; $\mu = 20$ cm^2/Vs; $m^* = 0.25\ m_0$; $d = 0.32$ μm.

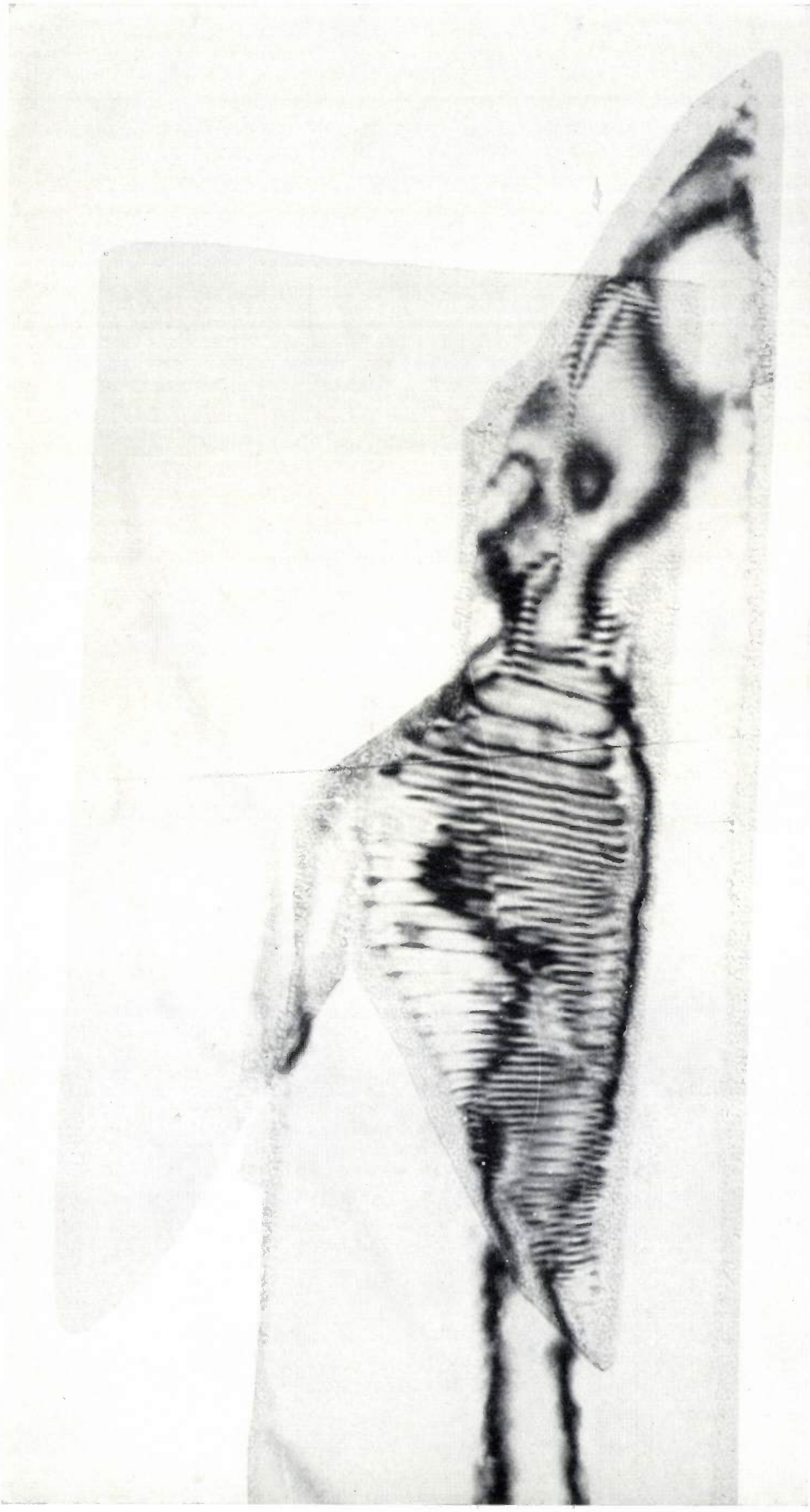
Summary. The luminous efficiency of tungsten lamps and other thermal radiators used as light sources is limited by the fact that most of the energy is radiated in the infra-red. Long ago efforts were made, without much success, to improve the luminous efficiency by using selective radiators with a relatively low emission in the infra-red. There are two possible types of selective radiator: one is the type which is completely transparent in the infra-red, and the other is the metallic reflecting type, which reflects all radiation in the infra-red. In this article the feasibility of such radiators is examined in some detail on the basis of modern theoretical considerations, largely from solid state physics, and making use of new data on various substances, partly obtained by measurements on single crystals. The four absorption processes in the solid state are examined in this connection, viz, absorption due to lattice vibrations, to free charge-carriers, to lattice defects, and to the excitation of charge carriers in semiconductors (absorption edge). The conclusion is that the prospects for transparent selective radiators are fundamentally limited. On the other hand, metallic reflecting layers used in conjunction with appropriate regenerative processes (comparable with the familiar tungsten-iodine cycle) are thought to have good chances of success.

Another useful device for raising the luminous efficiency is to surround a light source (thermal or non-thermal), which is required to meet certain geometrical requirements, with a filter that remains cold itself and selectively reflects the infra-red radiation back to the radiant body while passing the visible radiation. A filter of this kind, in the form of a thin layer of gold or stannic oxide, has already found successful application in sodium lamps. There are good grounds for believing that a suitable filter will also be found for the incandescent lamp.

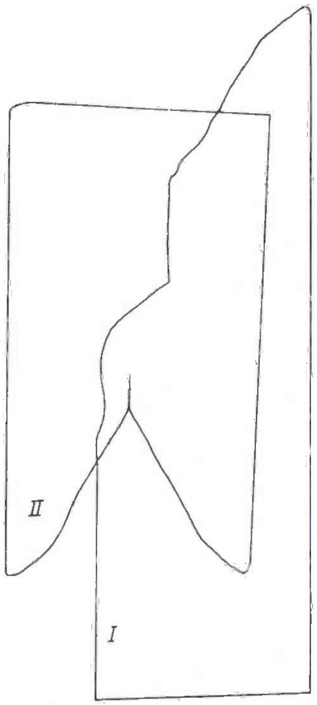
[11] For further particulars, see: M. H. A. van de Weijer, Recent improvements in sodium lamps, Philips tech. Rev. 23, 246-257, 1961/62.

[12] See H. Mayer, Physik dünner Schichten, Part I, Wiss. Verlagsges., Stuttgart 1950, p. 144. The latter step is admittedly subject to the restriction that the optical constants, owing to the "path length efficiency" (see Part II of the above book, 1955, p. 204) and also owing to structure effects, may differ from those of the bulk. In each individual case it is necessary to ascertain which electrical data are applicable to the thin layer.

Bragg reflections in electron photomicrographs



The photograph, taken with a Philips EM 100 B electron microscope, is of *synthetic* γ FeOOH, a substance which also occurs naturally as the mineral lepidocrocite. The overall magnification is about 200 000 times. The specimen consisted of small, extremely thin crystal wafers (thickness approx. 10 nm). As a rule the wafers are slightly bent, so that the planes of atoms in the wafer differ a little in orientation from place to place. Certain regions in the wafers may then be oriented so that part of the incident electron beam undergoes Bragg reflection. Electrons reflected at wide Bragg angles are intercepted by a diaphragm, and do not reach the photographic plate, so that these places show up in the picture as dark patches. The photograph represents two partly overlapping crystal wafers *I* and *II* (see the sketch). The two black strips in the non-overlapped bottom part of wafer *I* are due to Bragg reflections of this nature. In the part where the wafers overlap the play of Bragg reflections in the two crystals produces the remarkable skeleton-like pattern shown here.



An experimental "Plumbicon" camera tube with increased sensitivity to red light

E. F. de Haan, F. M. Klaassen and P. P. M. Schampers

621.397.331.222

The new type of television camera tube, recently introduced under the name of "Plumbicon", offers scope in principle for a wide range of variations. During experiments into the practical possibilities, part of the PbO in the photoconducting layer was successfully replaced by a PbO-PbS mixture which has a smaller energy gap, and thus the authors were able to make tubes with a much higher sensitivity to red light. This opens up prospects of a further important advance in the field of camera tubes. One of these tubes, which has been used with excellent results for colour television, is discussed in the article below.

The characteristic difference between a "Plumbicon" television camera tube and an ordinary vidicon is that the former has a photoconducting layer (target) with *rectifying* (or *blocking*) contacts. It is to these contacts that "Plumbicon" tubes owe their extremely low dark current. The contact on the gun side is made by doping the basic target material on the relevant surface of the photoconducting layer in such a way as to make it a *P*-type semiconductor. The other contact is formed by the signal plate, which is made of strongly *N*-type SnO_2, or by the adjoining sublayer, if it has become *N*-type by contact with the SnO_2.

It was explained in a previous paper [1] that, in a wide range of donor concentrations and charge carrier mobilities, the energy (or band) gap ΔE of the photoconducting material theoretically need not be greater than about 0.9 eV in order to be able to apply sufficiently strong blocking contacts; with this energy gap the dark current at room temperature would be about 10^{-8} A. The energy gap of the red modification of PbO, which is the principal material of the photoconducting layer in the tubes discussed in that article, was very much higher than this limit, amounting to 2.0 eV.

Apart from the dark current the spectral sensitivity is also governed to a great extent by the energy gap. The smaller the energy gap, the higher is the cut-off wavelength λ_g, i.e. the upper limit of the wavelength range within which a photoconductor absorbs. Expressed as a formula, $\Delta E = hc/\lambda_g$, where h is Planck's constant and c is the velocity of light. An energy gap of $\Delta E = 2.0$ eV corresponds to a cut-off at about 6200 Å. Although, then, the relatively wide energy gap of red PbO is advantageous from the point of view of the dark current, for a tube to be also sensitive to long-wave red light or to infra-red radiation it is evidently necessary to use a material which has a narrower energy gap.

Philips Laboratories at Eindhoven have now succeeded in making "Plumbicon" tubes which have a greater red sensitivity than the tubes previously described [1]. This has been done by using for the photoconductive layer a PbO material in which the oxygen atoms have partly been replaced by sulphur atoms. The energy gap of this material is narrower than that of pure PbO, being narrower the greater the fraction of PbS. One of these tubes will be discussed in this article. The tube in question, apart from its increased red sensitivity, differs very little in other respects from the tube described earlier. We shall refer to it here as "tube II"; the tube with a target material of pure PbO will be referred to as "tube I". We have chosen tube II as an example because of its eminent suitability for colour television.

Details of the tube

The "Plumbicon" with increased red sensitivity differs from tube I in that the target material on the gun side — i.e. the side where the *P* contact is — consists of a layer of PbO-PbS instead of PbO. On the side where the *N* contact is, PbO has been left unchanged,

Dr. E. F. de Haan (deputy director), Dr. F. M. Klaassen and P. P. M. Schampers are research workers at Philips Research Laboratories, Eindhoven, Netherlands.

[1] E. F. de Haan, A. van der Drift and P. P. M. Schampers, The "Plumbicon", a new television camera tube, Philips tech. Rev. **25**, 133-151, 1963/64 (No. 6/7).

and as a result of this the tube is almost identical with the other in respect of its sensitivity to blue. Because of the presence of PbS in the regions last to be reached by the light, however, the fraction of the red light that would pass through pure PbO is now absorbed. This increases the sensitivity to red light, the cut-off now lying at a wavelength about 1000 Å higher. Compared with tube I, then, tube II has a better spectral sensitivity distribution in the red.

When the new tube is used in a camera for colour television, the "red" channel is so much more sensitive than with tube I that it is no longer this channel but the "blue" one which determines the overall sensitivity of the camera. The overall sensitivity is in fact about three times higher, one result of which is a better signal-to-noise ratio of the Y signal. Using a colour camera equipped with tube II it is possible under normal conditions to obtain an excellent television picture with an on-scene level of illumination of only 500 lux [2]. If some noise can be tolerated, a level as low as 100 lux is sufficient. These levels of illumination are no higher than those needed for black-and-white television using an image orthicon.

Fig. 1 shows the spectral sensitivity distribution of tube II. As can be seen, the curve has *two* peaks: the first corresponds to the spectral sensitivity of pure PbO, the other to that of the PbO-PbS mixture.

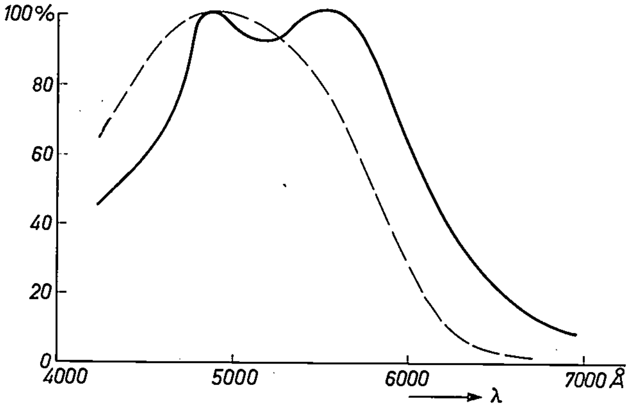


Fig. 1. Spectral sensitivity distribution of an experimental "Plumbicon" camera tube, in which the PbO of the photoconducting layer has partly been replaced by PbS (tube II). This considerably increases the sensitivity to red light. The left peak corresponds to the spectral sensitivity of PbO, the other to that of the PbO-PbS mixture. The dashed line gives the spectral sensitivity distribution of a tube with a PbO layer made by the standard process [1].

As a result of the greater sensitivity to red light the sensitivity to white light is also higher, being about 400 μA/lm [3]. It can be seen in *fig. 2* that the photocurrent i_f as a function of the applied voltage again exhibits the saturation effect. The coefficient γ from the formula $i_f \propto E^\gamma$, which gives the relation between the photocurrent and the illumination E, has a value close to unity, as in tube I.

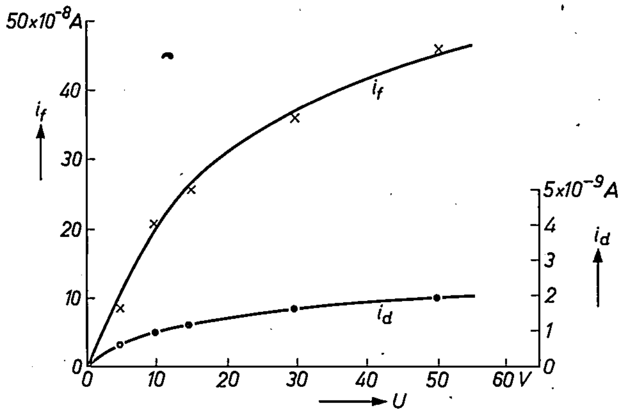


Fig. 2. The photocurrent i_f and the dark current i_d of tube II as a function of the voltage U across the target at a given level of illumination.

Now that all the light is absorbed in the layer, there is no longer any loss of definition due to the scattering of red light reflected from the back of the target. The resolving power of the tube is therefore even better than that of tube I.

A consequence of the new tube's high red sensitivity is that its dark current contains a not insignificant component, caused by the light from the cathode. This unwanted component, being a parasitic photocurrent, is of course more pronounced the higher the value of λ_g. By taking suitable measures, however, for example by using a low cathode temperature and a narrow diaphragm in the gun, it can be kept reasonably low. The dark current of tube II, even with an applied voltage of 100 V, is therefore no higher than about 3×10^{-9} A (fig. 2). As regards the speed of response, tube I and tube II are more or less equivalent; see the recorded curve in *fig. 3*. The life of the tubes II so far made also seems to be on a par with that of tubes I.

Without going back on our intention to discuss only one type of tube, a final word might be appropriate in regard to one of the other tubes we have made. This other tube was not made with a view to its suitability for a given application, but solely to demonstrate the extent to which a "Plumbicon" tube can be given a

[2] The factors which, apart from the sensitivity of the camera tube, govern the illumination required will be dealt with in an article by A. G. van Doorn, shortly to be published in this journal.

[3] Measured in the same way as described in the article mentioned in footnote [1]. Here too, white light is understood to be black-body radiation at 2640 °K.

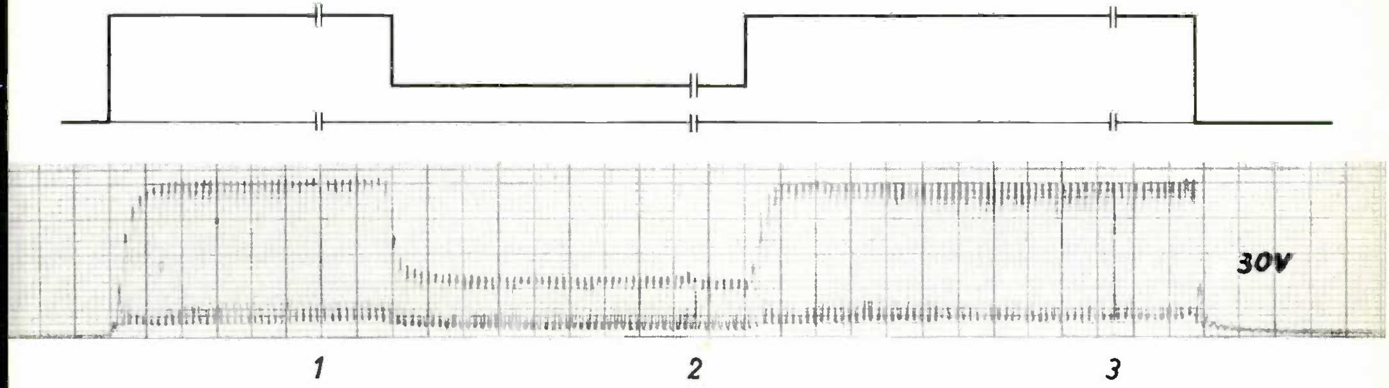


Fig. 3. The response of tube II at an applied voltage of 30 V. The curve was obtained by illuminating only the upper half of the tube. The signal current consists of a train of pulses spaced at intervals of 1/50 s (half-a-frame time because of the interlaced scanning). The programme of illumination is indicated above the recording, and began in each case with a dark period lasting one minute. At points *1*, *2* and *3* the recording was interrupted for 10, 10 and 30 seconds respectively.

particular spectral sensitivity distribution. The sensitivity chosen in this case was that of the human eye. The result is shown in *fig. 4*. As can be seen, the differences are very slight; only in the blue is the tube relatively more sensitive than the eye.

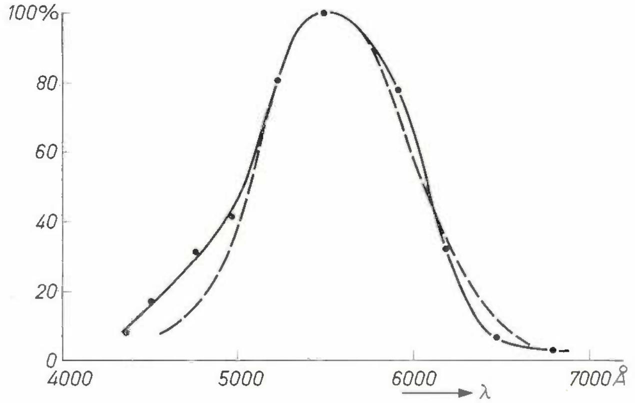


Fig. 4. Illustrating the extent to which it is possible to give a specified shape to the spectral sensitivity distribution curve of a "Plumbicon" tube with a PbO-PbS target. The curve was measured on a tube which was designed in an attempt to match the spectral sensitivity distribution as closely as possible with that of the human eye (dashed curve).

Summary. The upper limit of the spectral sensitivity of a "Plumbicon" camera tube can be shifted to a higher wavelength by adding PbS to the PbO of the target. Some details are given of an experimental tube in which this has been done. The photoconducting layer of this tube (tube II) consists partly of pure PbO and partly of a PbO-PbS mixture. Its characteristics are almost identical with those of a tube in which pure PbO is used (tube I) but the sensitivity to red light is much higher and the resolution is somewhat improved. Tube II is just as fast as tube I (speed of response about 3/50 s). The value of γ is close to unity. The dark current is about 3×10^{-9} A.

Phase theory

II. Quantitative considerations on binary systems

J. L. Meijering

541.12.01

The calculation of a binary phase diagram — at least of certain details, since a complete calculation is scarcely practicable — makes it possible both to *verify* the experimentally determined diagram and to *add to our knowledge of it.* We shall give examples of both possibilities. As regards obtaining additional knowledge, we shall give as an example a prediction of metastable miscibility gaps, which would not easily have been brought to light without that prediction.

Each point will be approached from two sides: first considering the behaviour of the Gibbs' free energy as given and drawing conclusions from it concerning the phase diagram, and then taking the phase diagram as our starting point and drawing inferences regarding the free energy, or more in particular the enthalpy or entropy. As far as possible our considerations will be quantitative, but where appropriate we shall include qualitative aspects in our discussion. Since what can be understood quantitatively in a phase diagram represents only a fraction of the total, the picture would be all too fragmentary if we were not to do so.

Further treatment of some subjects from Part I

In Part I of this article the following expression was given as a first approximation for the Gibbs' free energy G of a binary phase:

$$G = H - TS$$
$$= ax(1-x) + RT\{x \ln x + (1-x)\ln(1-x)\}\,. \quad (1)$$

The factor a, a measure of the energetic interaction between the atoms or molecules of the components, is provisionally assumed to be positive.

Considering that this expression is symmetrical with respect to $x = \frac{1}{2}$, and does not therefore have general validity, we immediately allowed ourselves to depart from this formula, and in the rest of the article we took asymmetrical G-x and T-x diagrams as our starting points. We shall now retrace our steps for a moment and take this symmetrical formula as our basis. In view of the said symmetry with respect to $x = \frac{1}{2}$, at every temperature the double tangent line on the G-x curve is *horizontal.* As the tangent point indicates the concentration where, at the relevant temperature, the miscibility begins, we obtain the equation defining the boundary of the miscibility gap (T as a function of x) by putting $\mathrm{d}G/\mathrm{d}x = 0$. This gives:

$$T = \frac{a(1-2x)}{R \ln[(1-x)/x]}\,. \quad \ldots\ldots (2)$$

The miscibility gap thus calculated is shown in *fig. 1.*

The calculation is not nearly as simple when we turn to an asymmetric case, but it discloses an aspect that might otherwise easily be overlooked. We now take:

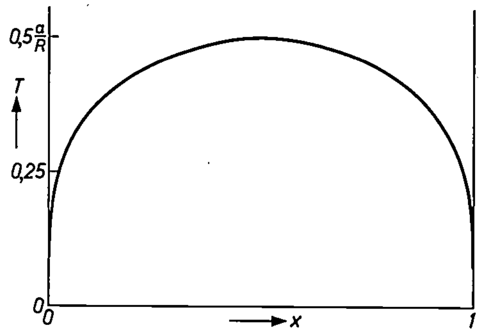


Fig. 1. Miscibility gap calculated with equation (2).

$$H = a(2x - x^2 - x^3)\,. \quad \ldots (3)$$

For simplicity we keep the same term for the entropy of mixing, i.e. Gibbs' expression:

$$S = -R\{x \ln x + (1-x)\ln(1-x)\}\,, \quad \ldots (4)$$

and we then have:

$$G = a(2x - x^2 - x^3) +$$
$$+ RT\{x \ln x + (1-x)\ln(1-x)\}. \quad \ldots (5)$$

Let x_1 and x_2 be the concentrations at the tangent points on the double tangent line, then the following equations can be derived:

$$RT \ln \frac{1-x_1}{1-x_2} = a(x_2^2 + 2x_2^3 - x_1^2 - 2x_1^3)\,, \quad . (6)$$

and

$$RT \ln \frac{x_1}{x_2} = 2a(x_1 + x_1^2 - x_1^3 - x_2 - x_2^2 + x_2^3)\,. \quad \ldots (7)$$

Whereas in eq. (2) the value of T can easily be calculated as a function of x, it is necessary to use graphical methods or the method of successive approximation in order to solve T and x_2 (or x_1) in (6) and (7). The

Prof. Dr. J. L. Meijering, formerly with Philips Research Laboratories, Eindhoven, is now Professor of Inorganic Chemistry and Metallurgy at Delft University, Netherlands.

reason is that it is not possible to eliminate x_2 from (6) and (7), as can be done in the symmetric case, where one always has $x_2 = 1 - x_1$. Equations (6) and (7) together represent a T-x_1-x_2 curve in three dimensions.

It becomes clear that the solubility of a substance is not only governed by the properties of the solution concerned, *but that the solubility depends essentially on the phase or phases with which the solution is in equilibrium*. This is obvious enough, but it is all too often forgotten. There is not much point in drawing conclusions from the comparison of the solubilities of oxygen in copper and nickel if one fails to take into account that Cu_2O and NiO are the respective coexisting phases. A solubility is not purely and simply a property of the solution.

In binary systems, the tie lines — these are the lines which, in a two-phase region, join points that indicate the concentrations of two coexisting phases — therefore play an important part; the custom of drawing several of them·has more significance than mere cross-hatching: they remind us that the two points in each case appertain to one another.

In extension of another topic from Part I we shall now give the *quantitative* expressions for the dependence of the solubility x_e on temperature and on pressure. These expressions were first worked out by Van der Waals. Although the derivation is not difficult, it will be sufficient here to refer to the literature [1]. We will, however, show that the expressions given here are in complete agreement with the qualitative considerations in Part I. The relations read:

$$\frac{\mathrm{d}x_e}{\mathrm{d}T} = \frac{\mathrm{d}S/\mathrm{d}x - \Delta S/\Delta x}{\mathrm{d}^2G/\mathrm{d}x^2}, \quad \ldots\ldots \quad (8)$$

$$\frac{\mathrm{d}x_e}{\mathrm{d}p} = -\frac{\mathrm{d}V/\mathrm{d}x - \Delta V/\Delta x}{\mathrm{d}^2G/\mathrm{d}x^2}. \quad \ldots\ldots \quad (9)$$

The quantities on the right-hand side should be taken at the concentration $x = x_e$. ΔS, ΔV and Δx represent $S' - S$, $V' - V$ and $x' - x$, the quantity with the prime relating to the coexisting phase. The influence of the coexisting phase, referred to above, is clearly apparent in these formulae.

With the aid of a figure substantially the same as *fig. 2*, we considered in Part I how the solubility of B in the solid phase varies with temperature. We saw that this depended on the slope of the line tangent to the lower S-x curve ($= \mathrm{d}S/\mathrm{d}x$) compared with the slope of the straight line joining the tangential point to the appertaining point of the coexisting phase ($= \Delta S/\Delta x$). If the first slope is less steep (as at the equilibrium KL), the solubility decreases with temperature; if it is steeper (at MN), the solubility increases. We note that formula (8) leads to the same conclusions, *provided that d^2G/dx^2 is positive*. The same applies, with due alteration of details, to eq. (9).

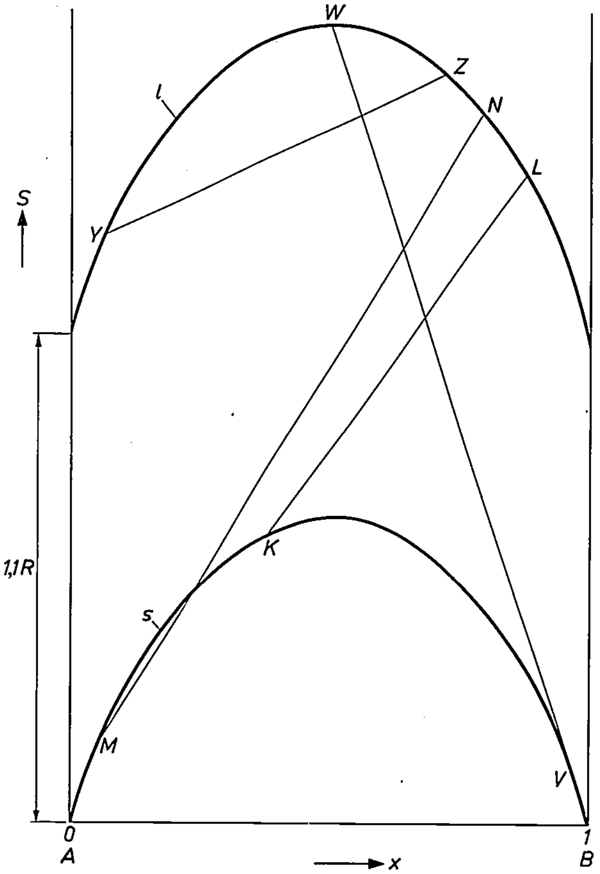


Fig. 2. Idealized S-x diagram for alloys of normal metals, given in Part I as fig. 22, to which reference may be made for further particulars. The lines VW and YZ are used later in this article.

This brings us to a very important point: $\mathrm{d}^2G/\mathrm{d}x^2$ cannot in fact be other than positive, for where it is negative the system is an *unstable* one. Consider *fig. 3*, for example. There, $\mathrm{d}^2G/\mathrm{d}x^2$ is negative between the points of inflexion A and B on the curve. In this region every local fluctuation in the concentration will cause the free energy to decrease, with the result that a separation process sets in giving an increasing phase separation. If the initial state is D, for example, then point E on the line FG indicates a possible intermediate stage in this separation process. This irrevocable though sometimes slow process only comes to an end at point H on the double tangent line OP. For comparison, let us consider an initial state between an inflexion point and a tangent point, e.g. Q. Phase separation on a small scale, e.g. up to R on line ST, should now result in an increase of the free energy, and since state R is less stable than Q, this phase separation does not occur. A more stable state can, however, be reached by separation up to point U on the double tangent line.

[1] See J. L. Meijering, Philips Res. Repts. 3, 281, 1948.

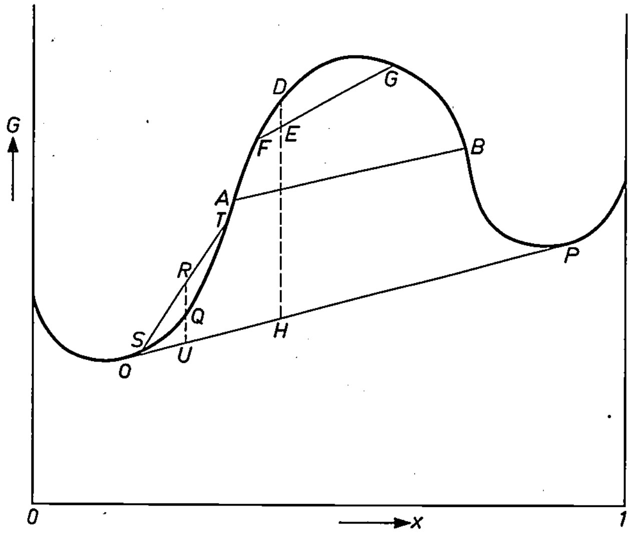


Fig. 3. To distinguish between unstable, metastable and stable regions in the G-x diagram.

In short, as far as the homogeneous phase is concerned, we have an *unstable* region between the inflexion points of the curve, a *metastable* region between tangent point and inflexion point, and a *stable* region beyond the tangent points.

The locus of the inflexion-point concentrations in the T-x diagram is termed the *spinodal* curve (the dashed curve in *fig. 4*). It forms the boundary between the unstable and the metastable regions and touches the boundary of the miscibility gap at the critical point C. The spinodal curve is given by $\mathrm{d}^2G/\mathrm{d}x^2 = 0$, and at the critical point both $\mathrm{d}^2G/\mathrm{d}x^2$ and $\mathrm{d}^3G/\mathrm{d}x^3$ must be zero. If G is given by eq. (5), we find for the spinodal curve:

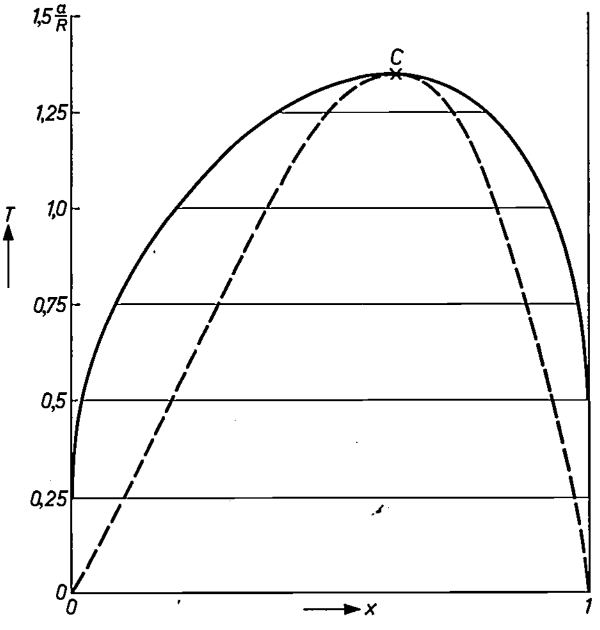


Fig. 4. Calculated miscibility gap and *spinodal* curve (dashed line), which separates the unstable from the metastable region. C is the critical point. The Gibbs' free energy was assumed to be given by eq. (5).

$$T = x\,(1-x)\,(2+6x)\,(a/R)\,,$$

and for the critical point:

$$x = \frac{2+\sqrt{13}}{9} = 0.623,$$

and

$$T = \frac{140+52\,\sqrt{13}}{243}\,\frac{a}{R} = 1.348\,\frac{a}{R}\,,$$

see fig. 4. The boundary of the miscibility gap in fig. 4 is found, using (6) and (7), by the method of successive approximation.

Having thus calculated from a given curve for the free energy the corresponding miscibility gap, we shall now look at the question from another angle. We shall consider a particular miscibility gap as given, and see how we can use this as a starting point for obtaining further knowledge of the phase diagram.

The phase diagram in the proximity of a critical point

Solidus curve above a miscibility gap

Let us consider a T-x diagram which shows in its lower part a miscibility gap relating to a solid phase. Suppose, too, that the solidus curve of this phase runs just above the critical point. From the foregoing we know that $\mathrm{d}^2G/\mathrm{d}x^2$ (and $\mathrm{d}^3G/\mathrm{d}x^3$) of the relevant phase is zero at the critical point. On the solidus curve, therefore, $\mathrm{d}^2G/\mathrm{d}x^2$ cannot differ very much from zero. This means according to eq. (8) that $\mathrm{d}x/\mathrm{d}T$ is large: near the critical point the solidus curve in such a case is nearly horizontal. Examples are to be found in the Al-Zn and Au-Pt systems; see *fig. 5*.

If a solidus curve is just in contact with a miscibility gap, then the critical point is likewise the point of inflexion of the solidus curve, and the common tangent line is horizontal. *Fig. 6* shows the form of the T-x diagram when the solidus curve is relatively somewhat lower. In that case a *peritectic* three-phase equilibrium occurs. Note that, as opposed to the eutectic, in a three-phase equilibrium of this kind the liquid phase is "peripheral" with respect to the two coexisting solid phases.

Liquidus curve above a miscibility gap

Liquidus curves which are remarkably flat have been found in various metallic systems, for example Cu-Fe, Cu-Co and Cu-Li, and also in the water-salicylic acid system, which has become a classic example in the group of non-metallic systems.

We can now proceed to reverse the above argument. The shape of the liquidus curve indicates that *as far as the liquid phase is concerned* $\mathrm{d}^2G/\mathrm{d}x^2$ is small. A critical point of a miscibility gap of a liquid phase will therefore be near at hand. It should be realized that only where this miscibility gap is *above* the liquidus curve

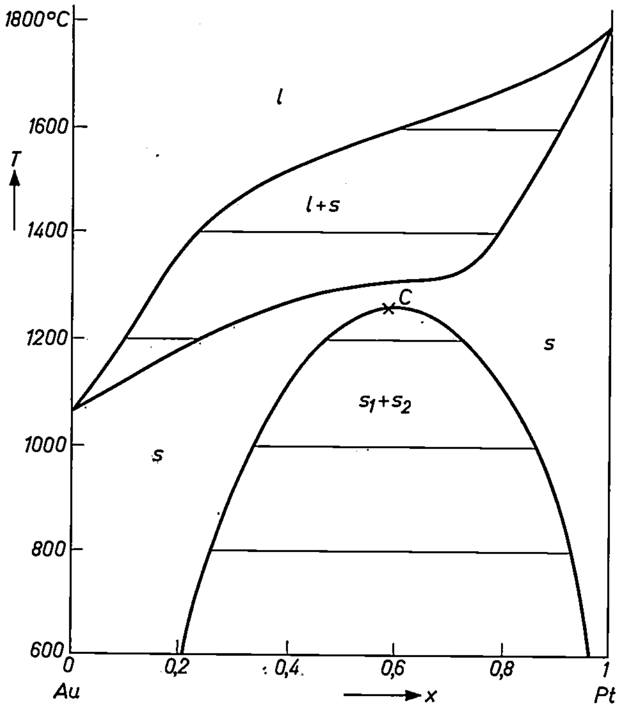


Fig. 5. *T-x* diagram of the Au-Pt system, where the solidus curve is fairly horizontal near the critical point *C*. A solidus curve (or liquidus curve) of this shape indicates the proximity of a miscibility gap.

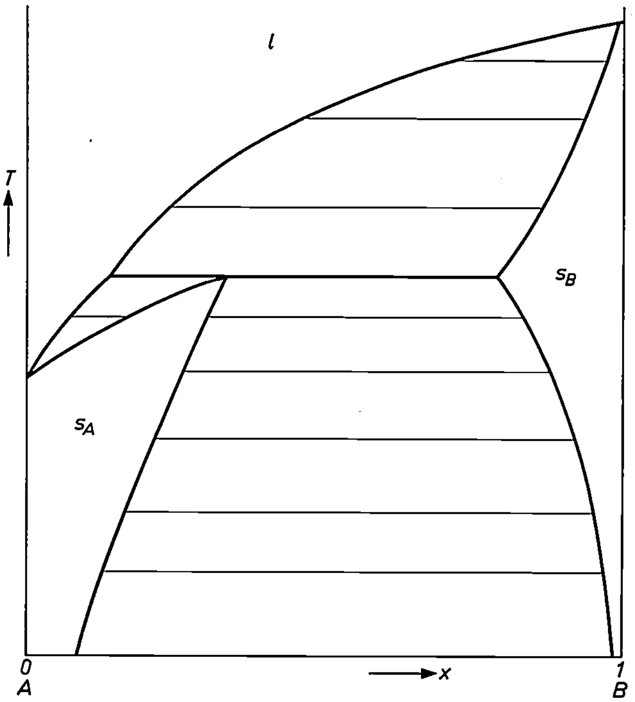


Fig. 6. *T-x* diagram with a *peritectic* three-phase equilibrium ($l + s_A + s_B$). The liquid phase here is "peripheral" in relation to the two solid phases. This diagram can be imagined as produced from fig. 5 by letting the solidus curve drop a little further.

— we shall return presently to this possibility — can it in fact be stable; in the case we are about to consider, where the miscibility gap is *below* the liquidus curve, it is metastable (*fig. 7*). This metastable miscibility gap is brought to light only upon *undercooling* of the stable liquid phase.

In this context we should mention Nakagawa's elegant experiment [2]. Nakagawa started from molten Cu-Fe and Cu-Co alloys with a Cu content of roughly 50 at.% which exhibit a strikingly flat liquidus curve By undercooling these alloys 20 and 90 °C respectively, he was indeed able to produce a separation into two liquid phases. What makes this experiment so interesting is that the temperatures at which the separation takes place can be fairly accurately estimated.

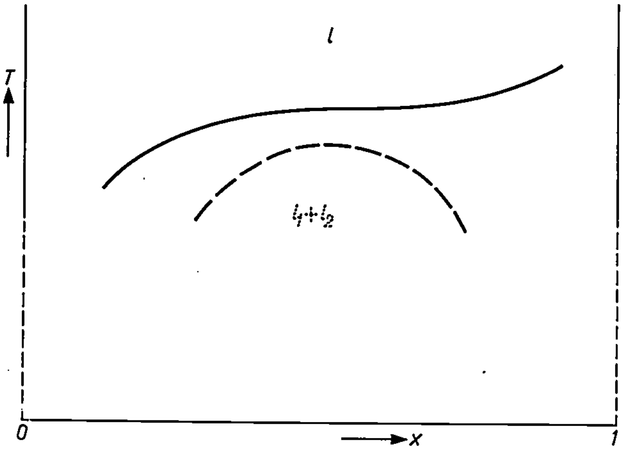


Fig. 7. Metastable miscibility gap *below* a liquidus curve. The flatness of the liquidus curve indicates the proximity of a miscibility gap of the *liquid phase*. Since this miscibility gap is here *below* the liquidus curve, it is bound to be metastable. Compare this figure with fig. 21*b* in Part I.

The estimation is based on eq. (8). By extending the tangent line at the point of inflexion of the relevant liquidus curve [3] we find the value of $\mathrm{d}x/\mathrm{d}T$ in the region of the critical point. Assuming that the liquidus curve lies δT higher than the critical temperature, and that δT is small, we may write for $\mathrm{d}^2G/\mathrm{d}x^2$:

$$\frac{\mathrm{d}^2G}{\mathrm{d}x^2} = \frac{\partial^3 G}{\partial T \partial x^2}\,\delta T\,.$$

Inserting $\partial G/\partial T = -S$ gives:

$$\frac{\mathrm{d}^2G}{\mathrm{d}x^2} = -\frac{\partial^2 S}{\partial x^2}\,\delta T\,. \quad . \quad . \quad . \quad . \quad (10)$$

Taking Gibbs' expression (eq. 4) for S, we obtain at the given concentration, $x = 50$ at.%:

$$\frac{\mathrm{d}^2G}{\mathrm{d}x^2} = R\,\delta T/x\,(1-x) = 4R\,\delta T\,.$$

To estimate the numerator of eq. (8) we turn to fig. 2.

[2] Y. Nakagawa, Acta metallurgica 6, 704, 1958.

[3] The liquidus curves of the Cu-Fe, Cu-Co and Cu-Li systems are given in the book by M. Hansen and K. Anderko, Constitution of binary alloys, McGraw-Hill, New York 1958.

Since it is known that mixed crystals with which the liquid mixtures coexist at 50 at. % Cu contain little Cu, we may roughly represent the two-phase equilibrium by the line VW. At $x = 0.5$ we have $\mathrm{d}S/\mathrm{d}x = 0$ and $\Delta S/\Delta x$ is roughly equal to $-R\,(1.1 + \ln 2)/0.5$, so that the result is:

$$\mathrm{d}x/\mathrm{d}T \approx 0.9/\delta T \quad \text{or} \quad \mathrm{d}T/\mathrm{d}x \approx 1.1\ \delta T.$$

Owing to the many approximations used, the result is certainly not reliable to within 10%, and for convenience we shall therefore assume $\mathrm{d}T/\mathrm{d}x$ to be equal to δT. As regards the two systems in question, Cu-Fe and Cu-Co, the theoretical and experimental values are in fairly good agreement. For the Cu-Li system, likewise with 50 at. % Cu, it is possible to predict, from the shape of the liquidus curve [3], that phase separation will occur at about 200 °C undercooling. This prediction still awaits experimental confirmation.

It might be objected that the condition of a small δT is not fulfilled in these systems. From eq. (10) it can be seen, however, that this condition need not be fulfilled if $\partial^2 S/\partial x^2$ does not depend on T. Nor is this the case if Gibbs' relation is a good approximation of the S-x curve.

Liquidus curve below a miscibility gap

We have just anticipated the possibility of a miscibility gap in the liquid phase lying *above* a flat liquidus curve (*fig. 8*). Such a miscibility gap, contrary to the one first discussed, is stable and is therefore not particularly difficult to determine by experiment. This type of miscibility gap, which has what is termed a *lower critical solution temperature*, is interesting for quite another reason. In particular, it is not immediately understandable why in this case an *increase* of temperature brings about the phase separation. It indicates, as we shall now explain, an "abnormal" S-x curve.

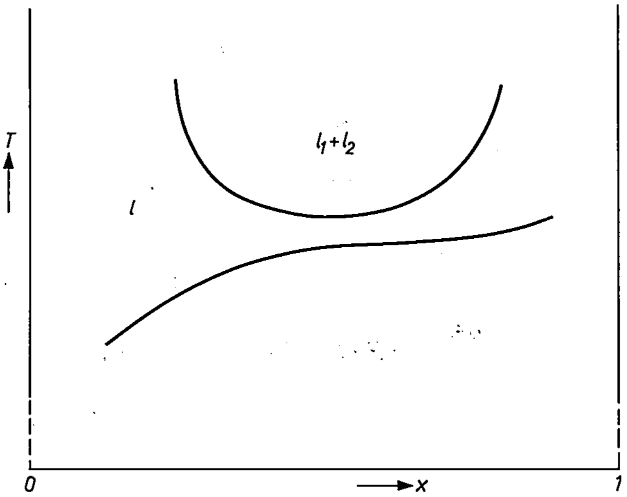


Fig. 8. Miscibility gap of a liquid phase *above* a liquidus curve. In such a case the miscibility gap is stable. An example is CO_2-orthonitrophenol.

Hitherto we have considered as normal the shape of the S-x curve in accordance with the Gibbs' function. For our purposes it is sufficient to say of this curve that, looked at from above, it is everywhere "convex". This is to say that all tie lines between points on this curve (e.g. YZ in fig. 2) and which represent heterogeneous equilibria, run *below* the S-x curve of the homogeneous phase. Since a temperature increase ultimately leads to the formation of the state with the greatest entropy, in this case the homogeneous phase, a temperature increase in the present case will obviously cause the mutual miscibility to increase. (An analogous example was given in fig. 23b in Part I, relating to the retrograde solidus curve. As long as the line MN runs *below* the curve of the homogeneous solid phase, the solubility of B in the solid phase increases with increasing temperature.) Conversely, *decreasing* mutual miscibility with increasing temperature, as encountered in the latter type of miscibility gap, must be brought into relation with an S-x curve part of which, seen from above, looks "concave" [4] (*fig. 9*). Here, connecting lines are possible that run entirely or partly *above* the curve itself.

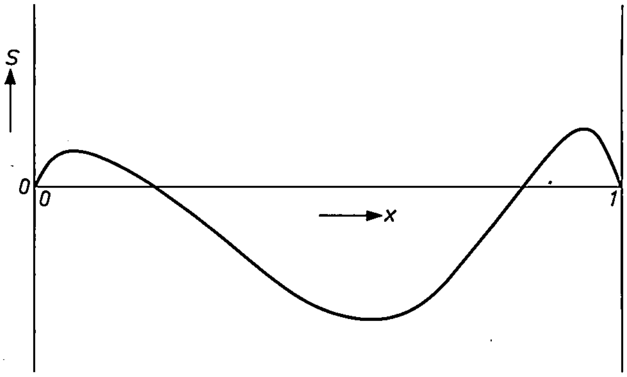


Fig. 9. *Abnormal* entropy-of-mixing curve. A curve of this kind accounts for miscibility gaps as in fig. 8, where the phase separation increases with increasing temperature.

If the S-x curve is of the type in fig. 9, the H-x curve will not usually be convex either (so far we have confined ourselves to this possibility by taking the factor a in equations (1) and (3) to be positive) but will again be concave (*fig. 10*). This means that the interaction between the atoms or molecules of the components is energetically favourable. An energetically favourable interaction certainly occurs just as often as an unfavourable one. The reason that we are only now considering this case is that in combination with a normal S-x curve it causes no phase separation — although it does in combination with an "abnormal" S-x curve.

A good example of the latter is found in the triethylamine-water system, in which a tendency to hydrate formation exists. Mixing of the components in the

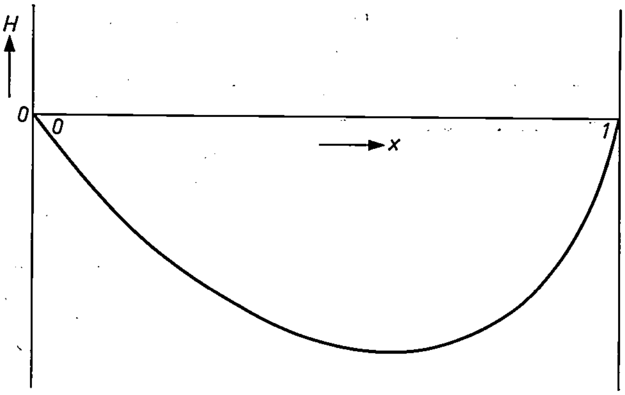


Fig. 10. Enthalpy-of-mixing curve in the case of energetically favourable interaction between the atoms or molecules of the components.

liquid phase lowers both the energy and the entropy. This is because the energetically favourable hydrate formation reduces the mobility of the molecules, thus producing a *stiffening* effect and hence a loss of entropy. Because of the energy term, the miscibility is good at low temperature; when the temperature is raised, the entropy term becomes dominant and the miscibility decreases. This provides an explanation for the occurrence of phase separation when the temperature is raised.

A very special case is found when the above-mentioned types of *S*-*x* and *H*-*x* curves change their shape upon an increase of temperature, the *S*-*x* curve becoming a normal, convex type and the *H*-*x* curve convex. This might happen in the triethylamine-water system, for example, because of the stronger dissociation of the hydrate molecules as the temperature is increased. This can give rise to a *closed* miscibility gap, i.e. one that has both a lower and an upper critical solution temperature. A classic example of this is the ovoid miscibility gap in the nicotine-water system.

In *solid* phases no cases are known of a lower critical solution temperature, and only a few *S*-*x* curves that are not completely convex. So many crystal structures are possible for a solid phase that if such a tendency to a lowering of entropy and energy leads to the appearance of a second phase, the latter will always have a different crystal structure. Separation of this kind into phases of *different* structure cannot possibly produce a critical point.

Before proceeding, let us briefly recapitulate. As the key to some calculations of the phase diagram in the neighbourhood of the critical point we have used the fact that $\mathrm{d}^2G/\mathrm{d}x^2$ at that point is zero. We have followed this up with various considerations, and from the form of the miscibility gaps encountered we have drawn inferences regarding the shape of the *S*-*x* and *H*-*x* curves.

For our further calculations we shall limit ourselves to "dilute solutions", which will enable us to employ a variety of simplifying assumptions. As a first example we consider calculations relating to elastic "*lattice loosening*" (i.e. lowering of the Debye temperature). We choose this as our first example because it links up in a somewhat surprising way with what has gone before, the *S*-*x* and *H*-*x* curves presenting more or less a mirror image of the stiffening just discussed.

Some laws relating to dilute solutions

Lattice loosening

Whereas stiffening of a liquid mixture is accompanied by a lowering of energy and entropy, lattice loosening of a solid phase, which is due to the substitution of atoms that do not fit readily into the lattice, is accompanied by an increase of energy and an "excess of entropy". By the latter we mean that the entropy increase with x is greater than that according to Gibbs. We shall denote this excess by S_{ex}:

$$\begin{aligned} S_{\mathrm{ex}} &= S - S_{\mathrm{Gibbs}} \\ &= S + R\{x \ln x + (1-x)\ln(1-x)\}\,. \quad . \quad . \quad (11) \end{aligned}$$

For studying the phenomenon of lattice loosening we compared the solubility x_{e} of a number of elements (Mg, Cu, Si, Mn, Cr and Ni) in solid aluminium. The solubility in all these cases is low.

We again take formula (8) as our starting point:

$$\frac{\mathrm{d}x_{\mathrm{e}}}{\mathrm{d}T} = \frac{\mathrm{d}S/\mathrm{d}x - \Delta S/\Delta x}{\mathrm{d}^2G/\mathrm{d}x^2}\,.$$

First, we shall show how this formula can be considerably simplified by introducing certain assumptions. We shall not explain each step of this simplifying process in detail, but try to make the broad lines clear.

Where x is small we can write for H and S of a binary phase:

$$H = fx + gx^2 + \ldots \quad . \quad . \quad . \quad . \quad . \quad (12)$$

and

$$S = -R\{x \ln x + (1-x)\ln(1-x)\} + hx + ix^2 + \ldots \quad . \quad . \quad . \quad (13)$$

(taking $G = 0$ for $x = 0$). Simple calculation shows that for small x we may therefore equate $\mathrm{d}^2G/\mathrm{d}x^2$ with RT/x_{e}. For eq. (8) we can now write:

$$\frac{\mathrm{d}x_{\mathrm{e}}}{\mathrm{d}T} = \frac{\mathrm{d}S/\mathrm{d}x - \Delta S/\Delta x}{RT/x_{\mathrm{e}}}\,. \quad . \quad . \quad (14)$$

For an equilibrium, $\mathrm{d}G/\mathrm{d}x - \Delta G/\Delta x$ must be zero, while $G = H - TS$, and therefore:

[4] This also applies where only one of the solubilities decreases with increasing temperature.

$$\mathrm{d}S/\mathrm{d}x - \Delta S/\Delta x = \frac{\mathrm{d}H/\mathrm{d}x - \Delta H/\Delta x}{T}. \quad . \quad (15)$$

Insertion of (15) in (14) gives:

$$\frac{\mathrm{d}x_\mathrm{e}}{\mathrm{d}T} = \frac{\mathrm{d}H/\mathrm{d}x - \Delta H/\Delta x}{RT^2/x_\mathrm{e}}. \quad . \quad . \quad . \quad (16)$$

$\mathrm{d}H/\mathrm{d}x - \Delta H/\Delta x$ is independent of T provided that:

a) the composition x' of the coexisting phase is almost constant;

b) the specific heat c_p versus x does not depart too much from a linear function.

For convenience we assume further:

c) that x' is not only virtually constant but roughly equal to unity (which amounts to treating the coexisting phase as a component), and

d) that the two coexisting phases have the same structure.

We can then equate $\Delta H/\Delta x$ (and $\Delta S/\Delta x$) to zero and write:

$$\frac{\mathrm{d}x_\mathrm{e}}{\mathrm{d}T} = \frac{\mathrm{d}H/\mathrm{d}x}{RT^2/x_\mathrm{e}}, \quad . \quad . \quad . \quad . \quad . \quad . \quad (17)$$

or:

$$\frac{\mathrm{d}(R \ln x_\mathrm{e})}{\mathrm{d}(-1/T)} = \mathrm{d}H/\mathrm{d}x. \quad . \quad . \quad . \quad . \quad . \quad (18)$$

Since $\mathrm{d}H/\mathrm{d}x$ does not depend on T, integration yields:

$$\log x_\mathrm{e} = a - b/T. \quad . \quad . \quad . \quad . \quad . \quad (19)$$

where a and b are constants. In the systems referred to, and also in others, it has in fact been found that the logarithm of x_e is a linear function of $1/T$.

Before saying anything about the values of a and b for the various systems, we shall first comment on their physical significance. The constant b, the slope of the curve, is an *enthalpy* quantity, and in particular a measure of $\mathrm{d}H/\mathrm{d}x$, as can be seen immediately from eq. (18).

The constant a, which is found by extrapolating the point of intersection of the curve with the axis $1/T = 0$, is an *entropy* quantity, being a measure of $\mathrm{d}S_\mathrm{ex}/\mathrm{d}x$.

The relation between a and $\mathrm{d}S_\mathrm{ex}/\mathrm{d}x$ can be derived as follows. With the aid of (18) we find:

$$R \ln x_\mathrm{e} = -\frac{1}{T}\,\mathrm{d}H/\mathrm{d}x + R \ln x_\mathrm{extrap},$$

where x_extrap is the extrapolated value of x for $1/T = 0$. Using eq. (15), in which $\Delta S/\Delta x$ and $\Delta H/\Delta x$ are taken as zero, we arrive at:

$$R \ln x_\mathrm{e} = -\frac{\mathrm{d}S}{\mathrm{d}x} + R \ln x_\mathrm{extrap}.$$

Filling in (11), after some working-out, yields:

$$\frac{\mathrm{d}S_\mathrm{ex}}{\mathrm{d}x} = R \ln x_\mathrm{extrap} = \frac{Ra}{\log \mathrm{e}}. \quad . \quad . \quad (20)$$

When Mg, Cu, Si, Mn, Cr and Ni are dissolved in solid aluminium in that order, both a and b are found to increase. This behaviour is summed up in the rule: the more difficult it becomes to substitute in a solid a given small quantity of foreign atoms (greater $\mathrm{d}H/\mathrm{d}x$), the "looser" becomes the crystal lattice, which entails a higher S_ex [5].

The lattice loosening can be thought of as due to an increase in the percentage of empty lattice sites (vacancies) as a result of the substitution of the foreign atoms. The removal of an atom may be expected to cost less energy if its neighbour atom is one with which it has a less favourable energetic interaction. It is obvious that vacant lattice sites offer the atoms more opportunities to move in the lattice. It is indeed found that the modulus of elasticity of a metal shows a relatively marked drop when not readily soluble elements are dissolved in it [6].

It should again expressly be stated that we are concerned here only with *elastic* lattice loosening. The resistance to *plastic* deformation will on the contrary be increased by the presence of vacancies, as well as by the presence of other irregularities in the lattice.

We shall now continue with our derivation of fairly generally known laws applicable to dilute solutions, such as Raoult's and Van 't Hoff's laws.

Raoult's law

For H and S of a binary phase with small x we again write:

$H = fx + gx^2 + \ldots$ and

$S = -R\{x \ln x + (1-x)\ln(1-x)\} + hx + ix^2 + \ldots$

For the chemical potential [7],

$$\mu_1 = G - x\,\frac{\mathrm{d}G}{\mathrm{d}x},$$

we find, using these equations:

$$\mu_1 = RT \ln(1-x) + \alpha x^2 + \ldots \quad . \quad . \quad (21)$$

The parameters f and h now disappear, while g and i are expressed in α. By expanding $\ln(1-x)$ to $-x - \frac{1}{2}x^2 - \ldots$, we can write for μ_1:

$$\mu_1 = -RTx - (\tfrac{1}{2}RT - \alpha)x^2 + \ldots \quad . \quad . \quad (22)$$

Provided x is sufficiently small we may neglect the second and higher powers of x, thus eliminating all "mixing parameters" specific to a given system:

$$\mu_1 = -RTx. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (23)$$

This equation states among other things that in dilute solutions the partial pressure of the solvent decreases linearly with the (molar) concentration of the solute, *irrespective of the nature of the solute*. This is Raoult's law. How dilute the solution must be for this depends on the parameter α in (21).

Van 't Hoff's law

Using eq. (23) we shall now derive a relation that tells us something about the two equilibrium curves which, in a binary *T-x* diagram, begin at a melting point, a transition point or a boiling point. For the sake of the argument we take a melting point at the absolute temperature T_0, which is the point of origin of the liquidus and solidus curves (see *fig. 11*).

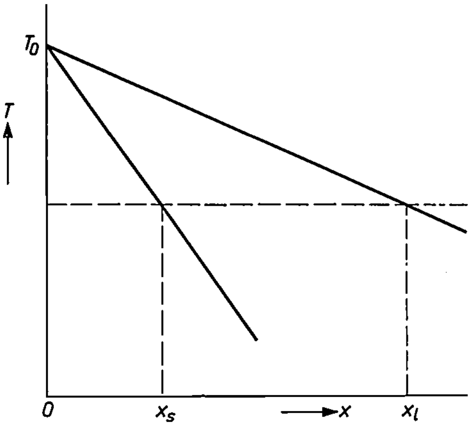


Fig. 11. For calculating Van 't Hoff's law.

We find a relation between the initial slopes of these curves from the condition that the values of μ of both coexisting phases must be equal. At the melting point T_0 of the pure substance A ($x = 0$) we have $G_{\text{liquid}} - G_{\text{solid}} = \mu_{1\ \text{liquid}} - \mu_{1\ \text{solid}} = 0$. If we change the temperature slightly to T, we then have $\mu_{1\ \text{liquid}} - \mu_{1\ \text{solid}} = (T_0 - T)\Delta S$, where ΔS is the entropy of fusion. If the equilibrium is to be maintained, the concentration of B in this solid must change from 0 to x_s and in the liquid to x_l. With the aid of eq. (23) we can write as the condition of equilibrium:

$$-RT\,x_l + RT\,x_s + (T_0 - T)\Delta S = 0,$$

or

$$RT\,(x_l - x_s) = (T_0 - T)\Delta S.$$

Since this expression only holds for small x_s and x_l, we may write it as follows:

$$\left(\frac{\mathrm{d}x_s}{\mathrm{d}T} - \frac{\mathrm{d}x_l}{\mathrm{d}T}\right)_{T=T_0} = \frac{\Delta S}{RT_0} = \frac{\Delta H}{RT_0{}^2}. \quad . \quad (24)$$

This is Van 't Hoff's equation, more familiar as the formula for freezing-point depression. For in the case where x_s is always zero (no mixed crystal formation) we can use this equation for calculating how far the melting point will drop as the result of dissolving a small quantity of a substance.

If mixed crystals *are* formed, this calculation is no longer possible and the equation gives us only the *difference* between the slopes of the two equilibrium lines concerned. For the rest, its application is very general; it also holds, for example, where the melting point (or the transition or boiling point) *rises* due to the addition of a second component.

In order to establish the two initial slopes separately — which would enable us to calculate the freezing-point depression in the case of mixed crystals — we should have to take μ_2 into account, which would mean that specific "mixing parameters" of the system would enter into the equation. The advantage of Van 't Hoff's equation is precisely that we can use it *without* further knowledge of the binary system, all we need to know being the entropy of fusion (or heat of fusion) and the melting point of the *pure* component.

The formula thus offers a simple means of checking certain phase diagrams. We shall give an example to explain this point, before turning finally to another method of verification, used for the highly important Fe-C diagram.

Verification of binary phase diagrams

In the *T-x* diagram of the Ag-Au system, as represented in the book by Hansen and Anderko [3] (see also fig. 9 in Part I), the solidus and liquidus curves have the shape of a "cigar", which is about 15 °C "thick" in the middle. From eq. (24) it is at once apparent that the cigar should be much more pointed at the ends. From considerations based on fig. 2 it can be derived that the melting range in the middle should be smaller than 2 °C [8]. A figure of 1.5 °C was in fact recently reported by White [9]. The solidus curve of the first diagram mentioned was found to be too low. This is often the case if the melt does not cool down very slowly. The first parts of the relevant alloy to solidify are rich in Au, and the Au content decreases as the temperature goes down. If, however, because of too rapid cooling, the diffusion in the solid state is unable, as it were, to keep pace with the change in concentration, zones relatively poor in Au and rich in Ag will be left behind. The "final freezing point" then lies at a lower temperature than if the whole had been cooled very slowly (and than the "initial melting point" after a homogenization anneal).

In the Fe-C system, too, the solidus curve of γ-Fe (*EJ* in *fig. 12*) used to be drawn on the whole much too low. It can immediately be seen that this diagram is wrong if one remembers that, although the curves *EJ*

[5] In reality, assumptions (c) and (d) are not valid for the systems under consideration, but this does not affect the conclusion arrived at here; see article quoted in footnote [1].

[6] W. Köster and W. Rauscher, Z. Metallkunde **39**, 111, 1948; C. Zener, Thermodynamics in physical metallurgy, Amer. Soc. Metals 1950, page 16.

[7] See equation (11) in Part I.

[8] J. L. Meijering, Proc. Symp. Nat. Phys. Lab. Teddington, 1958, Part II, paper No. 5A.

[9] J. L. White, Trans. Metallurg. Soc. AIME **215**, 178, 1959.

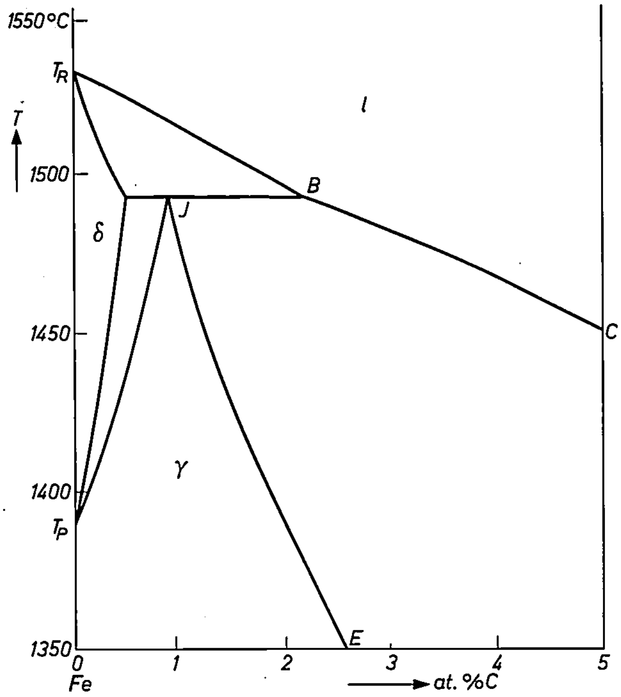


Fig. 12. Part of Fe-C diagram, as formerly drawn.

and *CB* cease to be stable at *J* and *B*, they do in principle continue, and indeed extend to the metastable melting point of γ-Fe. In other words, the metastable extensions of *EJ* and *CB* should intersect somewhere on the vertical Fe axis.

The question now remaining is: at what point on the axis? To answer this question, let us look at *fig. 13.* The diagram shows the mutual differences in Gibbs' free energy ΔG of the phases γ, δ and l of Fe as a function of temperature. The difference relates in all cases to the free energy of phase δ, that is to say the horizontal line $\Delta G = 0$ refers to this phase. It is further assumed that the ΔG-T lines of the other phases are also straight, so that their slope ΔS is independent of temperature. As the temperature range concerned is relatively small, this assumption is justified.

The calculation boils down to working out the triangle *PQR*. In the diagram, T_P is the transition temperature $\gamma \rightarrow \delta$, T_R is the melting temperature $\delta \rightarrow$ l, and T_Q is the unknown melting temperature $\gamma \rightarrow$ l. The slopes of *PQ* and *RQ* are given by experimentally determined values of ΔS, viz, the transition entropy for $\gamma \rightarrow \delta$ and the entropy of fusion for $\delta \rightarrow$ l. The entropy of fusion is 2.0 cal/degree, which is in good agreement with Richards' rule mentioned in Part I (1.1 $R = 2.2$ cal/degree). As expected, the entropy of transition is much smaller, being 0.13 cal/degree.

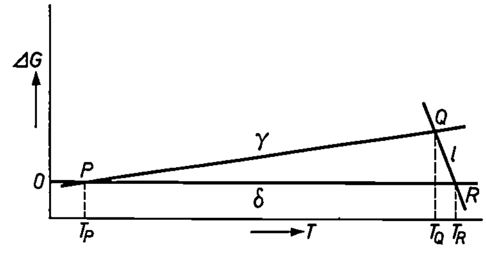


Fig. 13. For calculating the metastable melting point T_Q of γ-Fe.

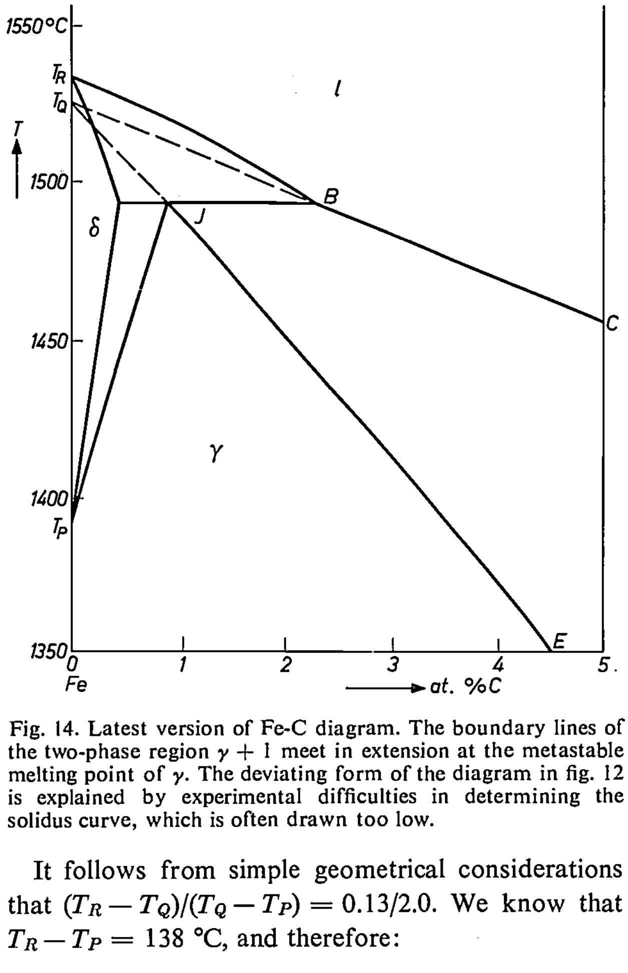


Fig. 14. Latest version of Fe-C diagram. The boundary lines of the two-phase region γ + l meet in extension at the metastable melting point of γ. The deviating form of the diagram in fig. 12 is explained by experimental difficulties in determining the solidus curve, which is often drawn too low.

It follows from simple geometrical considerations that $(T_R - T_Q)/(T_Q - T_P) = 0.13/2.0$. We know that $T_R - T_P = 138$ °C, and therefore:

$$T_R - T_Q = 138 \times 0.13/(2.0 + 0.13) = 8\ °\mathrm{C}.$$

The metastable extensions of the lines bounding the two-phase region γ + l in the Fe-C diagram should therefore meet 8 °C below the stable melting point on the Fe axis. In reality then, the curve *EJ* should arrive at *J* with a much smaller slope, and have the form roughly as shown in *fig. 14.*

Summary II. Among the subjects dealt with are calculations of the spinodal curve, the critical point and the boundary of a miscibility gap; calculations relating to phase diagrams with flat solidus or liquidus curves; the theory of elastic "lattice loosening" of the solid phase; the laws of Raoult and Van 't Hoff, and simple methods of verifying phase diagrams.

Volume 26, 1965, No. 2 pages 33-60 Published 12 th March 1965

PHILIPS TECHNICAL REVIEW

VOLUME 26, 1965, No. 3

A klystron multiplier for generating 0.8 mm waves

B. B. van Iperen and W. Kuypers

621.385.623.5

At the beginning of the 'fifties, when the most vigorous period in the development of microwave tubes was over, the view was generally held that it would be difficult, if not impossible, to generate wavelengths shorter than a few millimetres by means of the principles employed at that time. In most research centres concerned with the problem of generating shorter waves, other ways and means of doing this were therefore sought and studied.

In recent years, however, the "classical" principles have again been gaining ground. Thanks to improved theoretical insight and new technological possibilities, it has proved feasible to use the same principles for generating submillimetre waves. An example of this new development is discussed in this article. Here a tube is described that is capable of continuously generating waves of 0.8 mm with a power output of some tens of mW.

Introduction

One of the most widely employed sources for generating cm and mm waves is the reflex klystron. Tubes of this type are simple to use and deliver a continuous output which is amply sufficient for many purposes [1]. It has not previously been possible, however, to make continuous-wave reflex klystrons for the submillimetre region. The shortest wavelength achieved with these tubes is about 2 mm. The main reason for this limitation is that the reflex klystron is itself an oscillator and must therefore satisfy the starting condition for oscillation [2]. According to this condition the sum of electronic admittance and resonator admittance must be negative. Since the resonance admittance of the cavity resonator is inversely proportional to the square root of the wavelength, it becomes increasingly difficult at shorter wavelengths to satisfy the oscillation condition. In addition, there are also other serious problems of a less fundamental nature, such as the extremely small dimensions dictated by the wavelength, and the cathode loading and heat dissipation. It does not seem likely, therefore, that it will be possible in the foreseeable future to generate wavelengths of less than 1 mm with the reflex klystron.

In recent years research has been undertaken at the Philips Research Laboratories, Eindhoven, into the possibility of generating microwaves with the aid of a two-cavity klystron that does not work as an oscillator but as a frequency multiplier. This method has the great advantage that it is *not* necessary to satisfy the oscillation condition. The designer of the tube then has more freedom in the choice of certain parameters, so that other problems can be solved more easily than in the case of a reflex klystron.

The method is based on the fact that the velocity-modulated electron beam of a klystron can contain in addition to the fundamental frequency of the driving signal a large number of its harmonics. If the driving signal and the transit time between the two resonant cavities is suitably chosen, these higher harmonics can be relatively strong [3]. According to a simple theory, which disregards the interaction forces between the electrons, the amplitude of the tenth harmonic in the beam current can even be greater than half the magnitude of the first harmonic. In reality the situation is even more favourable. If the output cavity (catcher) is tuned to the tenth harmonic of the driving frequency,

Drs. B. B. van Iperen and Ir. W. Kuypers are research workers at Philips Research Laboratories, Eindhoven.

[1] B. B. van Iperen, Reflex klystrons for wavelengths of 4 and 2.5 mm, Philips tech. Rev. **21**, 221-228, 1959/60.

[2] See the article in reference [1], p. 225 et seq. (eq. (3)).

[3] This was first noted by D. L. Webster, J. appl. Phys. **10**, 501-508, 1939.

the beam delivers energy at ten times the frequency of the driving signal (*fig. 1*).

Some years ago tubes working on this principle were described, which were suitable for the generation of cm- and mm-waves [4]; the shortest wavelength reached was 5 mm, with an output power of 0.1 mW using the second harmonic. We have found, however, that submillimetre waves can also be generated with a klystron multiplier provided the velocity of the electrons is drastically increased [5]. In the tube described in this article, the beam voltage V_0 — 2 to 2.5 kV in reflex klystrons — is not less than 25 kV. This tube generates 0.87 mm waves having a continuous output power of 35 mW [6].

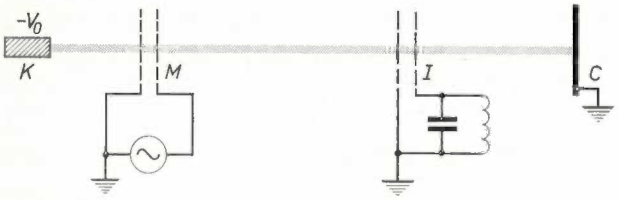


Fig. 1. Principle of klystron operation. K cathode. V_0 electron accelerating voltage (beam voltage). M two modulation (buncher) grids, between which the electron beam is velocity-modulated as a result of the alternating voltage — the driving signal — applied to M. In the drift space between M and I the velocity-modulation gradually gives rise to density modulation, because the electrons accelerated in M catch up with the earlier electrons that have been slowed down: the electron beam, which was a "direct current" at M, has become at the induction grids I an "alternating current" with the frequency of the modulating signal. In certain circumstances this alternating current contains many higher harmonics. The tube can be made to work as a *frequency-multiplier* by tuning the resonant circuit connected to I to the frequency of one of these higher harmonics. The tenth harmonic is the one used in the 0.8 mm tube. In reality, M and I are not pairs of grids with separate resonant circuits connected to them, but gridless interaction gaps in resonant cavities.

In the following section we shall explain the significance of raising the beam voltage, and say something about the factors that govern the optimum value of the beam current. In conclusion we shall discuss the construction and the principal characteristics of the new 0.8 mm tube.

Theoretical considerations

As stated, the greatest obstacle to the construction of tubes for very short wavelengths can be overcome by using a frequency multiplier instead of an oscillator. At the same time this makes the above-mentioned increase in the beam voltage V_0 possible. This considerably reduces the other difficulties, such as the fact that at very short waves the dimensions of the resonant cavities and their distance apart become extremely small, the limitations in regard to the permissible loading of the cathode, etc.

By increasing V_0 one increases what is termed the electronic wavelength, i.e. the distance travelled by the electrons in one cycle of the driving or output signal; this distance is proportional to $\sqrt{V_0}$. We will now explain the advantages to be gained by increasing the electronic wavelength.

a) If V_0 is increased the electron beam can be given a larger diameter, thus reducing the cathode loading. This is immediately apparent from the following consideration. To ensure good interaction between beam and resonant cavity the transit time of an electron in the alternating electric field of the interaction gap (see *fig. 2a*) must not be much longer than half a cycle of this field. So in the first place, the gap width d must not be much larger than half an electronic wavelength. This has not yet met the above-mentioned requirement, however, for even if d is relatively small the electric field extends along the axis of the resonant cavity over a length which is of the same order of magnitude as the diameter $2r$ of the interaction space. For good interaction, therefore, this diameter too should not be much greater than half an electronic wavelength. The same consequently applies to the beam diameter. (In our tube the beam diameter is governed by the electronic wavelength related to the *output signal*, this wavelength being much smaller than the one related to the input signal.)

b) The output power increases in proportion to V_0. This is bound up with the characteristics of the catcher cavity. At wavelengths below about 1 mm there are disadvantages in using resonant cavities as drawn in fig. 2*a*. In the first place, the re-entrant parts are then difficult to make because of their small dimensions. Moreover, there is a danger that these parts will be overheated by electron bombardement. We therefore

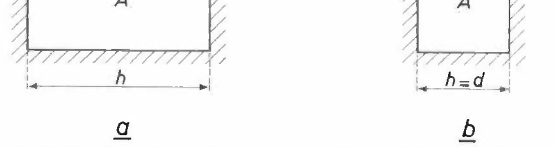


Fig. 2. *a*) Cross-section of a resonant cavity of the type frequently used in klystrons (re-entrant cavity), showing an electron beam (the shaded strip) in interaction with it. The beam passes the resonant cavity A through the hole B, in the wall of which a slit of width d is cut. The part of the hole near the slit, where the electric field is relatively strong, is the actual interaction space (see the sketched paths of the electric lines of force). For good interaction between resonant cavity and beam the diameter $2r$ of the interaction space and the gap width d should at the most be roughly equal to half the "electronic wavelength" of the beam. *b*) Example of a flat resonant cavity, as used in the tube for 0.8 mm waves. The letters have the same meaning as in (*a*). The re-entrant parts are missing here, so that the gap width d is equal to the height h. A resonant cavity of this kind can be made for extremely short wavelengths and gives better heat dissipation.

use what is called a "flat cavity" (fig. 2*b*). The resonant frequency of such a cavity is primarily determined by the dimensions perpendicular to the beam. In the first instance the height h can be freely chosen. It is advantageous to choose it as high as possible, because this increases the resonance impedance (which is proportional to h^2) and thus the output power also. The extent to which h can be increased, however, is limited by the fact that h here is equal to the gap width, and must not therefore be greater than about half an electronic wavelength. Thus, in order to obtain a reasonable output power it is equally important to have a high V_0, for as stated the output power is proportional to the resonance impedance and hence to V_0.

c) When the beam voltage is raised, the optimum distance between the centres of the two resonant cavities also increases; at constant input power the increase is more than proportional. By choosing the beam voltage in the region of 25 kV one ensures that there is sufficient space to accommodate resonant cavities of optimum dimensions together with their tuning devices and connections, and also that the heat generated in the cavities is adequately dissipated.

The beam current

The extent to which the electron beam current in a klystron can be increased in order to step up the power output is limited by the influence of mutual repulsion between the electrons. If the density of the beam current is too high, this repulsion tends to oppose electron bunching in the beam and the formation of harmonics is disturbed. Calculation of the influence of this effect is a complicated problem that can only be solved with the aid of an electronic computer, and its treatment does not come within the scope of this article [7].

Our calculations have shown, remarkably enough, that to some extent a higher current density initially favours the production of higher harmonics; the adverse effect referred to only appears later on. Moreover, the optimum beam current is found to depend, among other things, on the rank number of the relevant harmonic (i.e. on the factor by which the frequency is multiplied) and on the driving power; the voltage V_0, on the other hand, has little influence. These calculations show that in our tube the optimum beam current for a driving power of about 10 W is roughly 20 mA.

An experimental tube for 0.8 mm

A simplified cross-section of the tube is shown in *fig. 3*. From left to right can be seen successively the electron gun *K*, the buncher cavity *M* with tuning device and input waveguide, the block *I* containing the catcher cavity and the output waveguide, and the collector *C*. The catcher cavity itself is too small to be shown in this figure. Also to be seen are the pole pieces *P* of the electromagnet which encloses the tube

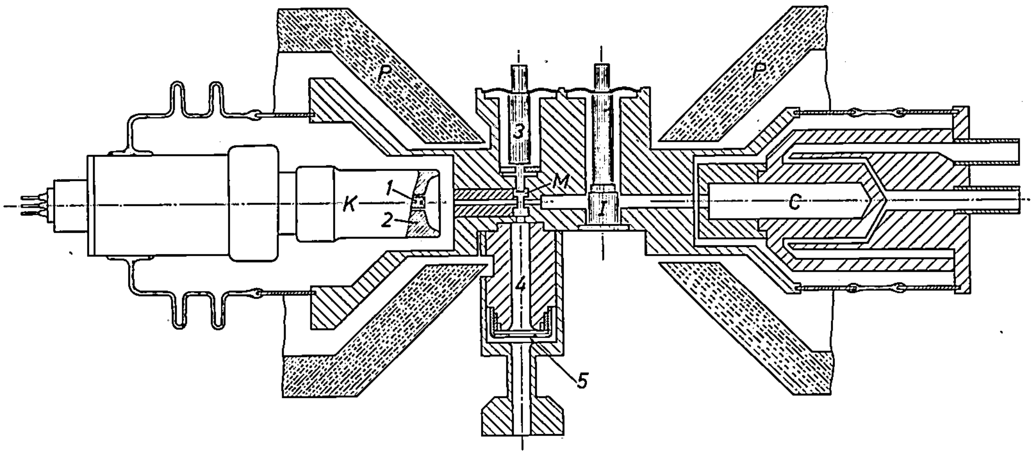


Fig. 3. Simplified cross-section, 1.5 times full[size, of the klystron multiplier for 0.87 mm. *K* electron gun with cathode *1* and focussing electrode *2*. *M* modulation cavity (buncher) tuned to 8.7 mm. *3* tuning plunger, movable by means of a diaphragm in the outside wall of the tube. *4* input waveguide with glass vacuum-tight window *5*. *I* the block containing the catcher cavity with the output waveguide and accessories. (A cross-section of the catcher cavity on a larger scale can be seen in fig. 6.) *C* collector (water-cooled). *P* pole pieces of focussing magnet.

[4] See for instance W. H. Cornetet, Jr., IRE Trans. on Electron Devices **ED-6**, 236-241, 1959.

[5] See B. B. van Iperen, Proc. IEEE **51**, 935-937, 1963. In this article a 2.5 mm tube is described in which this increase in beam voltage has been applied for the first time.

[6] See also B. B. van Iperen and W. Kuypers, Experimental CW klystron multiplier for submillimetre waves, Philips Res. Repts **20**, 462-468, 1965 (No. 4).

[7] See B. B. van Iperen and H. J. C. A. Nunnink, Harmonics in velocity-modulated cylindrical electron beams, Philips Res. Repts **20**, 432-461, 1965 (No. 4). See also B. B. van Iperen and H. J. C. A. Nunnink, C.R. 5me Congrès int. tubes hyperfréquences, Paris 1964.

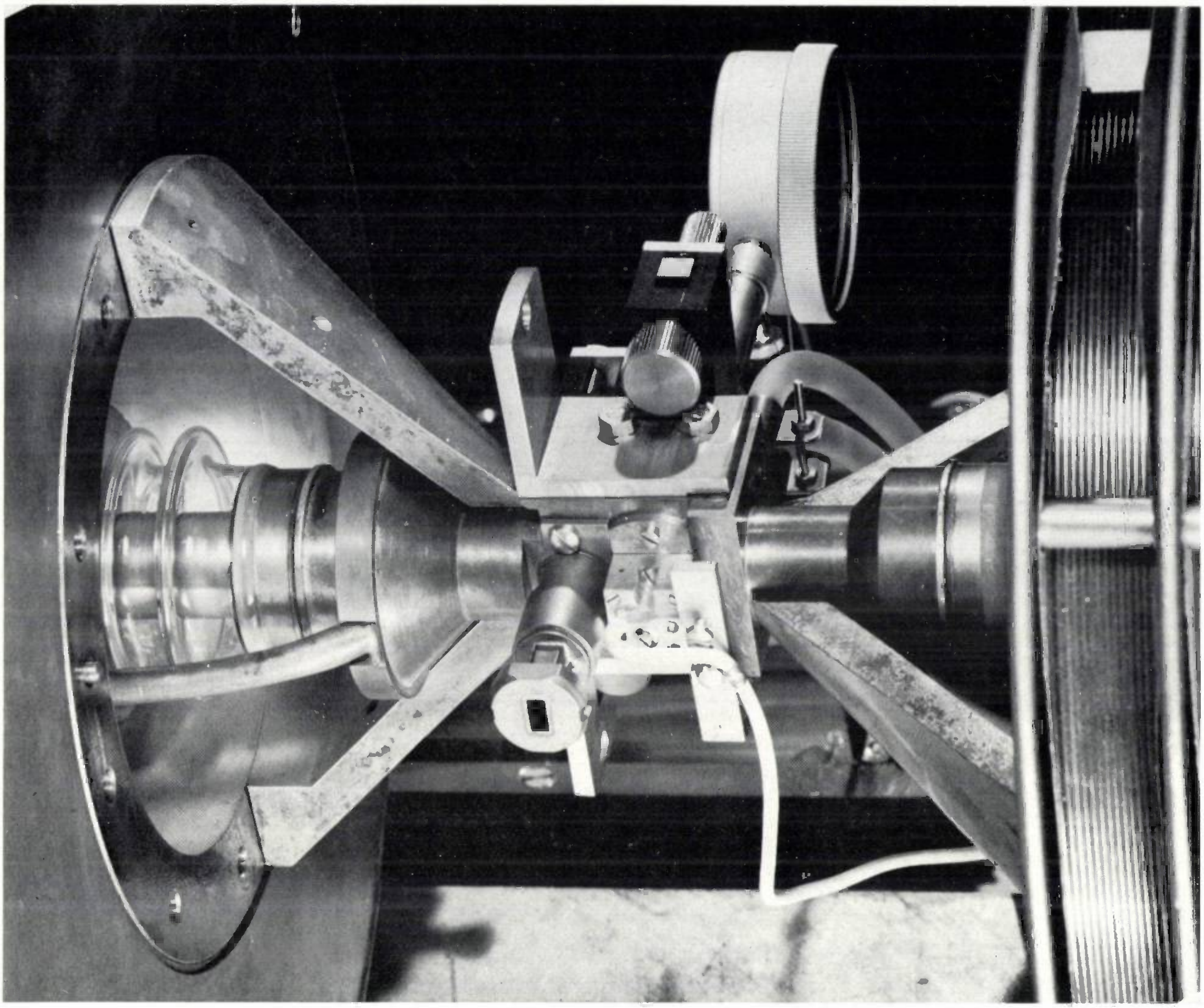


Fig. 4. The 0.8 mm klystron in the magnet. To make the ends of the tube visible, half of each of the tapering pole pieces has been removed; on the left is the gun section. The connection for the input waveguide can be seen at the front in the centre. On the right, partly behind the white cable, is a bolometer employing a thermistor and matching transformer, for measuring the output power. In the centre, above the tube, is the tuning knob for the catcher cavity. In the background is a micrometer for reading the position of the tuning plunger in the catcher cavity.

and which supplies the magnetic field needed for focussing the electron beam. *Fig. 4* shows a photograph of the tube inside the magnet.

The electron gun is a Pierce gun, which gives an electron beam current of 20 mA at 25 kV. The beam diameter is about 0.2 mm. Since the beam converges, the current density at the cathode surface is limited to 7 A/cm^2; this value is well below the permissible maximum for a modern L-cathode. To obtain a well-shaped beam only a circular part of the front face of the cathode is used for emission, the remainder having been made non-emissive by a special treatment [8]. This non-emissive portion thus acts as a focussing electrode at cathode potential.

The driving signal is supplied by an 8 mm klystron. The catcher cavity is a flat resonant cavity of rectangular form, which oscillates in the TE$_{012}$ mode. Because of this shape and resonator mode it is possible to make the catcher cavity tunable. This can be understood as follows.

In the TE$_{012}$ mode the alternating electric field is in the x direction (see *fig. 5*) while the alternating magnetic field in this direction is zero. Both fields are independent of x but vary sinusoidally in the y and z directions, in such a way that a half wave is along the y axis and a full wave along the z axis inside the cavity. The beam aperture is located at a position of maximum electric field, i.e. at a distance $l/4$ from one of the end faces (l being the length of the cavity along the z axis). At a

[8] The process used for this was found and developed by Dr. R. Levi, Philips Metalonics, Mt. Vernon, N.Y., U.S.A. It is described by E. S. Rittner and R. Levi in J. appl. Phys. 33, 2336, 1962.

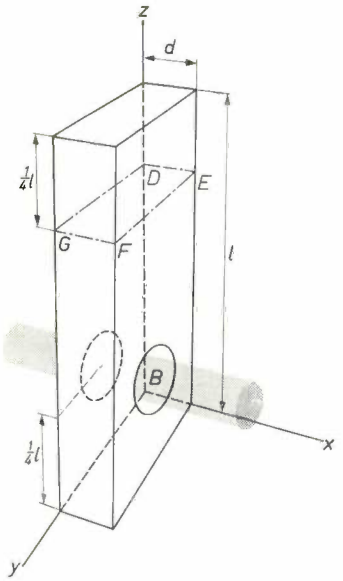


Fig. 5. Rectangular flat cavity, used as the catcher cavity for 0.87 mm waves. The cavity is excited in the TE$_{012}$ mode, the alternating electric field then being in the x direction and the alternating magnetic field in this direction being zero. B apertures. The cavity can be cut along the lines $DEFG$ without disturbing the resonator mode. This fact is turned to use for making the cavity tunable.

distance $l/4$ from the end faces the y component of the magnetic field is zero. Since the current in the wall is at right-angles to the adjacent magnetic field, the z component of the wall current is also zero at that position. The resonant cavity can thus be cut through at that point without upsetting the resonator mode.

The latter fact is turned to use for making the catcher cavity tunable. For this purpose it is composed of two parts, one of which is $3l/4$ long and the other $l/4$. The first part contains the beam aperture and is a fixed part of the tube. The other can be moved in the z direction by means of a diaphragm in the wall of the tube. In this way the cavity can be tuned by about 5% without significantly changing the impedance.

The construction of the catcher cavity is shown in the perspective drawing in *fig. 6*. The figure also shows the output waveguide, which is separated from the actual resonant cavity by two partitions, with a coupling hole between them. The other end of this tapering waveguide is made vacuum-tight by a mica window 5 μm thick. The small dimensions — the volume of the resonant cavity is about 0.1 mm^3 — and the high demands made on the dimensional accuracy and on the quality of the surfaces, make it necessary to use special fabrication techniques such as "hobbing", "optical" milling and diffusion welding. By these means it proves possible, however, to achieve a quality factor and an impedance which are not far below the theoretical values.

The most important characteristics

To conclude we shall mention some of the principal characteristics of the new tube. *Fig. 7* shows the result of measurements of the output power P_o at 0.87 mm as a function of the driving power P_i (wavelength 8.7 mm).

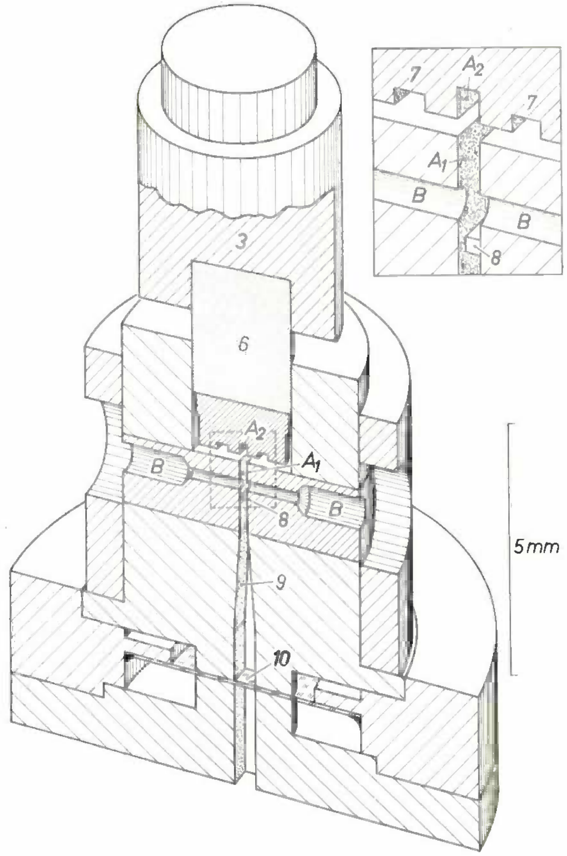


Fig. 6. Perspective sketch of the catcher cavity for 0.87 mm. The cavity has been cut along a plane through the beam axis and parallel with the xz plane (see fig. 5), and the front half removed. A_1 rigid section of the cavity. A_2 movable part of the cavity, which is fixed to the tuning plunger by means of a sapphire rod *6*. The plunger can be moved axially via a diaphragm in the tube wall. *7* choke grooves to prevent leakage of the high-frequency field through the gap between A_1 and A_2. *8* partition between resonant cavity and output waveguide. The coupling hole is between this partition and the one in the cut-away half. *9* tapered output waveguide. *10* mica window, 5 μm thick, for vacuum seal. B beam aperture.

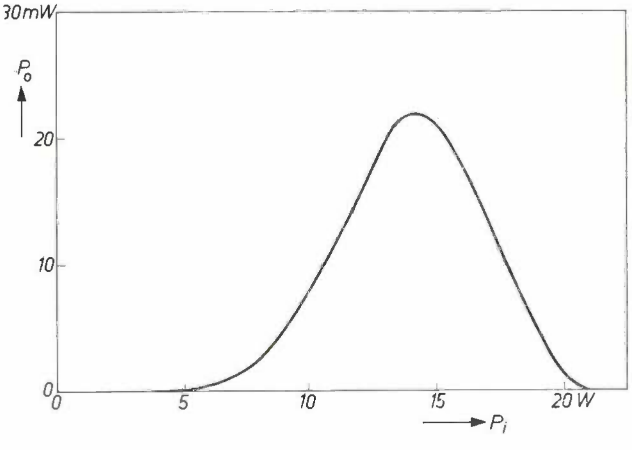


Fig. 7. The continuous wave output power P_o at a wavelength of 0.87 mm as a function of the driving power P_i at 8.7 mm. For this plot the beam voltage was 25 kV, the beam current 20 mA and the focussing magnetic field 0.37 Wb/m^2 (3700 gauss). The curve is rather broad, and therefore P_i does not have to be stabilized so drastically in order to obtain a reasonably constant output power.

The curve relates to a beam voltage of 25 kV and a beam current of 20 mA. This plot is in reasonable agreement with the calculations.

If one compares the maximum output power attainable at this beam voltage and current with the driving power which it calls for, it can be seen that the conversion of 8 mm waves into 0.8 mm waves involves a conversion loss of about 26 dB. In view of the wavelength and the high order of harmonic involved, this is a very satisfactory value.

As can be seen, the curve in fig. 7 is fairly broad; to obtain a reasonably constant output power it is therefore not necessary to stabilize the driving power too rigorously.

There are two reasons, which apply to every klystron, why the output power cannot go on rising indefinitely with the input power. In the first place, the amplitude of each of the harmonics in the beam is never larger than twice the DC value of the unmodulated beam. In the second place, the position of the point where the bunching is strongest, and thus where the maximum output power can be expected, depends on the input power: there is only one value of the input power where that point coincides exactly with the location of the catcher cavity. This explains why the curve shows a maximum.

It has been found that, with a low driving power, at the most 0.5% of the beam electrons are lost on the way; the rest reach the collector. As the driving power is raised the loss increases, and at the optimum value of the driving power it amounts to roughly 7%. These figures relate to a focussing magnetic field of 0.37 Wb/m^2 (3700 gauss).

Fig. 8 shows a plot of $P_{0\,\mathrm{max}}$, the maximum output power at 0.87 mm, and $P_{\mathrm{i\,max}}$, the associated driving power at 8.7 mm, as functions of the beam voltage at constant perveance — i.e. the ratio of the beam current to $V_0^{3/2}$. It can be seen from the figure that an output power of 1 mW is already achieved at a beam voltage of 16 kV, where the beam current is 10 mA and the driving power 8 W. At 28 kV, the highest voltage at which measurements were taken, the output power is 35 mW at a driving power of only 15 W. The measured values of the driving power are within 20% of the values calculated using the theory given in the article referred to [7]. This theory also seems to be confirmed by the measurements as far as the output power is concerned. The difference in this case was estimated at no more than 30%. An accurate comparison here is not possible at present, since the necessary measuring methods in the submillimetre range are not yet sufficiently advanced.

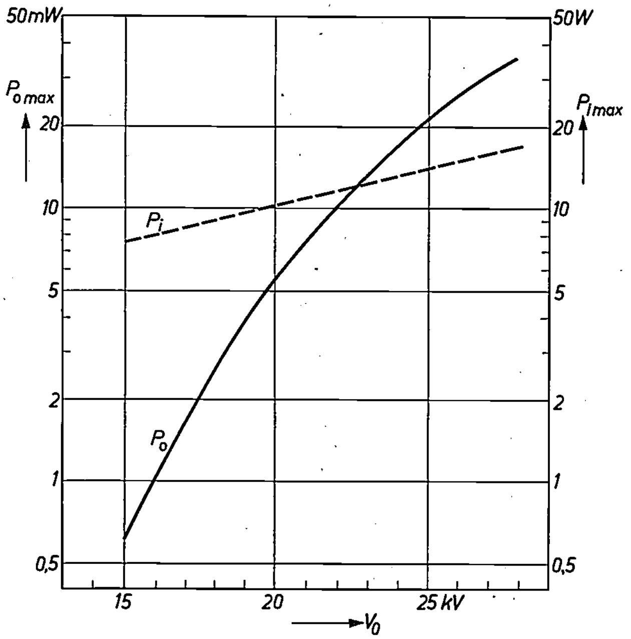


Fig. 8. The maximum output power $P_{0\,\mathrm{max}}$ and the appertaining driving power $P_{\mathrm{i\,max}}$ as a function of the beam voltage V_0 at constant perveance (in this case 5×10^{-9} A/V$^{3/2}$).

All in all it may be said that it has proved possible to generate 0.8 mm waves with reasonably high continuous power by means of a klystron frequency multiplier. There are reasonable prospects that it will be feasible with this method to penetrate still further into the submillimetre range within the foreseeable future.

Summary. Submillimetre waves (wavelength 0.87 mm) can now be generated with a continuous power of some tens of mW by means of a klystron frequency multiplier (the catcher cavity being tuned to the tenth harmonic of the driving frequency). This result was mainly obtained by choosing the beam voltage (25 kV) much higher than in reflex klystrons and earlier klystron multipliers (2-3.5 kV). This has made it possible, for example: a) to increase the beam diameter (this is about 0.2 mm), which reduces the cathode load; b) to use a flat resonator as the catcher cavity (better cooling and easier to make) without substantially sacrificing output power; c) to increase the distance between the buncher and catcher cavities. The latter is rectangular, oscillates in the TE$_{102}$ mode, and can be mechanically tuned by about 5%. The beam (20 mA) is focussed by a magnetic field of 0.37 Wb/m^2 (3700 gauss). The maximum power output achieved (35 mW) required an input power of only about 15 W.

Developments in the field of electronic computers during the last decade

W. Nijenhuis and H. van de Weg

681.14-523.8

The rapid development of electronic computers as regards computing speed, reliability and capacity is mainly due to the advances made in the technology of transistors and diodes, magnetic memory cores and magnetic recording. These various points are discussed in this article, and the possible lines along which present-day techniques may develop are indicated.

Introduction

The electronic computer plays an indispensable part in modern society; we could not imagine being without it. If we realize that the first electronic computer, ENIAC, was completed in 1947 and that Remington Rand installed the first commercial machine in 1951, then it will be clear that the development of these machines since then has been explosive. This rapid development comes out very strongly on scanning through the latest edition of "A survey of domestic electronic digital computing systems" from the middle of 1961; but it may also be seen that a great deal of pioneer work and experimentation was done in those early years: of the list of computers given in this publication, 85 of the 170 types considered are represented by a single model only, while about 90% of the total number of machines manufactured belong to only 15 types. We may assume that in that initial period money was spent too easily for an attractive idea, a brilliant brainwave, without thinking out the consequences with respect to a complete system beforehand, with the result that the pilot model often did not go into series production.

The limited number of successes makes it easier to pick out a few technical lines of development while disregarding the many side-tracks, and to give a brief account of these lines, as is intended here. We shall also largely omit specialized developments, particularly those in the military sector. These may be interesting, and may sometimes be the precursors of the techniques of tomorrow as a result of the sometimes exorbitantly high demands which are made; but the price factor is so subordinated to other arguments in these cases that they cannot be taken as a real reflection of the present-day techniques.

After the first period of amazement that it was *possible* to make an electronic computer was over, all attention was focussed on the problem of *reliability:* next to the price the reliability is the most important criterion for judging a particular construction. On the one hand the reliability of the components is important, and on the other hand the reliability of the connections, e.g. plug connections and soldered connections. The ENIAC may be cited as an example of how things were with the reliability in the early days. When this machine was switched on, usually at least a few of the 18 000 electronic valves failed. A problem was often worked out several times, the solution being considered to be correct if it was the same twice.

It proved possible to increase the reliability of the components by improved quality control, but initially this progress hardly kept pace with the increasing complexity of the machines, which led to much greater chances of something breaking down. Many will still remember the time before 1957, when computers only did 7 hours' useful work in each planned 10 hours, and the average time between two breakdowns was 2 to 4 hours. It was only once in a blue moon that a machine would work for a whole day without trouble.

The situation is now much better. In the first place, in addition to the above-mentioned quality control, better methods of construction have been introduced. These will be discussed below in more detail. But apart from this, various ways have been worked out to deal with an error once it occurs. "Diagnostic" programmes have been developed which make the computer itself indicate which part of it has failed. The computer

Ir. W. Nijenhuis and Ir. H. van de Weg (deputy director) are research workers at Philips Research Laboratories, Eindhoven. — This article is chiefly based on a talk which Ir. Van de Weg gave to the Dutch Computer Society in April 1964.

has also been split into a number of functional units and groups of functional units, with corresponding alarm devices to localize the site of the trouble as closely as possible. As a result, one can now find in sales and hire contracts for electronic computers a guarantee that the average useful machine time will be more than say 90%. The actual performances are usually even better. With a modern electronic computer, it should be possible to repair any trouble in less than half an hour.

As a special example in this connection, we may mention the new electronic telephone exchange of the Bell Laboratories in the United States [1]. Electronic telephone exchanges have much in common with electronic computers. They also use memories, in order to be able to store subscribers' numbers temporarily and they are built up of the same types of unit as computers (mainly flip-flops and gate circuits). These units are nowadays mounted on phenolic or glass-epoxy boards, with printed wiring ("printed-wiring cards"). The designers of the Bell system have aimed at making their diagnostic programmes so refined that the trouble can be pinned down to the printed-wiring board in question, so that maintenance can be done by unskilled workers. It must be conceded that in a certain sense an even greater reliability must be demanded of a telephone exchange than of normal commercial computers but we can expect similar maintenance methods for computers in the future.

Although the remarks made so far about the reliability have been on the whole optimistic, we must not overlook a few less happy points. In the first place, it is true that the reliability of the electronic circuits is now very reasonable, but this is much less so with the electromechanical equipment which we find at the periphery of the computer. Most trouble is still caused by this equipment, and this can only partially be prevented by frequent maintenance. Secondly, no computer is yet completely reliable in the sense that the user can be sure that it will work at a given time, e.g. Wednesday from half past two to five o'clock. This guarantee cannot yet be given for a computer, but it *is* demanded of the above-mentioned telephone exchange.

The reliability can be made especially great by increasing the *redundancy* of the circuit. For example, each valve can be replaced by two valves in parallel. A resistance *R* can be realized (as is also done for amplifiers in under-sea cables) by means of four resistors *R* connected two by two in series and in parallel, so that if one of these should break down or short-circuit there will still be a connection via resistors. One has another form of redundancy if the system contains more than one switching unit for a given function. The system can determine during operation whether a switching unit is working properly, and if not can look for an intact switching unit; the result is that a defect need not cause the whole system to cease operation. Computer techniques could gain a lot from telephony techniques in this respect. Very little redundancy can be seen in the circuits of present-day computers, even in those cases where it is desired to have a computer in service 24-hours a day without interruption. In this case a whole computer is often kept in reserve.

We will now go into some more detail about the history of the various techniques introduced into the field of electronic computers in the course of their development, which apart from a reduction in the price have mainly led to an improvement in reliability. We shall also pay a great deal of attention to the big increase in *speed* in nearly all parts of an electronic computer.

Components and logic circuits

Ten years ago, *valves* (mainly double triodes) were still in general use in computers, while active research was being carried out on the use of the point-contact transistor, invented in 1948. In Britain [2] and the United States [3], some computers were indeed made with point-contact transistors. This transistor can act as a high-speed electronic switch, with switching times of 0.1 - 10 microseconds. It was however rather fragile and difficult to make reproducibly, so that it was quickly replaced by the *junction transistor*. This was originally made by the pulling method, until the simpler alloying procedure was discovered. The junction transistor, originally announced as a low-frequency amplifying element, was made suitable for faster and faster circuits by improved methods of fabrication. While the best alloyed transistors made possible switching times of about 1 μs, the introduction of the diffusion technique has led to transistors which allow switching times of a few nanoseconds. Some details about these various types of transistors and their speeds are given in *Table I*.

Table I. Switching time obtainable with various kinds of transistors.

Type of transistor	Year of Introduction	Switching time (nanoseconds)
Point-contact transistor	1948	10 000-100
Junction transistor:		
Pulled	1949	ditto
Alloyed	1954	ditto
Diffused	1956	100-5
Planar	1960	100-5
Field-effect transistor	1962	1000-20

The technique of epitaxial crystal growth and the use of masking techniques have contributed to the improvement of the properties of the transistor, e.g. to the lowering of the knee voltage of the transistor characteristic. Moreover, about 1957 the change was made from germanium to silicon, which led to a reduced leakage current and better stability. The "planar technique" allows *P-N* junctions in silicon to be protected against the influence of the atmosphere by a SiO_2 film during their production, thus allowing a further improvement in reproducibility and life. This technique also proved to offer further possibilities which we shall discuss below (in connection with microminiaturization).

These methods of fabrication also allowed the production of the field-effect transistor — long proposed on paper.

The development of *diodes* more or less followed that of the transistor, after the change-over in about 1950 from the old selenium rectifiers to germanium point-contact diodes. If diodes are to retain their rectifying effect even at high frequencies (high switching speeds) then a low value of the internal capacitance, among other things, is of great importance. Point contact diodes satisfied this demand, but like point-contact transistors they were not sufficiently robust. Sturdier diodes could be made by the alloying technique. Very fast diodes can be obtained by a method in which the alloying is brought about during the attachment of the lead — a gold wire in this case. In general, diodes can be made faster by introducing certain impurities ("killers") into the germanium or silicon, of which gold is the most important example. Finally, the introduction of the Schottky diode, made by applying a thin film of metal to the silicon, opened the possibility of increasing the speed of the diodes even further. In *Table II*, the switching times of the various diodes used are shown.

Table II. Switching times of various types of diodes.

Type of Diode	Year of Introduction	Switching time (nanoseconds)
Point-contact diode	1947	100-0.1
Alloyed	1950	≈5000
"Gold-bonded diode"	1955	50-5
Planar	1960	50-5 *)
Schottky diode	1963	<0.2
Tunnel diode	1958	5-0.1
Boff diode	1962	0.2

*) If small and with gold killer.

This table also includes details of two types of diodes which are not used for their rectifying effect, viz the tunnel diode and the Boff diode. A few remarks follow concerning these two types.

The tunnel diode, discovered by Esaki in 1958 has two different functions as a result of the form of its characteristic, viz, as bistable element and as amplifier. It initially seemed very promising as a high-frequency logic element, owing to its inherently very fast switching mechanism, but the difficulty of making it reproducibly and the low voltage difference between the two stable states gave so much trouble in large systems that the interest in this component quickly faded. At present the tunnel diode is mainly regarded as a useful auxiliary element for increasing the speed of transistor circuits, e.g. the circuit for transmitting the carry to the following digit position in an adder; further as a memory element for small hyper-high-speed memories and for very high-speed counting circuits, mainly for applications in nuclear physics.

The Boff diode (also called the "snap-off diode") which acts as a delay element giving delays of the order of a few tenths of a microsecond, is of recent date. It is based on the fact that an initially conducting diode which is suddenly exposed to a voltage in the cut-off direction still passes current for a short time in this direction, the time depending on the value of the original forward current. The conduction in the cut-off direction ends with a very steep trailing edge which gives the very short switching time. The possibilities of this diode, e.g. in combination with the tunnel diode, are being investigated at present [4].

We shall now give a brief discussion of the most widely used *passive elements* (resistors and capacitors), together with a consideration of the development of the circuits.

At the start of the '50's, normal radio components were still used in combination with valves, which gave rather big constructions: the average packing density was about 0.1 components per cm^3. As the components became more reliable and stable, their design was altered to match the functional construction of the logic units in which they were used. This allowed the packing density to be increased to about 0.5 components per cm^3. The introduction of the transistor made it possible to reduce the dissipation to much smaller values, as a result of which resistors for lower powers (0.05 watt) and thus with smaller dimensions could be used.

[1] No. 1 Electronic Switching System, Bell Syst. tech. J. **43**, 1831-2592, 1964 (No. 5, parts 1 and 2).
[2] E. H. Cooke-Yarborough *et al.*, Proc. IEE **103 B**, suppl. No. 3, p. 361, 1956.
[3] J. H. Felker, Proc. IRE **40**, 1584, 1952.
[4] B. E. Sear, Charge controlled nanosecond logic circuitry, Proc. IEEE **51**, 1215-1227, 1963.

Use of these components, combined with printed wiring on a laminated or glass-epoxy board, led to the above-mentioned printed-wiring cards, which allowed the packing density to be increased by a further factor of 5, to 2.5 components per cm^3. This also greatly simplified the fabrication: all components could now be soldered on to the board in a single operation by dip soldering. If necessary, the units thus obtained could be protected against climatic influences by encapsulating them in plastic.

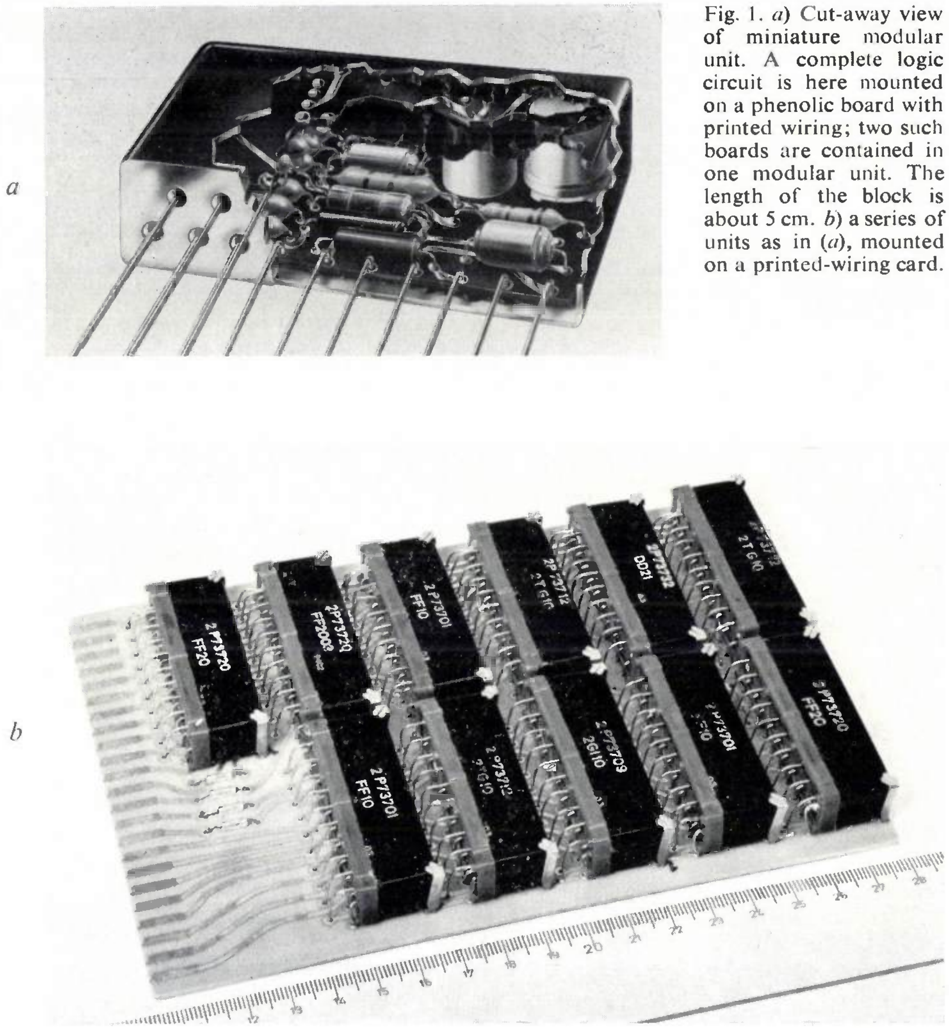


Fig. 1. *a*) Cut-away view of miniature modular unit. A complete logic circuit is here mounted on a phenolic board with printed wiring; two such boards are contained in one modular unit. The length of the block is about 5 cm. *b*) a series of units as in (*a*), mounted on a printed-wiring card.

If the connections between these units are also provided by printed-wiring cards, a very compact construction is obtained. This technique has already been in common use for more than 5 years, and is known as the *miniature technique* (*fig. 1a, b*).

We are now however in the middle of a spectacular further development as regards miniaturization. Miniaturization has always been an aim in the electronic industry. In some cases it was an end in itself, e.g. for airborne equipment; in computers it is also a means of improving the reliability. In the above-mentioned development, known as *microminiaturization*, the circuits are made by an integrated procedure, i.e. the connections between a number of components are made at the same time as the components themselves. One can then expect higher reliability because less soldered connections are needed in the circuits. One also hopes that the smaller circuits will give higher switching speeds.

We may distinguish two basically different methods for making these integrated circuits, viz the *thin-film* technique, based on the methods used for making resistors and capacitors, and the *monolithic technique*, which is based on the modern planar technique used for manufacturing transistors.

In the thin-film technique, a number of resistors and capacitors, together with the necessary connections, are evaporated on to a glass substrate in the form of a thin film. The output contacts of these units are also evaporated on to the glass. One difficulty connected with this technique is that the transistors and diodes needed must still be soldered on to the configuration in one way or another. Larger units are made e.g. by piling a number of plates together (*fig. 2*). These *microcircuits* allow the packing density to be increased by a further five times, to about 15 components per cm^3.

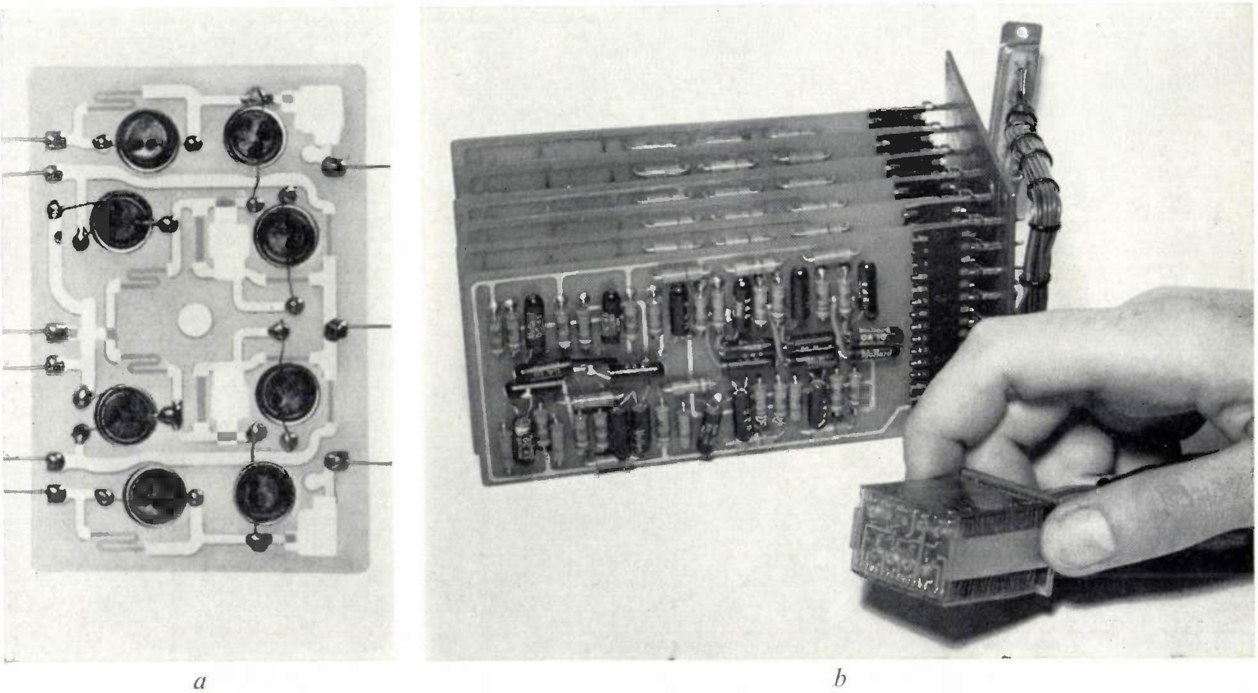


Fig. 2. Microminiaturization. *a*) Microcircuit on glass substrate. Length 3 cm. The circuit shown (one stage of a shift register for 100 kc/s) contains two transistors, six diodes, 10 resistors and four capacitors. *b*) A number of microcircuits stacked together to form a larger unit. In the background may be seen an equivalent unit in the normal miniature technique (the boards here are not encapsulated to form modular units as in fig. 1).

In the monolithic technique, the diodes and transistors are made very easily by forming them by the planar technique in a silicon crystal; but here it is the resistors and capacitors which are not so easy to make. This trouble is got round by using as a capacitor a *P-N* junction biased in the reverse direction, and as a resistor a diffused strip of a semi-conductor material of appropriate conductivity, insulated from the rest of the crystal by a *P-N* junction biased in the reverse direction. In principle, this technique can give much greater packing densities than even in the microcircuits; if we estimate the improvement in packing density as being increased by a further factor of 5, we certainly are on the safe side. The factor that can be obtained depends mainly on the art of connecting these little "chips", which only have a surface area of a few mm^2, with miniature leads to other units (*fig. 3*).

By 1960, this technology was so far advanced that the manufacturers (in the first place Texas Instruments and Fairchild) saw that microminiaturization might well become an economic proposition. This expectation has not yet been realized, but hopes are still high that the prices of these circuits will fall so that in a few years they will be able to compete with conventional circuits.

It cannot yet be predicted which of the two above mentioned techniques will finally win the race. If it proves possible to evaporate still smaller resistors and capacitors with narrow tolerances on to the monolithic circuits, this will give very elegant units. However, a possible factor in favour of the thin-film technique is the fact that a method has now been developed for making field-effect transistors and diodes (the Schottky diode) by a thin-film technique too.

There are still practical problems in connection with the applications of these small circuits. There is the question of the heat dissipated, which imposes a limit on the packing density. Further, we have the above-mentioned problem of reliably *connecting up* a number of these small units with the aid of miniature techniques This would allow larger functional blocks to be made, which could then be combined by conventional means (printed wiring, plugs and sockets). In this connection, the question immediately arises as to how many units should optimally be included in one block. Without going into details, we may say that here a number of contradictory factors, which are partly dependent on the state of the art, play a rôle: bigger blocks may be cheaper to manufacture, but the smaller a block, the more easily can one dispense with the demand that it should be repairable; in this connection the price of the block, the life and reliability of its components, the costs of skilled maintenance personnel and the price of the necessary spare units all play a part. Since part of the micro-logic must be used in combination with conventional equipment in the memory and peripheral equipment of the computer, additional demands are sure to be made on the assembly method chosen.

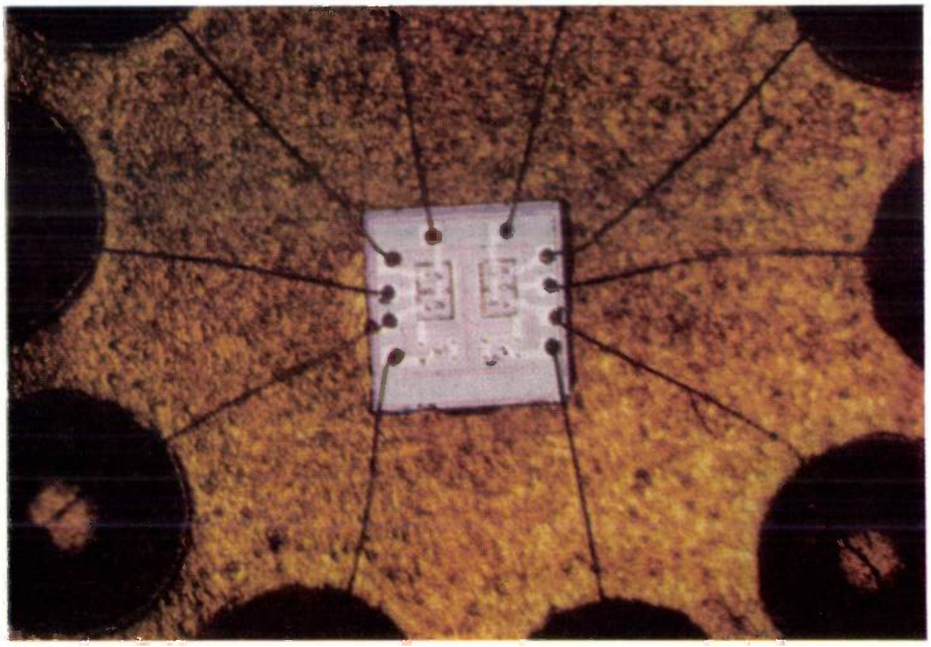


Fig. 3. Microminiaturization by the crystal technique. The little square in the middle (a "chip" 1 × 1 mm^2) contains four transistors and four resistors. The connections leading to normal contacts placed round the chip may also be seen.

The whole field of microminiaturization is undergoing a revolutionary development at present, and as with all revolutions it is very difficult to predict what will be the final outcome.

We will close this section with a few remarks on the *types of circuit* which are used in computer sub-assemblies. The most common are the "nand" and the "nor" circuits. These can be realized in various ways, as may be seen from *fig. 4*, which shows the most usual solutions [5]. Many attempts are being made to find an optimum solution for these and various other circuits. Many different criteria can be used in this connection-speed, sensitivity to interference, the generation of interference, the influence of tolerances and drift of components, energy consumption, attainable "fan-out" (i.e. the number of similar circuits which can be controlled by the circuit in question). If however one considers how the possible circuits would work under the most unfavourable conditions ("worst-case" calculations), then it is found that no single version of the circuits can be optimal in all respects, so that one is forced to make a choice, retaining those properties that are most valuable for the intended application.

The above-mentioned calculations are nowadays normally carried out on an electronic computer, the transistor characteristic normally being approximately represented by a straight line. A few attempts have already been made to take into account the non-linear behaviour of the barrier layers in these circuits, which allows in particular a more accurate quantitative description of transient phenomena. It may be noted that electronic computers are being used more and more for developing better computers and for the improvement of the various computer-manufacturing processes, e.g. for the making of optimum wiring patterns, the making of drawings, parts lists and wiring lists, etc. — thus contributing to a kind of eugenics of these robots.

High-speed main memories

The most important contribution to the improvement of the reliability of electronic computers has doubtless been the introduction, some 10 years ago now, of *magnetic core memories*. The memories used until then contained acoustic delay lines usually with mercury (but sometimes with nickel wire or quartz) as transmission medium, or cathode-ray tubes, usually called Williams tubes after the inventor of this application. The drum memories, which in those days were also sometimes used for main storage in slower machines, will be discussed in the following section. Both of the other types of memories mentioned had serious disadvantages. For example, the temperature dependence of the delay lines was a problem; in the Williams tube, the problem was the poor persistence of the charge pattern on the screen, which necessitated continual regeneration.

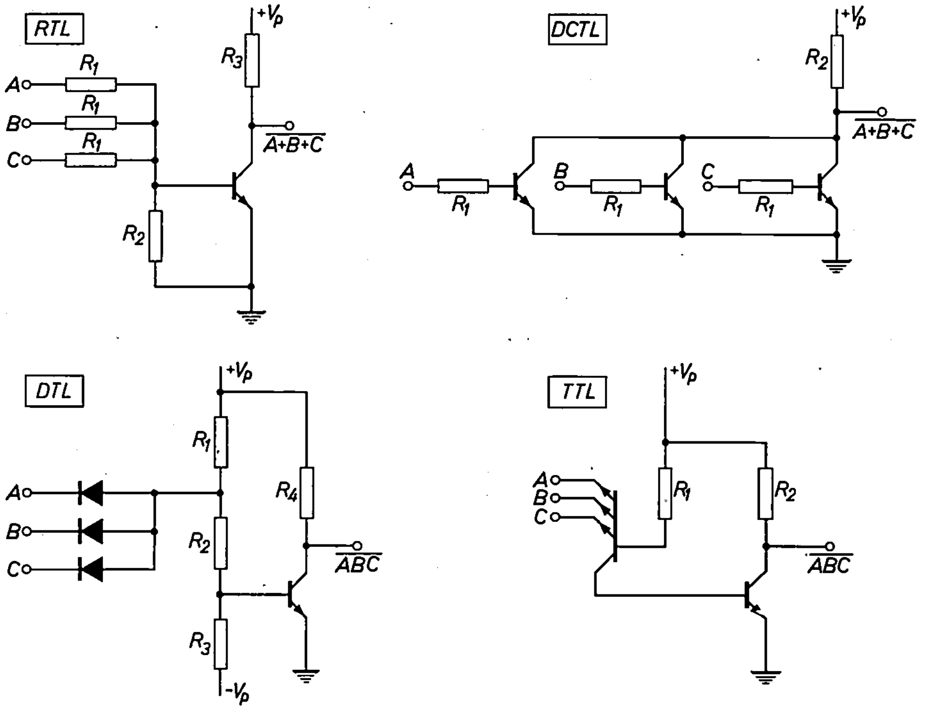


Fig. 4. Various logic circuits using transistors. The top two are "nor" circuits, and the bottom two "nand" circuits. The notation $\overline{A + B + C}$ denotes "not (A or B or C)", i.e. the value "0" (in this case represented by a *low* voltage) will appear at the output as soon as the value "1" (a *high* voltage) is present at at least *one* of the inputs $\overline{ABC}$. The notation ABC denotes "not (A and B and C)", i.e. the value "0" appears at the output only if the value "1" is present at *A and B and C*.

Just when the various technological problems of the Williams tube had been more or less solved, the magnetic core, made possible by the development of ferrites with a square hysteresis loop, came on to the scene, thus inaugurating a period of great and rapid progress in memory techniques.

If we disregard certain types of larger core for special applications, we may say that the first core which found practical application in computer memories had an external diameter of 0.080 inch i.e. about 2 mm. This made possible a cycle time of 10 μs. The cycle time consists of two switching times (for the destructive reading and for the writing back of the information) and a certain time between and after these two, needed for the decay of the transient phenomena in the amplifiers for the read currents and in other parts of the electronic circuit. In order to obtain shorter cycle times, one must try to limit the transient phenomena as well as to reduce the switching time proper. For the latter purpose, we must use cores with a higher coercive force and use a higher field strength for reading and writing. This means, however, since one wants to obtain the higher field strength with roughly the same currents, that one has to use smaller cores. This allows the memory matrices to be made more compact and since in a smaller matrix the voltage pulses occurring during writing are smaller, this also has a favourable effect on the transient phenomena in the amplifiers. For these various reasons, the cycle time could be reduced to 6 μs by use of a core 0.050 inch in diameter (*fig. 5*).

The development workers then concentrated mainly on improving the reliability. Better knowledge of the sintering process of the ferrite used allowed the properties of the memory element to be made much more reproducible. A better choice of the composition of the material gave less temperature dependence and a

[5] R. Foglesong, Integrated circuits design and application, Semicond. Prod. and Solid State Technol. 7, No. 3, 32-34 and 39-42, 1964.

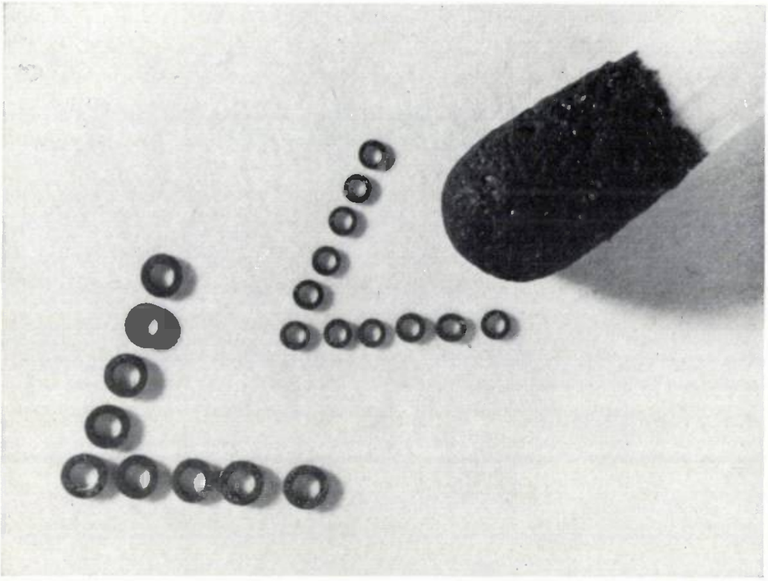


Fig. 5. Ferrite cores 0.050 and 0.030 inches in external diameter. The match-head indicates the scale.

squarer hysteresis loop. Much attention was also paid to improving the methods of threading the cores in the memory matrices.

The steady improvement in the mastery of the technology allowed even smaller cores to be put into use in 1961, the external diameter being reduced to 0.030 inch (fig. 5). As a result of this, and of the appreciable speeding up of the associated electronics which had been realized in the meantime, the cycle time was reduced to about 2 μs.

It goes without saying that the problems of threading the cores became greater: at least three wires had to be threaded through a hole of diameter 0.4 mm. The general opinion was that this represented the practical limit of these memories. Recently, however, it has proved possible not only to make still smaller cores (external diameter 0.020 inch), but also to thread these without too much difficulty (*fig. 6* and *7*). These cores give cycle times of about 1 μs.

The fundamental and practical limitations of this technique are now being examined. There is no doubt that still smaller ferrite cores can be made, but it is not yet known whether it would still be possible to thread them, and if so whether other difficulties will not then arise, e.g. that the resistance of the thin wires might become excessive. It is however to be expected that cycle times of about 0.5 μs will be possible within a few years. *Table III* gives the switching time for the above-mentioned cores, together with the necessary switching current and the year of introduction. One should beware of drawing too stringent conclusions from the differences in switching time given in this table, since the cores on which these values were measured differed appreciably as regards both the thickness and the ferrite used, while the currents used also varied, as may be seen from the table.

Many attempts have been made to find a better alternative to the magnetic core in high-speed memories, e.g. by giving the ferrite *another shape*, by looking for *other materials* with a square hysteresis loop and by use of materials with a basically *faster switching mechanism*.

For example, it has been proposed to make a memory of flat square sheets of ferrite containing a matrix of holes [6]. Part of the wiring is automatically fixed on to the sheets, while the rest can very simply be threaded through the holes. To the best of our knowledge the only large-scale application of these sheets is in the memories for the above-mentioned electronic telephone exchange at the Bell Laboratories.

Table III. Dimensions, switching time and switching current of successively used magnetic memory cores.

External diam. of core (inches)	Year of introduction	Switching time (μs)	Switching current (mA)
0.15	1953	10	400
0.080	1955	1.5	700
0.050	1958	1.0	500
0.030	1962	0.5	650
0.020	1964	0.2	850

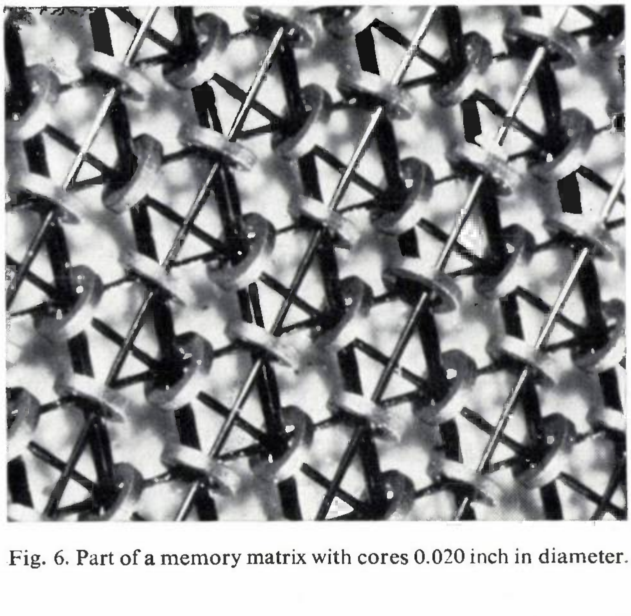


Fig. 6. Part of a memory matrix with cores 0.020 inch in diameter.

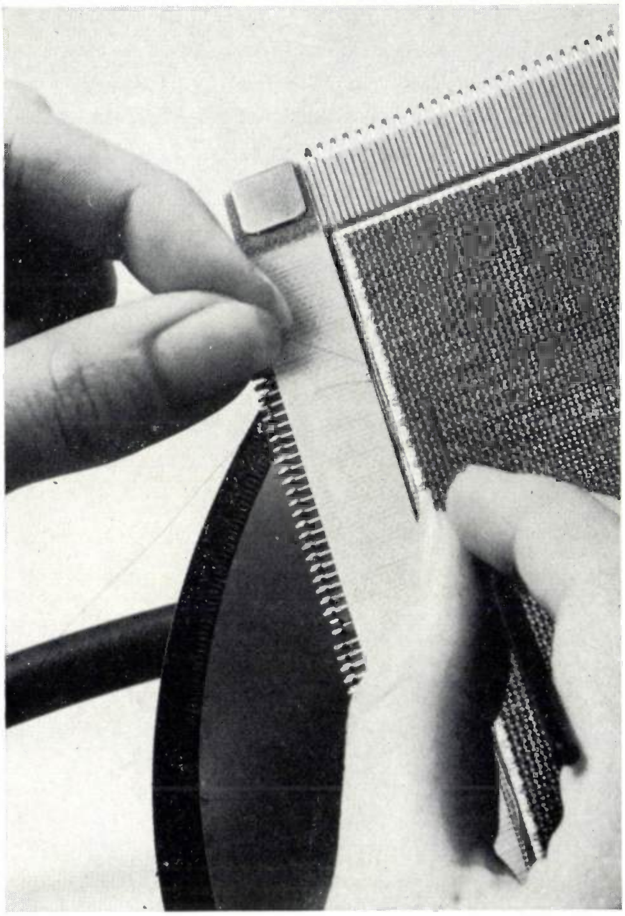


Fig. 7. Threading a memory matrix such as that of fig. 6.

Some dielectrics, e.g. barium titanate and GASH [7], have a roughly square hysteresis loop, but one property makes these substances inferior to the magnetic materials, viz, the fact that the memory elements based on them have two poles. With the magnetic materials one can make elements with separate input and output wires, which gives many more possibilities. Another disadvantage in e.g. barium titanate is the memory loss.

The most important memory element with a fast switching mechanism is the thin magnetic metal film (*fig. 8*), where the magnetization is not reversed by movement of Bloch walls but by the much faster rotation of the direction of magnetization in a Weiss region. The switching times which can be obtained in this way are of the order of nanoseconds, but no memory has yet been made which takes full advantage of this speed. The only thin-film memories on the market at present have cycle times of from 0.3 to 1 μs, while in a laboratory study a prototype of a memory with a cycle time of 0.1 μs has been realized. Thin films also have the disadvantage of a relatively very low output voltage, viz, about 1 mV, compared with tens of mV in the magnetic cores. The ferrite core is thus still leading for the moment, and it looks as if the thin films will rather lose than gain ground.

This does not mean that the thin film has already lost the race. It is conceivable that good technological mastery of the manufacturing process of the thin-film memories (including their control and read circuits) could eventually make these memories cheaper than corresponding magnetic-core memories. In this case, the price rather than the speed would be finally decisive. So far, however, this is not the case. And whether the thin film will still stand a chance in a few years' time will strongly depend on the progress made with the *cryogenic memory*, while the newest inventions in the field of ferrite memories, the *microferrites* and the *laminated ferrites* [8] also appear very promising.

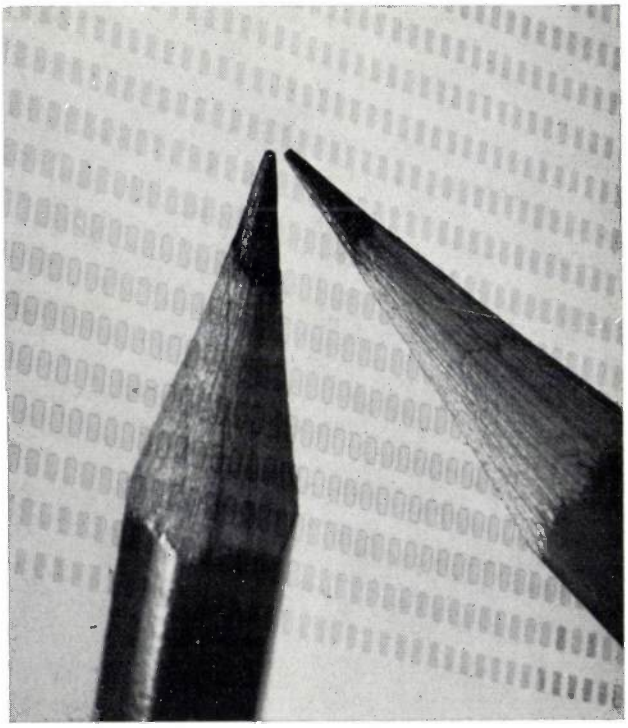


Fig. 8. A memory in which each element is formed by a very thin magnetic metal film on an aluminium carrier. The pencil point (which is reflected in the aluminium) indicates the scale.
The wiring for this memory is applied to various Mylar foils which are laid on the pattern of elements.

Cryogenic memories make use of substances which become superconducting at low temperatures. With

[6] J. A. Rajchman, Proc. IRE **45**, 325, 1957.
R. H. Meinken, Conf. Magnetism and magnetic materials, Boston 1956, p. 674.
J. A. Rajchman, Computer memories: a survey of the state-of-the-art, Proc. IRE **49**, 104-127, 1961.

[7] Guanidine Aluminium Sulphate Hexahydrate. — Guanidine is a substance which was first discovered in guano.

[8] R. Shahbender, C. Wentworth, K. Li, S. Hotchkiss and J. Rajchman, Laminated ferrite memory, AFIPS Conf. Proc. **24** (1963 Fall Joint Computer Conference), pp. 77-90.

thin-film techniques it is possible to make very small memory matrices on this principle (e.g. 128×128 bits on a surface of 5×5 cm^2). The circuits for the selection of the desired word from the memory can be made by similar means and form an integral whole with the matrix. Memory capacities of the order of 10^9 bits would seem to be possible in this way. However, the very low resistances limit the switching speed of larger units, which is determined by L/r relaxation times. The switching time will be of the order of 1-10 μs. One advantage of these cryogenic elements is that, thanks to the infinite ratio between resistances in the normal and superconducting state, no parasitic voltages are produced by the coincident currents i [9] used for switching, so that one can in principle make very large memory matrices. There are still plenty of technological problems, however. The problem of finding the most economical cooling method for this application must also be considered [10].

It is hoped that microferrites and laminated ferrites may offer possibilities for memories of the order of 10^7 bits. Cycle times of from 0.1 to 1 μs have been measured on a number of prototypes of this kind. The output voltages are of the same order of magnitude as with the metallic magnetic films. The read currents needed can vary from 50 to 200 mA.

All these figures show that memories, like the logic circuits, are tending more and more to a microminiature form, so that we may in the long run expect much smaller computers, where the logic and the memory are better matched and thus have a more integrated character than is the case at present. The increasingly noticeable attempts to put selection circuits in the memories themselves also work in this direction.

Apart from the high-speed main memories discussed in this section, there is also a need for a fast, inexpensive *permanent or semi-permanent memory*, from which the information can be non-destructively read very rapidly, while changing the memory contents is relatively slow (e.g. by photographic processes). The demand for memories of this kind will increase more and more in connection with the steady increase in the amount of permanent information which one wishes to feed to an electronic computer in the form of translation programmes, sub-routines etc. Another reason for the demand for nondestructively-read memories is the risk that in the normal procedure. i.e. destructive reading followed by writing back of the information in question, a disturbing pulse may change the contents of the memory. This is catastrophic for memories for use in space flight, but even in telephone exchanges it is highly undesirable, since it could result in a subscriber losing the correct number as a result of a permanent change in this number during writing back in a memory.

No completely satisfactory solution for memories of this kind has yet been found. Various methods have been worked out, but they all leave room for improvement. For example, in the telephone exchange which has already been mentioned several times, the Bell Laboratories used the twistor as a semi-permanent memory element for storing permanent subscriber data and programmes for the connecting of one subscriber with another. Growing interest also exists at present in the "Biax" element, one of the several multi-hole ferrite elements which have been developed (*fig. 9*), in connection with its use in a nondestructive-read memory which can be read quickly (about 100 ns) and can still be written fairly quickly (of the order of a few μs).

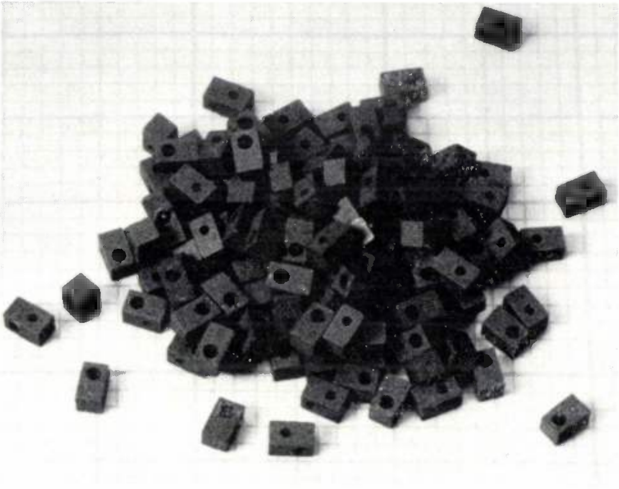


Fig. 9. A pile of "Biax" elements for a semi-permanent memory. Each element contains two holes at right angles to one another. The length of the element is about 2 mm.

For the sake of completeness, we may briefly mention here the reappearance of the delay line, in two forms, viz, the polygonal form in glass with a very low temperature coefficient and the magnetostrictive delay line [11]. Owing to the great increase in the speed of the logic circuits, these delay lines can, when used in not too large serial machines, provide the key to reasonably fast, and at the same time cheap, solutions.

Slow memories: drums, discs, magnetic tape

The need for larger memory capacities than can economically be provided by the fast main memories was already felt in the first electronic computers. The price per bit is here the decisive factor. Here too, magnetic recording still reigns supreme as regards simplicity and reliability, and this state of affairs may be expected to continue for years, although laboratory investigations are now being carried out on a number of interesting discoveries, e.g. thermoplastic recording. In fact, this method is probably more suitable for use in a semi-per-

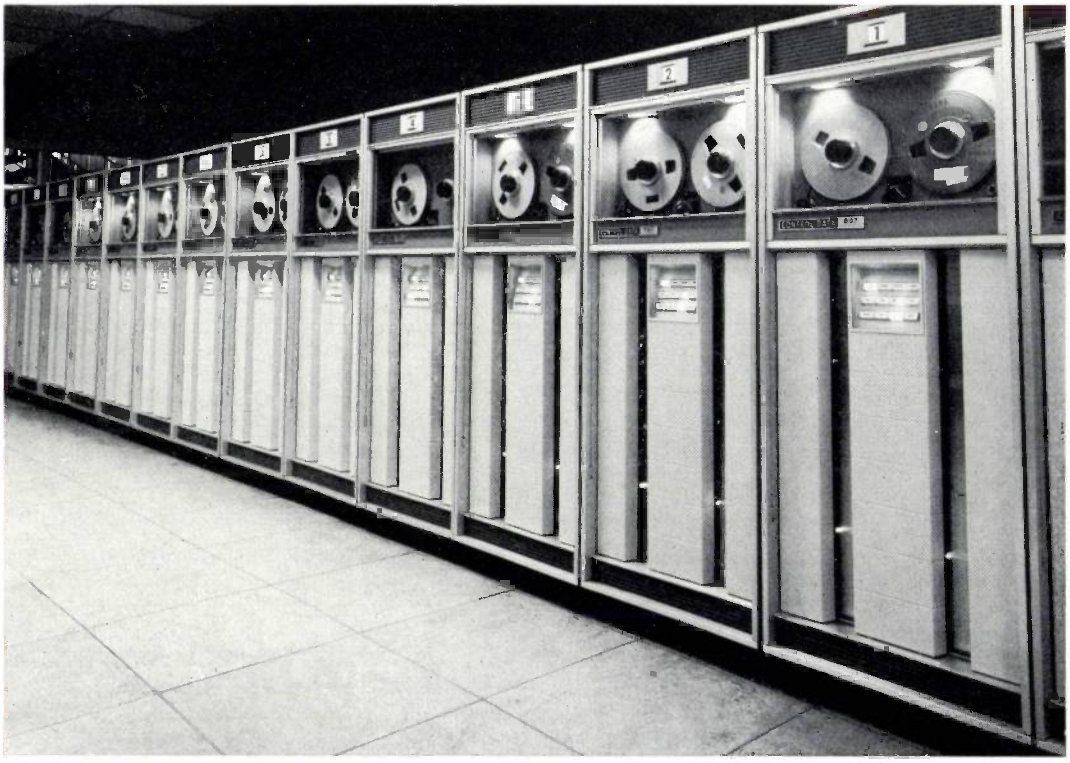


Fig. 10. Magnetic-tape units, used with the CDC 3600/3200 computer in Philips Computer Centre, Eindhoven.

manent memory, since the writing rate is on the low side.

Table IV gives some data about a number of representative memory types which will be briefly discussed in this section, to show how the present state of affairs compares with that of about 10 years ago.

Table IV. Typical data of magnetic tape, drums and discs in 1954 and 1964.

	1954	1964
ape:		
Reading and writing speed (bits/s per track)	12 000	170 000
Writing density (bits/mm)	4	60
Start-stop time (ms)	10	2-3
Drums:		
Mean access time (ms)	10-20	10-20
Writing density (bits/mm)	2-5	26-40
Capacity (bits)	10^5-10^6	10^7-10^8
Discs:		
Mean acces time (ms)	600	20-200
Writing density (bits/mm)	2-4	20-40
Capacity (bits)	10^8	10^8-10^9
Magnetic cards:		
Mean access time (ms)	—	100-600
Writing density (bits/mm)	—	ca. 10
Capacity (bits)	—	10^7-10^{10}

Magnetic tape is still much the cheapest information carrier, with a price of the order of 0.0003 dollar per bit. The writing rate for tape has increased by a factor of nearly 30 in these 10 years. The rate of 170 000 bits/s given in the table refers to the IBM's latest development, "hypertape".

Meanwhile, the magnetic-tape field has been undergoing a tacit standardisation of recent years: a number of firms have been marketing tape units which can directly read the tapes written by IBM machines, and vice versa. In this connection, the tape speed has been standardized at 75 and 150 inch/s and the writing density at 200, 556 and 800 bits/inch.

A great problem of a mechanical nature is the quick starting and stopping which is desired in order to leave as little unwritten tape as possible between blocks of information. Normally one needs a time of 7 ms for starting and stopping; modern techniques — which still however make the tape units rather expensive — can achieve times of 2 to 3 ms (*fig. 10*).

[9] For the concept of coincident current in the writing and reading processes, see e.g. H. J. Heijn and N. C. de Troye, Philips tech. Rev. **20**, 193-207, 1958/59, in particular p. 199.

[10] See also G. Prast, A gas refrigerating machine for temperatures down to 20 °K and lower, Philips tech. Rev. **26**, 1-11, 1965 (No. 1).

[11] For ultrasonic delay lines see e.g. C. F. Brockelsby and J. S. Palfreeman, Philips tech. Rev. **25**, 234, 1963/64 (No. 9).

The fact that magnetic tape entails a very high access time (of the order of seconds to minutes) owing to the fact that the information is written serially all along the tape has stimulated the demand for "*random-access*" *memories*, i.e. memories which can produce any desired piece of information without appreciable delay (appreciable, that is, for the application in question). The development here has been mainly in two directions; *drum memories* and *disc memories*.

We have already mentioned in passing the fact that not so long ago (about five years) some relatively small, slow computers still used drums as their main memories. Now this would be quite out of the question: the magnetic cores have advanced so far the question now arises whether they might not even provide an economic solution for the high-capacity memories for which drums are now used. For the moment, however, the drum still has a place as a big secondary memory with a capacity of the order of 10^7 to 10^8 bits and an average access time of about 10 ms (*fig. 11*).

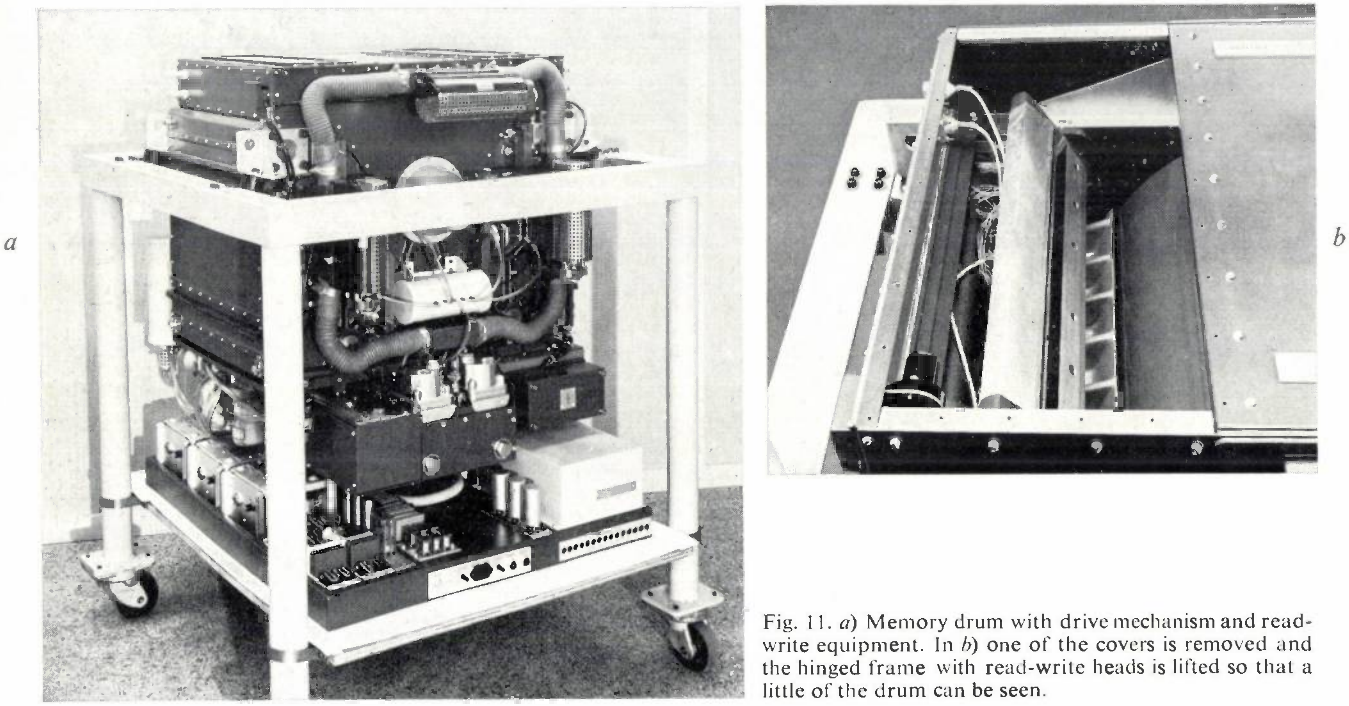


Fig. 11. *a*) Memory drum with drive mechanism and read-write equipment. In *b*) one of the covers is removed and the hinged frame with read-write heads is lifted so that a little of the drum can be seen.

Of recent years it has proved possible to increase the writing density on the drum considerably by a new way of mounting the read-write heads. Things are so arranged that the heads float stably on a thin cushion of air which rotates with the drum. This has allowed the distance between the heads and the surface of the drum to be reduced to a few microns. A further improvement in the drums is the saving in electronic reading and writing equipment obtained by providing means for shifting the heads mechanically over several tracks. This idea is in fact more than 10 years old, but it was difficult to realize properly without the floating-head technique.

A curious hybrid of the tape and the drum is the "carrousel tape" brought out by Facit. In order to reduce the access time, the tape is divided into 64 small pieces, each about 10 m long, each of which has its own little reel. The carrousel with the 64 reels must first turn until the desired reel is in front of the reading device, and then the tape of this reel must be fed past the reading device. The average access time is here 1 to 2 s and is thus a factor of 100 less than in normal tape units.

Disc memories, which came on to the scene nearly 10 years ago (Ramac 355), have undergone improvements similar to those of the drums, viz, an increase in the writing density and in the capacity. A special development in this field is the disc memory for the IBM 1440 computer, where a complete packet of discs can be exchanged. This thus gives a theoretically unlimited memory capacity, just as with magnetic tape, but with a much better average access time and the flexibility is much increased.

It may also be mentioned that some firms (IBM, NCR, RCA) have tried to increase the flexibility and decrease the access time in a big memory by slicing up the magnetic medium to give *magnetic cards*.

An important question which keeps on cropping up is whether there will still be a place for magnetic tape memories after the introduction of these big random-access memories. It has been suggested that one will probably still keep the magnetic-tape memory as a sort of spare memory, into which the disc memories can be "emptied" before being used for new purposes. Tape can also be used in order to provide a new starting point in case an error should occur in the machine during calculation which would make the information contained in the disc memory unreliable. With this conception, one need no longer make high mechanical demands on the tape units; e.g. short start and stop times would no longer be required. This allows the whole unit to be made much cheaper.

The possibility of having at one's disposal memory units in blocks of about 10^9 bits with access times of 0.1 to 0.2 s has given many people the idea of making a big computer with a very big memory, which by analogy with a telephone exchange could be used as a sort of *information exchange*, connected via a communication network (*data transmission network*) with a number of subscribers. The subscribers could be provided with equipment varying from simple keyboards to complete satellite computers, with the aid of which they could obtain access to and make use of all the computing and memory facilities of the exchange.

It is a fact that in America it is expected that the telephone network will undergo considerable expansion for the purposes of data transmission. It therefore seems to us that the information exchange is a real proposition. Computer techniques will then however have to make more use of redundancy than has been done in the present designs, in order to give the system as a whole the reliability which is so necessary for this plan. The operation of such an exchange will also require an enormous amount of initial programming work.

Peripheral equipment

In order to give an impression of the progress made in the field of peripheral equipment over the past ten years, *Table V* gives various important data about the most used input and output units. Inspection of these data will show that a much greater advance has been made in the input speed (by a factor of 20) than in the output speed (a factor 6). The reason for this is that the output units are subject to mechanical limitations, which can be eliminated in the input units by the use of optical systems. Here too there is a mechanical limit in the last resort, but this is related more to the handling of the information carrier (*punched cards* or *punched tape*), than to the reading mechanism. As is known, the tensile strength of paper tape limits the reading speed to about 4000 characters/s. There are a number of tape readers on the market for 1000 characters/s, so that not much more progress is possible here (*fig. 12*). Speeds of 2000 characters/s have been reached in laboratory set-ups. The punched card is a little stronger than the paper tape, but here too the present-day results lie quite close to the physical limit.

Table V. Speeds of various types of input and output units in 1954 and 1964.

	1954	1964
Punched-tape reader	100	2000 characters/s
Tape puncher	50	300 characters/s
Punched card reader	2	50 cards/s
Card puncher	1	5 cards/s
Line printer	2.5	15-20 lines/s

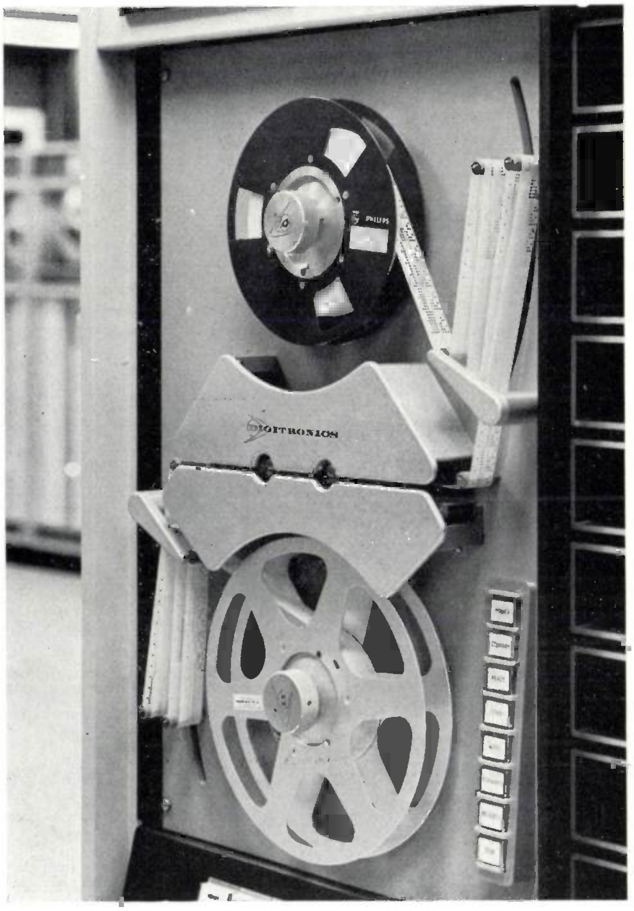


Fig. 12. Commercially available punched tape reader with a reading speed of about 1,000 characters per second, in use in Philips' Computer Centre, Eindhoven.

If one wishes to achieve greater speeds with the output units, one will have to make use of completely different principles. Investigations are being carried out on e.g. a method in which the punching is not done with a mechanical punch, but by means of the pressure wave caused by the passage of a spark [12]. A similar method should be able to form the basis of a high-speed printer. Another possibility is to replace the punching by

[12] See G. Haas, Problems and trends in the development of peripheral equipment for computers, to be published in Philips tech. Rev. **26**, 1965 (No. 4/5/6).

the application of black marks, but this has not proved very reliable so far, the margin between light and dark of the reflected light being rather small. The xerographic methods, which one would like to develop for output units, are also unreliable, but in another way: it is impossible to guarantee the production of a number of identical copies. An identical error in all copies is less objectionable than a good first print followed by others which might contain an error, or vice versa.

It may be asked whether it is really worthwhile spending an effort on the investigation of further improvements to this kind of input and output equipment, now that methods for the *automatic reading* of typed or printed matter are on the way.

Banks are already standardizing a method for the reading of stylized magnetic characters (type E13B); see *fig. 13*. This standardization began in the United States, and has been followed by Canada, Australia and Great Britain. Other suggestions have been made in Western Europe (type CMC7 in fig. 13); it is not yet clear how things will develop in Europe in this respect [13].

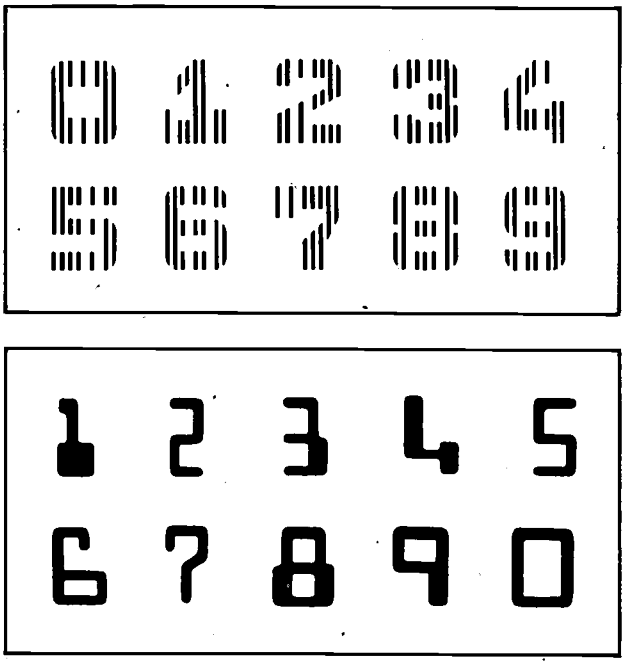


Fig. 13. Stylized magnetic digits which can be visually recognized and can also be read automatically by a scanning process. Above type CMC 7, below type E 13 B [13].

A great deal of development work has already been done on optical character readers, which will ultimately have a much wider range of applications. One obstacle to more general application at present is a lack of standardization of the types of letters to be read [14]. However, the IBM has already included an optical reader as direct input for the 1401 computer. This has a reading speed of 480 characters per second, and can deal with up to 400 documents per minute. This may possibly be taken as a standard to be followed for the moment by other computer manufacturers.

Will these optical readers perhaps change the entire character of the peripheral equipment — apart from the line printers which will still be needed? Or will there still be a place, albeit more modest, for the good old punched card as a very durable and flexible memory medium? We will not venture to make any predictions about this, but we do expect that within a few years a clear answer will be possible to these and many other questions which have remained open in this survey: the rapid development of the computer, which we referred to at the start of this article, is still progressing at full speed.

Summary. This article is based on a lecture given by one of the authors (v.d.W.) to the Dutch Computer Society. The developments discussed, which are to a certain extent still in full flight, concern 1) the improvement of the reliability, coupled with the realization of a more compact construction and higher switching speeds. As essential steps in this process are described the miniature technique and microminiaturization in "integrated circuits", in which transistors and diodes are made in one unit, together with resistors, capacitors and connections; 2) the realization of larger and larger memories, preserving a reasonable access time. The gradual reduction in the size and increase in the speed of magnetic memory cores and the successive new techniques (thin-film, etc.) are discussed in some detail; 3) the speeding up of the input and output of data. These processes are subject to mechanical limitations, and the progress here has so far been the smallest.

[13] W. J. Bijleveld, Automatic reading of digits, published by Stichting Studiecentrum voor Administratieve Automatisering, Amsterdam, March 1963.

[14] For proposals for the standardization of the reading of digits made by the Dutch Postal Order Service see: W. J. Bijleveld and A. J. van der Toorn, Methoden voor het met de hand invullen van automatisch te lezen getallen (Methods for writing by hand numbers to be read automatically; in Dutch), Ingenieur **76**, A 693-702, 1964 (No. 46).

Nitrogen in silicon iron

J. D. Fast

546.17:621.318.13

Professor Fast has been a frequent contributor to our journal. The article below is a chapter from part I of his recently published book "Interaction of metals and gases" [*]. *This chapter deals with methods of bringing about in silicon iron a special crystal orientation which gives the material favourable magnetic properties. New insight was obtained from experiments carried out by the author — about ten years ago — in which he added for the above mentioned purpose small quantities of a second phase to the metal. His experiments gave the impetus to extensive investigations in various countries, which have led among other things to better control of the production of grain-oriented silicon steel, used for making transformer cores.*

Introduction

Nitrogen can play an important part in silicon steel that is used as the "iron" core of transformers, electric motors and electric generators. The steel which is used in transformers generally contains about 3 % silicon.

In the earliest transformers the core was made from unalloyed steel which was soft-annealed to make it as far as possible free of internal stresses. The coercivity and hysteresis losses in this material were relatively large as a result of the many inclusions it contained, especially inclusions of iron carbide. Also the eddy current losses were relatively high due to the small electrical resistivity. It was especially unfortunate that the coercivity and hysteresis losses spontaneously increased with the course of time. This "magnetic ageing" was caused by the slow precipitation of nitrogen in the form of iron nitride [1].

The silicon steel used in transformers suffers smaller eddy current losses than unalloyed steel due to its greater resistivity. Also the hysteresis losses are smaller because silicon encourages the formation of graphite which is magnetically less harmful than iron carbide, since for the same number of carbon atoms the total volume of inclusions is much smaller. It is also of importance that silicon steel exhibits no ageing phenomena because the nitrogen which is present as an impurity occurs, after a suitable annealing treatment, in the form of a very stable silicon nitride [2]. The composition of this nitride is Si_3N_4 [3]. However, this precipitate too must be considered undesirable since, like all other precipitates, it has an unfavourable effect on the coercivity and the hysteresis losses. In the following, however, we shall see that by the deliberate addition of nitrogen to silicon steel one can profit from the presence of this element to obtain a magnetically favourable crystal orientation.

The role of Si_3N_4 in making grain-oriented silicon steel sheet

. Until comparatively recently virtually all the silicon iron sheet for transformers was obtained by hot-rolling The directions of easy magnetization, <100>, of the separate crystals in the sheet are then almost randomly distributed over the various directions in space (*fig. 1*), so that the hysteresis losses are relatively large. In principle it would be most desirable to make the cores

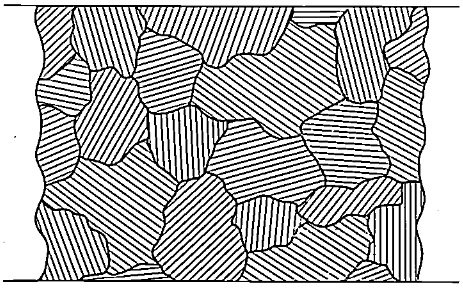


Fig. 1. The figure shows schematically that in hot-rolled silicon steel sheet the crystals show no preferential orientation with respect to the plane and direction of rolling.

Prof. Dr. J. D. Fast is a research worker at Philips Research Laboratories, Eindhoven, and a Professor Extraordinary of Physical Chemistry at the Technical University, Eindhoven. On 9th January 1965 Prof. Fast received from the Technical University of Delft the degree of Doctor honoris causa "for his outstanding services in the field of scientific and technical research on metals".

[*] J. D. Fast, Interaction of metals and gases I, published by Philips Technical Library, Eindhoven, and Academic Press, New York, 1965.

[1] J. D. Fast, Philips tech. Rev. **13**, 165, 1951/52; J. D. Fast and L. J. Dijkstra, Philips tech. Rev. **13**, 172, 1951/52.

[2] J. D. Fast, Philips tech. Rev. **16**, 341, 1954/55.

[3] W. C. Leslie, K. G. Carroll and R. M. Fisher, Trans. AIME **194**, 204, 1952; H. A. Sloman, J. Iron Steel Inst. **182**, 307, 1956.

from large laminar single crystals (bounded by cube faces) in such a way that the magnetic flux always follows a direction of easy magnetization (*fig. 2*).

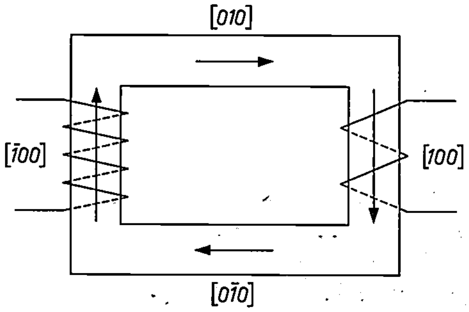


Fig. 2. Symbolic representation of a transformer, the core of which is built up of single crystal sheets in such a way that the magnetic flux can everywhere follow a direction of easy magnetization, a <100> direction.

Technically it is impossible to make single crystal sheet in large quantities. But it has been found possible to make on a large scale (in quantities of thousands of tons each month) polycrystalline silicon iron sheet, in which all the crystals have nearly the same orientation (*fig. 3*). This orientation is such that the crystals lie with a (110) plane approximately parallel to the surface of the sheet and with a [001] direction, a direction of easy magnetization, approximately parallel to the direction of rolling (*fig. 4*). This texture is often referred to as Goss texture after its inventor, Goss, but also as (110)[001] or cube-on-edge texture [4]. The latter name is illustrated by *fig. 5*, in which the orientation under discussion is demonstrated with the help of a number of cubes, which symbolize the unit cells. The Goss texture is obtained by cold-rolling silicon iron from a certain thickness and by subjecting it to certain heat-treatments.

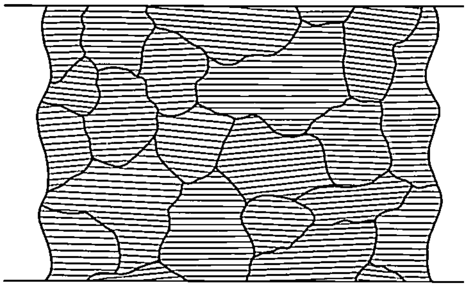


Fig. 3. Schematic representation of the crystals of cold-rolled 3% silicon iron sheet in which all the crystals have about the same orientation, viz. the orientation (110)[001] (cf. fig. 4).

For many years the way in which the crystal orientation in the Goss sheet is brought about was not understood and the production of the sheet did not always give the desired results. Our own experiments showed that it is impossible to obtain the Goss texture in silicon iron sheet made from pure iron and transistor-quality silicon. From this we concluded that the presence of impurities in the material is of essential importance [5]. In further experiments we added measured quantities, in each case of one element, to pure silicon iron alloys.

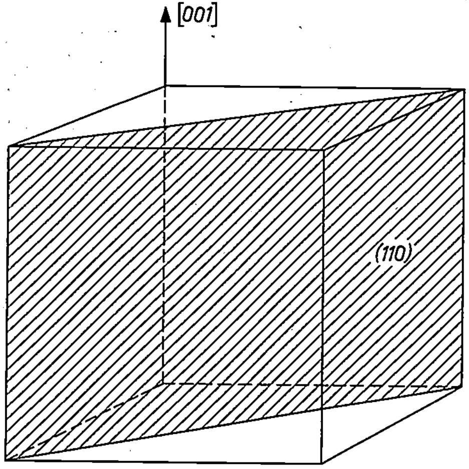


Fig. 4. In silicon iron sheet having Goss texture the crystals are oriented in such a way that they lie with a (110) plane approximately parallel to the rolling plane and with a [001] axis (cube axis) approximately parallel to the direction of rolling (cf. fig. 5).

It was found that the desired texture is readily obtained by introducing nitrogen (in quantities of a few hundredths of a percent) and heat-treating the metal before cold-rolling in such a way that it contains a finely divided precipitate of Si_3N_4, which is mainly present at the grain boundaries [6]. *Fig. 6* shows an electron microscopic photograph of 3% silicon iron, in which a precipitate of this type is present.

After cold-rolling and after primary recrystallization at 600° to 800 °C, both the pure silicon iron sheet and

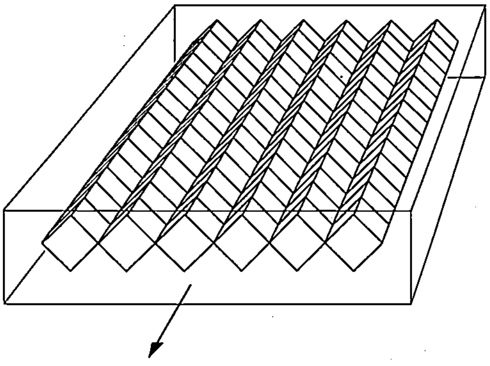


Fig. 5. Schematic representation of the crystal orientation in magnetic steel with Goss texture (cube-on-edge texture). The arrow indicates the direction of rolling.

the sheet containing Si_3N_4 contain only very few crystals with the orientation (110)[001]. In the pure alloy the primary recrystallization is followed at high temperature (e.g. 900 °C) by normal grain growth which exhibits no preference for a particular orientation. In silicon iron containing nitrogen the normal grain growth is inhibited by the Si_3N_4 precipitate. If the metal is heated in the appropriate temperature range in pure hydrogen these inhibiting inclusions slowly coagulate and go into solution, and at a certain point the few favourably oriented crystals begin to grow but not, as yet, the others. In order words secondary recrystallization (exaggerated grain growth) occurs, by means of which a few primary crystals with orientation (110)[001] grow at the cost of the other crystals to many times the sheet thickness. The driving force for growth of of these grains is the low gas-metal interfacial energy of the (110) surfaces in an atmosphere of pure hydrogen [7]. This surface energy is less than that of any (*hkl*) plane different from (110).

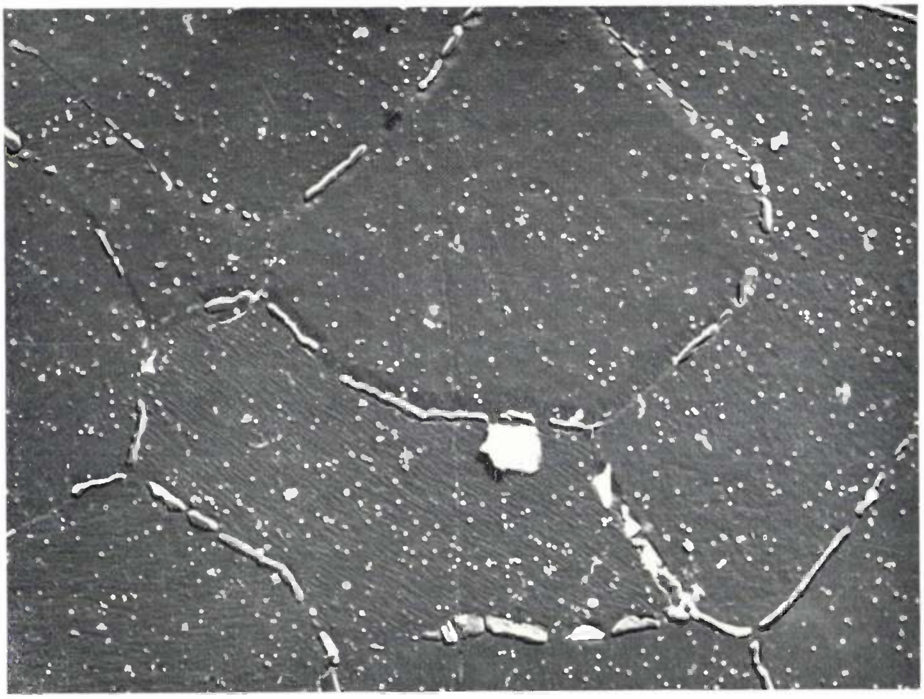


Fig. 6. Electron micrograph of 3% silicon iron in which a precipitate of Si_3N_4 is present along the grain boundaries. The micrograph corresponds to an area of 17 × 19 microns.

The selective grain growth under discussion occurs in a particular temperature range. If the nitrogen-bearing silicon iron is heated immediately before or after the primary recrystallization at too high a temperature (e.g. 1250 °C), then the active inclusions dissolve very rapidly, so that one obtains mainly normal grain growth, which results in a poor texture. Therefore one must first heat at a lower temperature (900 °-1000 °C). As soon as the (110)[001] crystals have a sufficient start in size on the other crystals, the temperature can be raised to accelerate further growth.

The Si_3N_4 particles, which are so useful for producing the Goss texture, are unfavourable for the final magnetic characteristics of the material. They must therefore be removed from the silicon iron after they have accomplished their grain-growth function. This takes place automatically in the final heat-treatment in an atmosphere of pure hydrogen, since virtually all the nitrogen then leaves the metal [5][8].

In the commercial 3% Si-Fe alloys, MnS is the most important impurity inhibiting normal grain growth after primary recrystallization [9]. Various other inclusions can also perform this task [10]. Here again it is of primary importance that they be present in the metal in

[4] N. P. Goss, Trans. Amer. Soc. Met. **23**, 511, 1935; R. M. Bozorth, Trans. Amer. Soc. Met. **23**, 1107, 1935; C. G. Dunn, Cold working of metals, Amer. Soc. for Metals, Cleveland 1949, pp. 113-120.

[5] J. D. Fast, Philips Res. Repts. **11**, 490, 1956.

[6] See also: J. D. Fast and J. J. de Jong, J. Phys. Radium **20**, 371, 1959.

[7] J. E. May and D. Turnbull, Trans. AIME **212**, 769, 1958; J. L. Walter and C. G. Dunn, Trans. AIME **215**, 465, 1959 and **218**, 1033, 1960.

[8] J. D. Fast and H. A. C. M. Bruning, Z. Elektrochemie **63**, 765, 1959.

[9] See the first article mentioned in note [7].

[10] H. C. Fiedler, Trans. AIME **221**, 1201, 1961; M. J. Markuszewicz, J. Iron Steel Inst. **200**, 223, 1962.

the desired size and distribution. For inclusions of MnS this has been shown most convincingly [11] by Fiedler. He experimented with a 3.3% silicon iron alloy which contained a little less than 0.1% MnS. This compound could be completely dissolved by heating the metal at 1325 °C. The most effective degree of dispersion could then be obtained in two ways: (*a*) by correct choice of the cooling rate, (*b*) by drastic quenching followed by precipitation heating at a lower temperature (1000 °C).

An advantage of Si_3N_4 over other grain-growth inhibitors is that after the final heat-treatment no impurities remain behind in the metal, so that the magnetic properties are particularly good.

The characteristics of silicon iron sheet with Goss texture deviate in some respects very little from those of single crystals. A disadvantage, however, is that the unfavourable [110] directions of the crystals lie in a direction perpendicular to the direction of rolling (cf. fig. 4). One can therefore only take full advantage of this material if the magnetic flux is everywhere parallel to the direction of rolling, i.e. if it is used in the form of ring cores wound from sheet and not, for example, in the form of E sheets.

For many applications it would be very desirable to have available silicon iron sheet with cube texture [12], i.e. sheet in which the crystals are so oriented that not only the direction of rolling, but also the direction perpendicular to it is a direction of easy magnetization ((100)[001] texture, *fig. 7*). In Germany and the U.S.A. research workers have already succeeded in making this material with (100)[001] texture on a small scale. In the production, the interaction between the metal surface and the surrounding gas atmosphere plays a very important part. If the oxygen activity of the gas exceeds a particular value, then it is no longer the (110) planes but the (100) planes which have a smaller surface energy than all other (*hkl*) planes, in other words, the driving force for growth of (100) grains is then greater than that of (110) grains.

A convincing demonstration of the above is given by experiments of Walter and Dunn [13] on the migration of (100)/(110) boundaries, i.e. boundaries between two grains, one of which has a (100) plane and the other a (110) plane parallel to the surface of the 3% silicon iron sheet. At 1200 °C the (100)/(110) boundaries advance into (100) grains in a good vacuum, then reverse their direction and migrate into (110) grains in an atmosphere of impure argon. The direction of migration reverses once again with (110) grains growing at the expense of (100) grains in a second vacuum anneal. These results are explained by the authors in terms of a change in concentration of oxygen atoms at the gas-metal interface during the anneals. The addition of oxygen atoms to the surface during the anneals in impure argon results in a decrease of the specific surface energy of the (100) oriented grains to a value lower than that of the (110) oriented grains. In a good vacuum or in pure hydrogen, however, the oxygen concentration at the surface is lowered to the point where the surface energy of the (110) grains has the lowest value.

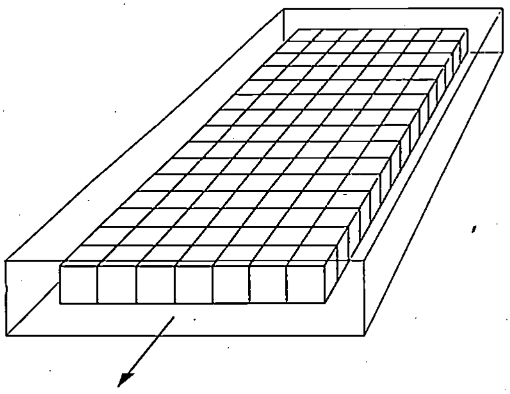


Fig. 7. Schematic representation of the cube texture (cf. fig. 5). The arrow indicates the direction of rolling.

The development of the cube texture in 3% silicon iron is, however, much more difficult and complicated than would be supposed from the foregoing. Control of the gas atmosphere in the final heat treatment is a necessary, but not sufficient condition for success. Up to now it has not been found possible to produce silicon iron sheet with cube texture economically in large quantities. The main difficulty seems to be getting the alloy into such a condition that, after primary recrystallization of the sheet, there is a sufficient number of crystals present with the required (100)[001] orientation. According to patents of the General Electric Company (U.S.A.) this aim can be achieved by starting with ingots having favourably oriented columnar crystals obtained by controlled directional solidification.

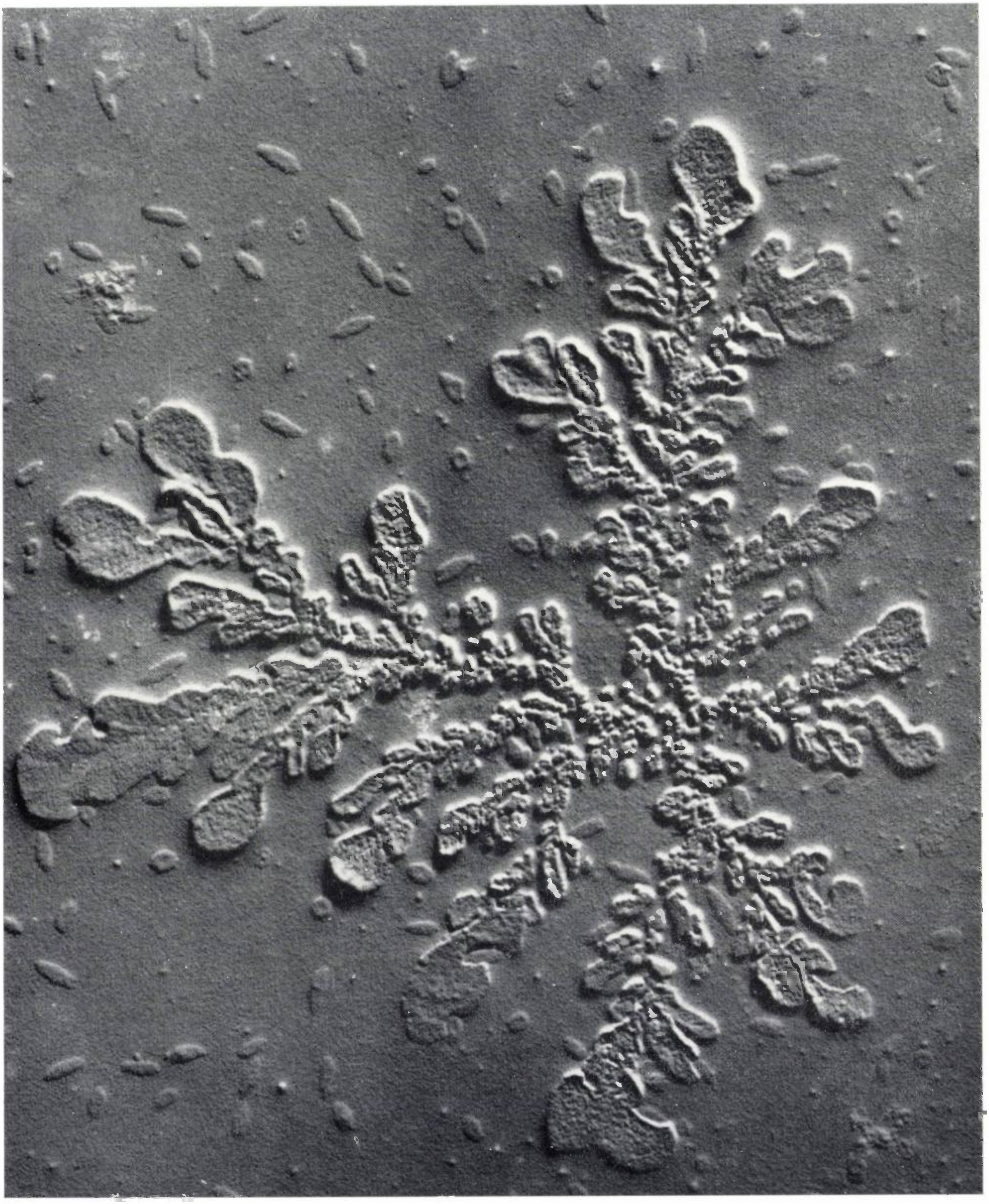
Summary. The article is a chapter from part I of the author's recently published book "Interaction of metals and gases". The chapter deals with the influence of nitrogen on crystal orientation and the associated magnetic properties of cold-rolled and recrystallized silicon iron, which is generally used for transformer cores. Given the right conditions, Si_3N_4 precipitates as a finely divided second phase which promotes the growth of crystals with a (110)[001] orientation ("cube-on-edge texture"). The MnS present in commercial types of silicon steel fulfils a function similar to that of Si_3N_4. Mention is also made of experiments aimed at producing silicon iron with a (100)[001] orientation ("cube texture"), which would be even better suited for the same purpose.

[11] H. C. Fiedler, Trans. AIME **230**, 95, 1964 (No. 1).

[12] F. Assmus, R. Boll, D. Ganz and F. Pfeifer, Z. Metallk. **48**, 341, 1957; F. Assmus, K. Detert and G. Ibe, Z. Metallk. **48**, 344, 1957; J. L. Walter, W. R. Hibbard, H. C. Fiedler, H. E. Grenoble, R. H. Pry and P. G. Frischmann, J. appl. Phys. **29**, 363, 1958; G. Wiener, P. A. Albert, R. H. Trapp and M. F. Littmann, J. appl. Phys. **29**, 366, 1958.

[13] J. L. Walter and C. G. Dunn, Acta metallurgica **8**, 497, 1960.

Glass fracture surface



Fracture surface of a fragment of glass from a television picture tube, taken with the electron microscope a few hours after fracture. Magnification approx. 20 000 ×. Within about an hour of fracture, traces of chemical attack are visible: they are probably local swellings caused by the action of water (layers of silica gel). After a few hours a kind of frostwork tracery as seen in the photograph often appears.

Recent scientific publications by the staff of the Philips laboratories and factories

Reprints of those papers not marked with an asterisk * can be obtained free of charge from the Philips Research Laboratories, Eindhoven, Netherlands, where a limited number of reprints are available for distribution. Requests need only quote the reprint number given in italics at the end of each entry.

P. Beekenkamp and **J. M. Stevels:** The structure of some glasses of the composition $M^{III}M^{V}O_4$.
Phys. Chem. Glasses **4**, 229-233, 1963 (No. 6). *3319*

H. Bienfait: Externe en interne communicatie bij het onderzoek. (External and internal communication in research; in Dutch.)
Meded. Dir. Tuinb. **26**, 542-551, 1963 (No. 10). *3322*

G. Blasse: New types of cation-order in the rocksalt lattice: the structure of Li_3SbO_4 and Li_3NbO_4.
Z. anorg. allgem. Chemie **326**, 44-46, 1963 (No. 1/2). *3313*

G. Blasse and **D. J. Schipper:** Antiferromagnetism of $CoRh_2O_4$ and $NiRh_2O_4$.
Physics Letters **5**, 300, 1963 (No. 5). *3293*

P. B. Braun and **J. A. Goedkoop:** An X-ray and neutron-diffraction investigation of the magnetic phase $Al_{0.89}Mn_{1.11}$.
Acta crystall. **16**, 737-740, 1963 (No. 8). *3285*

H. Breimer: The influence of main- and vestigial-sideband widths on picture quality.
I.E.E. Conf. Rep. Series No. 5 (Television Engineering), 1962, pp. 5-11, published 1963. *3279*

H. Bremmer: Scattering by a perturbed continuum.
Symp. on electromagnetic theory and antennas, Copenhagen 1962, pp. 665-699, Pergamon Press, Oxford 1963. *3298*

A. Bril and **W. van Meurs-Hoekstra:** Verwendung von Beugungsgittern in kleinen Spiegelmonochromatoren.
Z. Instrumentenk. **71**, 232-234, 1963 (No. 8). *3297*

C. M. van der Burgt and **H. S. J. Pijls:** Motional positive feedback systems for ultrasonic power generators.
IEEE Trans. on Ultrasonics Engng. **UE-10**, 2-19, 1963 (No. 1). *3291*

C. M. van der Burgt and **A. L. Stuijts:** Developments in ferrite ceramics with strong piezomagnetic coupling.
Ultrasonics **1**, 199-210, Oct./Dec. 1963. *3303*

H. B. G. Casimir: Reciprocity theorems and irreversible processes.
Proc. IEEE **51**, 1570-1573, 1963 (No. 11). *3317*

E. H. P. Cordfunke and **A. A. van der Giessen:** Pseudomorphic decomposition of uranium peroxide into UO_3.
J. inorg. nucl. Chem. **25**, 553-555, 1963 (No. 5). *3283*

J. Dieleman: Paramagnetic resonance of a photosensitive centre in CdS:Cu,Ga.
Magnetic and electric resonance and relaxation, Proc. XIth Coll. Ampère, Eindhoven 1962, pp. 409-413, North-Holland Publ. Co., Amsterdam 1963. *3289*

G. Diemer and **B. Bölger:** Proposal for reduction of diffraction losses in *P-N* lasers.
Physica **29**, 600-601, 1963 (No. 6). *3287*

J. van Dijk, V. G. Keizer and **H. D. Moed:** Synthesis of β-phenylethylamine derivatives, VIII. Four diastereoisomers of 1-(4′-hydroxyphenyl)-2-(1″-methyl-3″-phenylpropylamino)propanol.
Rec. Trav. chim. Pays-Bas **82**, 189-201, 1963 (No. 2). *3284*

S. Duinker: Basic network elements for the synthesis of non-linear systems.
Monograph on radio waves and circuits, ed. S. Silver, pp. 320-329, Elsevier, Amsterdam 1963. *3290*

P. Eckerlin, I. Maak and **A. Rabenau:** Über Mischkristallbildung in den Systemen $(NH_4)_3AlF_6$-$(NH_4)_3GaF_6$ und LiAl-LiGa.
Z. anorg. allgem. Chemie **327**, 143-146, 1964 (No. 3/4). *A 93*

C. Fengler: The reflection of a pulse at an Epstein profile.
Proc. int. Conf. on the ionosphere, London 1962, pp. 400-405, publ. Inst. Phys./Phys. Soc., London 1963. *H 30*

S. Garbe: Desorptionsvorgänge in Ionisationsmanometern bei Beschuss von ölbedeckten Oberflächen mit langsamen Elektronen.
Vakuum-Technik **12**, 201-205, 1963 (No. 7). *A 83*

S. Garbe: Desorption experiments in an ultra high vacuum system, pumped by molecular sieve trapped oil diffusion pumps.
Physik und Technik von Sorptions- und Desorptionsvorgängen bei niederen Drücken, Vorträge 2. Europ. Symp. "Vakuum", Frankfurt/M. 1963, pp. 295-304, publ. R. A. Lang, Esch (Taunus). *A 86*

S. Garbe: The influence of the gas ambient on the emission properties of oxide-coated cathodes in receiving valves.
Suppl. Nuovo Cim. **1**, 810-824, 1963 (No. 2). *A 94*

J. A. Geurst: The reciprocity principle in the theory of magnetic recording.
Proc. IEEE **51**, 1573-1577, 1963 (No. 11). *3316*

W. van Gool: Vapour pressure measurements of Cd and Cd-Ag alloys at 950 °C.
Proc. Kon. Ned. Akad. Wet. **B 66**, 209-215, 1963 (No. 4). *3299*

W. van Gool: Theoretical considerations about the determination of the structure of lattice defects by phase equilibria.
Proc. Kon. Ned. Akad. Wet. **B 66**, 311-331, 1963 (No. 5). *3306*

W. J. A. Goossens and **H. J. G. Meyer**: Enkele basisbegrippen uit de fysica van halfgeleiders, II, III. (Some basic concepts of semiconductor physics; in Dutch.)
Ned. T. Natuurk. **29**, 387-399 and 409-418, 1963 (Nos. 10 and 11). *3309* and *3315*
(Sequel to *3025*.)

D. Gossel: Multivibratorschaltungen mit Transistoren für extrem grosse kontinuierlich steuerbare Frequenzvariation.
Nachrichtentechn. Z. **15**, 511-525, 1962 (No. 10). *H 28*

H. G. Grimmeiss and **H. Scholz**: Efficiency of recombination radiation in GaP.
Physics Letters **8**, 233-235, 1964 (No. 4). *A 88*

R. Groth and **E. Kauer**: Lichterzeugung mittels thermischer Selektivstrahler.
Z. angew. Phys. **16**, 130-143, 1963 (No. 2). *A 78*
See also Philips tech. Rev. **26**, 33-47, 1965 (No. 2).

J. Haantjes: Pick-up and display tubes for colour television.
I.E.E. Conf. Rep. Series No. 5 (Television Engineering), 1962, pp. 55-58, published 1963. *3280*

N. Hansen: Gettereigenschaften von nichtverdampften Gettern mit Porendiffusion und Chemisorption.
Vakuum-Technik **12**, 167-173, 1963 (No. 6). *A 79*

N. Hansen: Non-evaporating getters with surface adsorption and pore diffusion.
Suppl. Nuovo Cim. **1**, 627-640, 1963 (No. 2). *A 92*

N. Hansen and **W. Littmann**: Änderung des Halleffekts bei der Chemisorption an aufgedampften Metallfilmen.
Ber. Bunsenges. phys. Chemie **67**, 970-975, 1963 (No. 9/10). *A 90*

G. E. G. Hardeman: Electron and nuclear spin resonance in *n*-type silicon carbide.
J. Phys. Chem. Solids **24**, 1223-1231, 1963 (No. 10). *3296*

P. A. H. Hart: Interception, scattering and multivelocity effects in a transverse-wave electron beam.
Microwaves, Proc. 4th int. congress on microwave tubes, Scheveningen 1962, pp. 222-227, publ. Centrex, Eindhoven 1963. *3327*

H. U. Harten and **R. Memming**: Potential distribution at the germanium electrolyte interface.
Physics Letters 3, 95-96, 1962 (No. 2). *H 25*

J. Hornstra: Dislocations in spinels and related structures.
Materials Sci. Res. **1**, 88-97, 1963. *3300*

B. B. van Iperen: Experimental CW klystron for multiplication from 30 to 2.5 millimeters.
Proc. IEEE **51**, 935-937, 1963 (No. 6). *3286*

M. H. Jørgensen, N. I. Meyer and **K. J. Schmidt-Tiedemann**: Conductivity anisotropy of warm and hot electrons in silicon and germanium.
Solid State Comm. **1**, 226-233, 1963 (No. 7). *H 33*

E. Kauer: Optical and electrical properties of LaB_6.
Physics Letters **7**, 171-173, 1963 (No. 3). *A 82*

A. Klopfer: Effect of an electric discharge on the rates of adsorption on titanium of nitrogen and carbon monoxide.
Vorträge 2. Europ. Symp. "Vakuum" (see S. Garbe *A 86*), pp. 271-277. *A 91*

M. Koedam, A. A. Kruithof and **J. Riemens**: Energy balance of the low-pressure mercury-argon positive column.
Physica **29**, 565-584, 1963 (No. 5). *3281*

J. A. Kok, J. W. Poll and **C. E. G. M. M. van Vroonhoven**: Breakdown tests carried out on liquefied gases.
Appl. sci. Res. **B 10**, 257-268, 1963 (No. 3/4). *3305*

H. de Lang and **G. Bouwhuis**: A gas laser with a non-degenerate configuration of three plane mirrors.
Physics Letters **5**, 48-50, 1963 (No. 1). *3276*

H. de Lang and **G. Bouwhuis**: Experimental analysis of Zeeman polarisation effects in the output of a He-Ne laser.
Physics Letters **7**, 29-30, 1963 (No. 1). *3301*

J. J. van Loef and **P. J. M. Franssen**: The Mössbauer effect in the hexagonal ferrite $BaO.6Fe_2O_3$.
Physics Letters **7**, 225-226, 1963 (No. 4). *3304*

B. Lopes Cardozo and **F. F. Leopold**: Human code transmission. Letters and digits compared on the basis of immediate memory error rates.
Ergonomics **6**, 133-141, 1963 (No. 2). *3278*

J. L. Meijering: Usefulness of a $1/\gamma$ plot in the theory of thermal etching.
Acta metallurgica **11**, 847-849, 1963 (No. 8). *3292*

R. Memming: Surface recombination at higher injection levels.
Surface Sci. **1**, 88-101, 1964 (No. 1). *H 31*

R. Memming: Formation of fast surface states by cupric ions at the germanium-electrolyte interface.
Physics Letters **7**, 89-90, 1963 (No. 2). *H 38*

L. Merten: Modell einer Schraubenversetzung in piezoelektrischen Kristallen, I. Allgemeine Theorie — Elektrisches Feld bei Ladungsfreiheit — Ladungsverteilung bei Feldfreiheit; II. Elektrisches Feld und Ladungsverteilung für eine Versetzung in einem Eigenhalbleiter im thermischen Gleichgewicht.
Physik kondens. Materie **2**, 53-65 and 66-79, 1964 (No. 1). *A 89*

B. J. Mulder and **J. de Jonge**: On the sensitization of the photoconduction of anthracene by organic dyes.
Proc. Kon. Ned. Akad. Wet. **B 66**, 303-310, 1963 (No. 5). *3307*

D. J. van Ooijen and **W. F. Druyvesteyn**: Analogon of Barkhausen noise observed in a superconductor.
Physics Letters **6**, 30-31, 1963 (No. 1). *3294*

H. P. Peloschek: Square loop ferrites and their applications.
Progress in dielectrics **5**, 37-93, 1963. *3282*

G. H. Plantinga: Pulsed magnetrons for 4 and 2.5 mm wavelength.
Proc. 4th int. congress on microwave tubes (see P. A. H. Hart *3327*), pp. 202-205. *3326*
See also Philips tech. Rev. **25**, 217-226, 1963/64 (No. 9).

G. Prast: A Philips gas refrigerating machine for 20 °K.
Cryogenics **3**, 156-160, 1963 (No. 3). *3324*
See also Philips tech Rev. **26**, 1-11, 1965 (No. 1).

S. C. Rademaker and **H. J. de Rouw:** Quelques possibilités d'identification des cordes dans le verre.
Silicates industr. **28**, 541-544, 1963 (No. 12). *3311*

H. G. Reik: Optical properties of small polarons in the infrared.
Solid State Comm. **1**, 67-71, 1963 (No. 3). *A 80*

H. G. Reik, E. Kauer and **P. Gerthsen:** Optical properties of lanthanumcobaltite explained by small polaron theory.
Physics Letters **8**, 29-30, 1964 (No. 1). *A 85*

K. J. Schmidt-Tiedemann: Experimentelle Untersuchungen zum Problem der heissen Elektronen in Halbleitern.
Festkörperprobleme I, pp. 122-174, Vieweg, Brunswick 1962. *H 26*

K. J. Schmidt-Tiedemann: Birefringence by free carriers in semiconductors.
Rep. int. Conf. on the physics of semiconductors, Exeter 1962, pp. 191-196, publ. Inst. Phys./Phys. Soc., London 1962. *H 27*

S. Scholz: The density-time relation in hot pressing.
Planseeber. Pulvermetallurgie **11**, 82-84, 1963 (No. 2). *A 84*

S. Scholz and **B. Lersmacher:** Der Verdichtungsablauf beim Drucksintern.
Ber. Dtsch. Keram. Ges. **41**, 98-107, 1964 (No. 2). *A 87*

J. A. Schulkes and **G. Blasse:** Crystallographic and magnetic properties of the systems lithium ferrite-aluminate and lithium ferrite-gallate.
J. Phys. Chem. Solids **24**, 1651-1655, 1963 (No. 12). *3312*

H. Severin: Spinwellen und Spinresonanzen in ferrimagnetischen Oxyden.
Festkörperprobleme I, pp. 260-273, Vieweg, Brunswick 1962. *H 24*

H. Severin: Ferrite, chemische Zusammensetzung, Kristallstruktur und Herstellungsverfahren.
Sprechsaal für Keramik - Glas - Email **95**, 683-688, 1962 (No. 24). *H 29*

A. Smit and **P. Westerhof:** Investigations on sterols, XXI. An alternative route for the synthesis of some 6-dehydro-$9\beta,10\alpha$-steroid hormone analogues.
Rec. Trav. chim. Pays-Bas **82**, 1107-1114, 1963 (No. 11). *3310*

M. J. Sparnaay: The reaction between water vapor and the germanium surface.
Ann. New York Acad. Sci. **101**, 973-982, 1963. *3277*

M. J. Sparnaay: The interaction between germanium and cupric ions in an aqueous solution.
Surface Sci. **1**, 102-109, 1964 (No. 1). *3308*

A. Stegherr, P. Eckerlin and **F. Wald:** Untersuchung der Schnitte Ag_2Te-Bi_2Te_3 und $AgBiTe_2$-$PbTe$.
Z. Metallk. **54**, 598-600, 1963 (No. 10). *A 81*

J. M. Stevels and **J. Volger:** Impurity-induced imperfections and the dielectric properties of quartz crystals.
Phys. Chem. Glasses **4**, 247-252, 1963 (No. 6). *3318*

W. A. J. J. Velge and **K. J. de Vos:** Influence of milling upon the magnetic properties of the intermetallic compound MnAlGe.
J. appl. Phys. **34**, 3568-3571, 1963 (No. 12). *3320*

A. Venema: The interaction of gases and solids in practical devices.
Vorträge 2. Europ. Symp. "Vakuum" (see S. Garbe *A 86*), pp. 42-53. *3323*

M. T. Vlaardingerbroek and **K. R. U. Weimer:** Some aspects of the interaction of an electron beam and a plasma.
Proc. 4th int. congress on microwave tubes (see P. A. H. Hart *3327*), pp. 322-326. *3325*

J. Volger and **C. W. Berghout:** Fysische eigenschappen en chemie van harde supergeleiders (supergeleiders van de tweede soort). (Physical properties and chemistry of hard superconductors (superconductors of the second kind); in Dutch.)
Ned. T. Natuurk. **29**, 322-330, 1963 (No. 9). *3302*

J. H. N. van Vucht, H. A. C. M. Bruning and **H. C. Donkersloot:** New compounds related to the superconductors V_3Ga and Nb_3Sn.
Physics Letters **7**, 297, 1963 (No. 5). *3321*

J. S. van Wieringen, Y. Haven and **A. Kats:** Paramagnetic resonance of colour centres in alpha-quartz containing germanium.
Magnetic and electric resonance and relaxation, Proc. XIth Coll. Ampère, Eindhoven 1962, pp. 403-408, North-Holland Publ. Co., Amsterdam 1963. *3288*

J. S. van Wieringen and **J. G. Rensen:** Influence of lattice imperfections on the paramagnetic resonance of V^{2+} and Cr^{3+} in MgO.
Paramagnetic resonance, Proc. 1st int. Conf., Jerusalem 1962, Part I, pp. 105-112, Academic Press, New York 1963. *3314*

A. L. Zijlstra: The viscosity of some silicate glasses in connection with thermal history.
Phys. Chem. Glasses **4**, 143-151, 1963 (No. 4). *3295*

Volume 26, 1965, No. 3 pages 61-88 Published 15th October 1965

PHILIPS TECHNICAL REVIEW

VOLUME 26, 1965, No. 4/5/6

In October 1964, Philips Zentrallaboratorium GmbH introduced to the scientific world its new laboratory buildings in Aachen and Hamburg. This seemed to us a welcome opportunity to offer our readers a general survey of the work being done by these two laboratories. Surveys of this kind — the reader will recall our special number devoted to the Eindhoven Symposium in September 1963 — are in our opinion useful because of the cross-section they present of extensive fields of research, highlighted here and there by notable results and achievements.

Dr. A.E. Pannenborg, director of the Philips Research Laboratories, Eindhoven, and until 1st May 1963 the Managing Director of Philips Zentrallaboratorium GmbH, opens this number with a review of the origins and development of the Aachen and Hamburg laboratories. Dr. H. Bruining and Prof. Dr. Ir. S. Duinker, the present directors of the two laboratories, follow with a more detailed account of the research programme, which is illustrated in the remaining part of the issue by 12 articles — 6 from the Aachen and 6 from the Hamburg laboratory.

History of Philips research laboratories in Germany

A. E. Pannenborg

001.891:62.001.5

The reconstruction of the Philips industries in Germany after the Second World War was carried out with considerable assistance from the parent company in the Netherlands. For the German Philips group the war had entailed the loss of many factories, partly because of their situation in the eastern part of the country and partly through war damage. The reconstruction was directed with energy and far-sightedness from the head office in Hamburg; old factories were restored and many new ones founded, as in Aachen, Wetzlar and Krefeld. Towards the middle of the 'fifties the expansion of the Philips industries in Germany, and the resultant growth of business, reached a point where it became both desirable and financially possible to strengthen the capacity of the group by a highly qualified research force. Up to that point, attention had mainly been concentrated on the manufacture of mass-produced articles. It was now rightly considered essential to supplement the programme henceforth by investing in plant of high quality and by the manufacture of special products. The decision was therefore taken to found a research laboratory.

Of course, the Philips Concern as a whole had long possessed its own research facilities. During the First World War the parent company in Eindhoven created a research department which, under the direction of Dr. G. Holst, developed favourably. In the 'twenties and 'thirties this Philips research laboratory already had many distinguished scientists on its staff. For that reason the foundation of a research laboratory in Germany does not appear as an isolated event but as a continuation of a tradition. Up to the outbreak of the Second World War the scientific potential accumulated in Eindhoven was sufficient to provide the Concern with the fresh impetus needed. The vigorous post-war development of Philips throughout the world made it necessary to expand the Concern's research capacity, and led to the foundation of Concern research laboratories in Great Britain [1] and in France [2]. Clearly, in setting up a new laboratory the aim was to assign to it a local task as well as specific tasks in a coordinated international programme.

In the considerations that led to the foundation of a German laboratory, it soon appeared that national and international interests were to some extent in opposition. Since the head office of all German Philips firms is in Hamburg, it seemed obvious that Hamburg

Dr. A. E. Pannenborg, director, Philips Research Laboratories, Eindhoven.

[1] Mullard Research Laboratories, Redhill.

[2] Laboratoires d'Electronique et de Physique Appliquée, Paris.

Old gateway on the "Bodenhof" estate.

should also be the home of the new research centre. On the other hand, in order to facilitate frequent contacts between scientific staff and to ensure close co-operation both nationally and internationally, particularly with the large research centre in Eindhoven, the distance between the laboratories had not to be too great. These considerations resulted in the decision to establish two research laboratories at the same time, one in Hamburg and one in Aachen. The first was to be primarily concerned with subjects of immediate or prospective importance to the factories belonging to the German Philips group. The Aachen laboratory was to be mainly engaged on work forming part of the comprehensive Concern research programme, which did not necessarily include subjects of specific interest to the German factories.

The Aachen laboratory was opened in 1955, and the Hamburg laboratory in 1957. As long as the laboratories were small, it was advantageous to manage them within the organization of the German Holding Company, i.e. the Allgemeine Deutsche Philips Industrie GmbH. When the initial phase was completed and the Aachen and Hamburg laboratories each had a staff of more than a hundred, they were jointly given the status of a subsidiary company, which was formed on 1st October 1960 under the name of Philips Zentrallaboratorium GmbH.

The Aachen laboratory

The city of Aachen is very suitable for the establishment of an industrial research laboratory. The size of the town and its cultural life meet the needs of the young scientist. Moreover, we believe that it is very useful for our laboratories to be situated in a university town, or in a town which has a technological institute of university status (Technische Hochschule). Experience has confirmed this, and we gratefully acknowledge here the support we have received in so many forms from the Rheinisch-Westfälische Technische Hochschule. Finally, the presence of a large Philips manufacturing centre in Aachen contributed greatly towards the quick and successful formation of the laboratory. Initially the laboratory was housed in a former Philips factory building, this being always regarded as a provisional arrangement. Even before the laboratories were founded, the "Bodenhof", formerly a private estate, was purchased with a view to erecting a laboratory building. The building work on this site was started in 1961, and by 1964 the entire laboratory staff of 250 had moved in. The preliminary plans for the new laboratory building were the work of Prof. W. Fischer. The plans were worked out in further detail by the company's own architectural department, which also supervised the erection work. Thanks to the early acquisition of the land the laboratory has an exceptionally favourable situation, within convenient distance of the centre of the city and the residential areas.

In accordance with the idea, already mentioned, that the Hamburg sister laboratory should mainly devote itself to new fields of specific interest to the German factories, the Aachen laboratory was given a programme within the traditional fields of Philips activity.

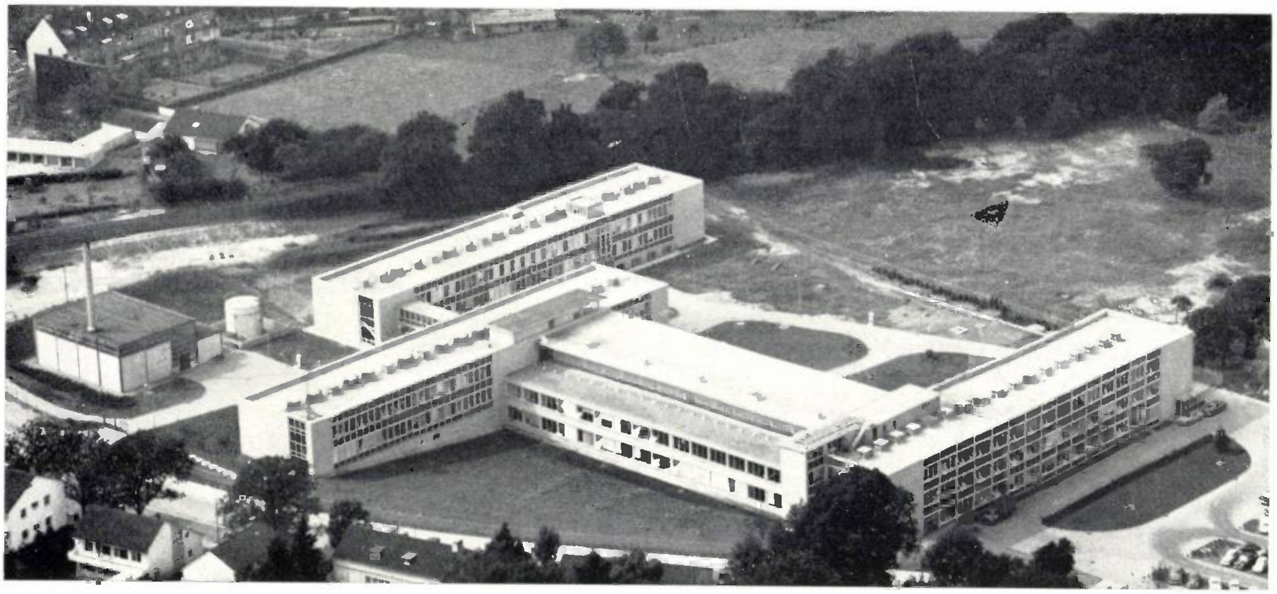
One of these was the production of light, an obvious choice since Philips owes its existence to the electric incandescent lamp, as still expressed in the name of the parent company. By the middle of the thirties the incandescent lamp had reached a high degree of technical perfection and consequently all major research projects in this field had by this time petered out, although of course the development laboratories were constantly working on the further improvement of the products. Research work had turned more towards the newer species of lamps, such as gas-discharge lamps and fluorescent lamps. After a gap of twenty years it appeared, however, that the tungsten lamp still presented a number of unresolved problems, which called for a further study of the thermal production of light. The investigations were especially concerned with the selective properties both of the radiating body and of the bulb.

The use of a high vacuum is a necessary condition in the manufacture of many Philips products. The technique of producing a vacuum in the pressure range from 10^{-4} to 10^{-6} torr had already been mastered in the laboratories and factories. At the time when the Aachen laboratory was founded there was a demand for considerably lower pressures, and moreover the means seemed to be available to produce them. The production and measurement of ultra-high vacua therefore became one of the topics of research. Valuable preliminary work had already been done in this field in the Eindhoven laboratories. For the accurate measurement of ultra-high vacua it is necessary to know not merely the total pressure but also the composition of the residual gases. The investigation revealed that the omegatron, already known in principle as a mass spectrometer, could be developed into an effective instrument for measuring partial pressures in the range between 10^{-5} and 10^{-11} torr. The opening up of the ultra-high vacuum range made it possible to embark on a thorough study of the many and various interactions between residual gases and solids, as occur for example inside a vacuum vessel. This led as a logical development to the study of getter substances, incorporated in sealed-off valves for the purpose of chemically absorbing residual gases in the system after evacuation and seal-off.

The most vigorous development in the electronics industry during the last 15 years has taken place in the applications of semiconductors. This development was ushered in by the invention of the transistor in 1948, but it has by no means remained confined to this device. It is therefore not surprising that the Aachen laboratory too should play its part in semiconductor research. Several Philips laboratories had already been engaged for some time in the investigation of germanium and silicon, the principal materials from which transistors and semiconducting diodes are made. The Aachen laboratory therefore turned to the study of semiconducting *compounds*. One of the fields in which such substances find application is thermoelectric cooling. Fundamental problems were tackled, in particular the properties of the substances and their control, and a start was made on the development of thermoelectric cooling devices. Some years later these devices were successfully put into production.

Another field of semiconductor research, one closely bound up with the generation of light, is *P-N* luminescence, i.e. the luminescence that can appear at the *P-N* junction of a semiconductor diode. The substances that can be used for producing such luminescent diodes have proved to be exceptionally difficult to control. We expect *P-N* luminescence to be of more importance in the new field called opto-electronics than in general lighting.

In all the above-mentioned fields, close cooperation is required between chemists, physicists and electrical engineers. Such cooperation is in line with the well-established and successful tradition of the Eindhoven Research Laboratories.

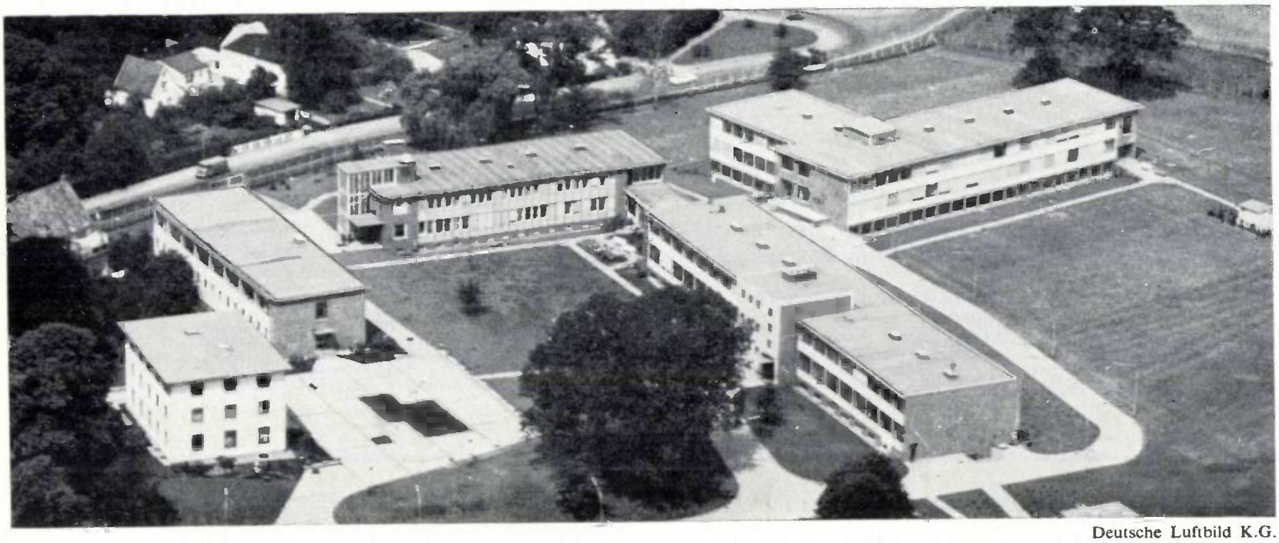


The Aachen laboratory.

Deutsche Luftbild K.G.

The Hamburg laboratory

Philips already had a small research laboratory in Hamburg before the war. It was founded under the name of "Studiengesellschaft", and occupied a former villa in Hamburg-Stellingen. It therefore seemed



Deutsche Luftbild K.G.

The Hamburg laboratory.

sensible to set up the new laboratory on the same site. Since the plans for the future were now much more ambitious, however, it was desirable to have more land available, and neighbouring plots were purchased. In the last 7 years various new buildings have been erected on this enlarged site, and old buildings either renovated or pulled down. The result is a fine, modern group of laboratory buildings at the edge of the Hamburg green belt. The staff of the Hamburg laboratory has now grown to nearly 250. The graduates among them, in keeping with the laboratory's general programme, are mainly electrical engineers and physicists.

As mentioned in the introduction, the Hamburg programme was to link up with existing branches of production, or to stimulate the manufacture of new products by the Philips industries in Germany. In addition, the laboratory was also prepared to undertake outside research projects.

The first major studies were in the field of microwave physics and engineering, as applied for example for radar purposes and in microwave radio links. The components and circuits were at the time basically well established, but much still remained to be done in order to determine the properties and behaviour of special materials at these frequencies. The materials concerned were mainly magnetic ceramics (ferrites), whose discovery was the result of earlier work in the Eindhoven research laboratories. At microwaves it is possible with ferrite substances to achieve effects that cannot be obtained with metallic magnets. The preparation of such substances was started in the Hamburg laboratory some years after its foundation, at which time the conditions were created for extensive experimental research. Other experiments related to the generation of very short microwaves by means of valve generators.

On the basis of this programme, links were established with the factories engaged on the manufacture of microwave components, ceramics and electron tubes. Another important activity in Hamburg is the manufacture of transistors and semiconductor diodes, and of the germanium and silicon used to make them. Although sufficient laboratory capacity was available in the factories for the development of new types and new circuit elements, the situation regarding the study of the surface properties of these solid-state devices was quite different. In the early years of the transistor it had been hoped that the properties of these devices would not change in the course of time, unlike electronic valves which, owing to the high temperature required for thermionic emission, have a limited life. Practice has shown, however, that the stability of transistors too is a problem. In this connection the surface state is an essential factor. A fundamental research programme was therefore set up for studying the surface properties in the interface between germanium or silicon and a liquid electrolyte.

A second item on the programme of semiconductor research was the behaviour of germanium under the influence of high electric fields. The study of the "hot electrons" thereby produced in germanium has yielded valuable contributions towards the understanding of the inner electronic structure of this element.

A third topic of research is the measurement and control of industrial processes by electronic means, i.e. industrial electronics. In this field there is a very wide scope for electronic applications. Digital techniques, developed for electronic computers, are now being used here in modified forms. One section of the relevant research group is therefore examining the possible ways in which digital circuits can be employed for solving a wide variety of control engineering problems.

Finally, Philips did not want to be outside the field of electronic computers, now rapidly increasing in importance. The scale of technical and scientific know-how required in this field is extremely wide, and calls for the coordinated efforts of all those within the Concern who can usefully contribute. The task that fell to the Hamburg laboratory was to deal with the electro-mechanical problems involved, in particular with the development of peripheral equipment (input and output devices).

The research programme of the Philips laboratories in Germany

H. Bruining and S. Duinker

001.891:62.001.5

The foregoing article by A. E. Pannenborg describes in broad lines the fields of research on which, in accordance with the initial plans, our laboratories in Hamburg and Aachen are engaged. The various investigations pursued as part of this research programme differ very considerably both in character and in subject matter. On the one hand, basic and exploratory problems are being studied with the object of gaining a deeper insight into certain fields of physics, chemistry and electrical engineering. Other work is aimed at mastering the technology of many new materials, substances and electronic components. A further research group is concerned with the applications of these materials and substances in all kinds of electrical, electronic and mechanical applications and equipment. Finally, at the other end of the range of activities, investigations are carried out into problems of a more practical nature, for which solutions are sought on the basis of new technological means. As regards the diversity of subjects, the wide scope of the Philips research programme is dictated by the Concern's sphere of interest, based on the three "main pillars": consumer goods, components and professional equipment.

In carrying out research work in such a diversity of fields, considerable efforts are made to ensure the most effective cross-fertilization between the various disciplines. This is necessary if only for the reason that the problems occupying a particular research group are very often affected to a considerable extent by the results and requirements of other groups. For instance, consider a research group working on the exploitation of physical effects of new materials with a view to their application in new types of circuit device. For this purpose it is necessary, right from the beginning, to take proper account of many aspects and demands of the future fields of application of these devices. To carry out such a programme it is obviously most essential to have good coordination and cooperation between the different research groups within the laboratory and the relevant departments within the Concern. Only in this way can the main objective of industrial research be reached, which is to support the existing production and to open up new spheres of activity for the enterprise as a whole.

As an introduction to the contributions in the present issue from members of our two laboratories, this article will describe at somewhat greater length the research programme already outlined by Dr. A. E. Pannenborg, and an attempt will be made to indicate the above-mentioned links between the various research activities. It will be useful to deal with these under two headings, *solid-state research* and *systems-oriented research*. The first is concerned with the physics and chemistry of the solid state and their applications in electrical and electronic engineering; the second is mainly concerned with control engineering, data processing and related subjects.

Solid-state research

It is common knowledge that electrical engineering largely owes its development to the existence of metals and alloys which possess particular or indeed extreme properties. Examples are the high conductivity of

Dr. H. Bruining is director of the Aachen laboratory of Philips Zentrallaboratorium GmbH.
Prof. Dr. Ir. S. Duinker is director of the Hamburg laboratory of Philips Zentrallaboratorium GmbH, and professor extraordinary at the University of Groningen.

niques, developed for electronic computers, are now being used here in modified forms. One section of the relevant research group is therefore examining the possible ways in which digital circuits can be employed for solving a wide variety of control engineering problems. .

Finally, Philips did not want to be outside the field of electronic computers, now rapidly increasing in importance. The scale of technical and scientific know-how required in this field is extremely wide, and calls for the coordinated efforts of all those within the Concern who can usefully contribute. The task that fell to the Hamburg laboratory was to deal with the electro-mechanical problems involved, in particular with the development of peripheral equipment (input and output devices).

The research programme of the Philips laboratories in Germany

H. Bruining and S. Duinker

001.891:62.001.5

The foregoing article by A. E. Pannenborg describes in broad lines the fields of research on which, in accordance with the initial plans, our laboratories in Hamburg and Aachen are engaged. The various investigations pursued as part of this research programme differ very considerably both in character and in subject matter. On the one hand, basic and exploratory problems are being studied with the object of gaining a deeper insight into certain fields of physics, chemistry and electrical engineering. Other work is aimed at mastering the technology of many new materials, substances and electronic components. A further research group is concerned with the applications of these materials and substances in all kinds of electrical, electronic and mechanical applications and equipment. Finally, at the other end of the range of activities, investigations are carried out into problems of a more practical nature, for which solutions are sought on the basis of new technological means. As regards the diversity of subjects, the wide scope of the Philips research programme is dictated by the Concern's sphere of interest, based on the three "main pillars": consumer goods, components and professional equipment.

In carrying out research work in such a diversity of fields, considerable efforts are made to ensure the most effective cross-fertilization between the various disciplines. This is necessary if only for the reason that the problems occupying a particular research group are very often affected to a considerable extent by the results and requirements of other groups. For instance, consider a research group working on the exploitation of physical effects of new materials with a view to their application in new types of circuit device. For this purpose it is necessary, right from the beginning, to take proper account of many aspects and demands of the future fields of application of these devices. To carry out such a programme it is obviously most essential to have good coordination and cooperation between the different research groups within the laboratory and the relevant departments within the Concern. Only in this way can the main objective of industrial research be reached, which is to support the existing production and to open up new spheres of activity for the enterprise as a whole.

As an introduction to the contributions in the present issue from members of our two laboratories, this article will describe at somewhat greater length the research programme already outlined by Dr. A. E. Pannenborg, and an attempt will be made to indicate the above-mentioned links between the various research activities. It will be useful to deal with these under two headings, *solid-state research* and *systems-oriented research*. The first is concerned with the physics and chemistry of the solid state and their applications in electrical and electronic engineering; the second is mainly concerned with control engineering, data processing and related subjects.

Solid-state research

It is common knowledge that electrical engineering largely owes its development to the existence of metals and alloys which possess particular or indeed extreme properties. Examples are the high conductivity of

Dr. H. Bruining is director of the Aachen laboratory of Philips Zentrallaboratorium GmbH.
Prof. Dr. Ir. S. Duinker is director of the Hamburg laboratory of Philips Zentrallaboratorium GmbH, and professor extraordinary at the University of Groningen.

copper and the magnetism of iron. The advances made and the new possibilities opened up in the course of the years could hardly be imagined if materials possessing such properties had not been available. For a company like Philips, operating in so many branches of electrical and electronic engineering, further development is only possible if endeavours to achieve new products and applications are backed up by intensive research in the field of solid-state physics. The research can be prompted by the desire to improve existing products or to develop new ones. On the other hand, a new product may be the unexpected outcome of a new effect, or a physical property discovered by chance.

If a substance, e.g. a novel type of semiconductor, is to be given a technical application, it is necessary to be able to control the properties of the substance. One should be able to produce it in a pure state and moreover in a form suitable for experiment or application, e.g. as a single crystal, as sintered material or as a thin film. This "control of materials" is an essential aspect of our work, not only because it is necessary to know the material, whose physical properties after all govern the application, but also because in any subsequent application reproducible results can only be obtained if the composition of the material and the process by which it is produced are exactly known.

As belonging to this "control of materials" we can also consider the research work concerned with the production and application of ultra-high vacua. A "thin film", whose structure and composition must in many cases be well known, can only be obtained reproducibly if the vacuum is sufficiently high or the gas atmosphere sufficiently well defined. The "thin film", by which is meant a layer with a thickness between 1 and 1000 atomic layers, will be of considerable importance in the technology of the future. To produce such films in defined conditions, vacua with pressures down to 10^{-9} torr or lower are often necessary.

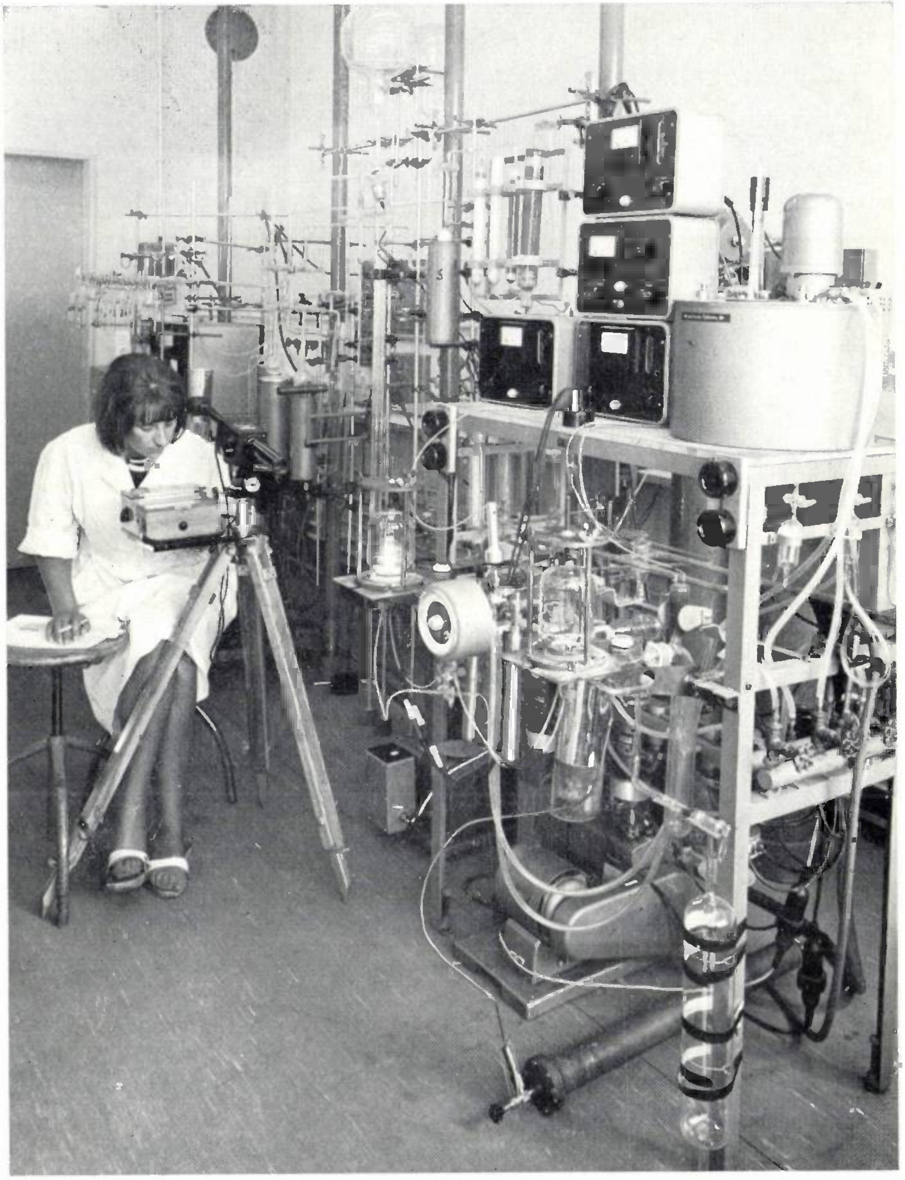
The control of materials has reached a particularly high stage of advancement in the semiconducting elements germanium and silicon. Nowadays the best and purest artificial single crystals are made from these elements. On the other hand, relatively little is known about the surface states of such crystals. Impurities are usually adsorbed on their surfaces, in the form at least of a monolayer, which may substantially affect the properties of the semiconducting material or of the device made from it. The smaller the sample the more noticeable the effect, and it is therefore of special importance in miniature devices. For the purposes of fundamental research the only conclusive experimental methods have so far been those in which the surface is immersed in a fluid, preferably water [1] and this condition is very remote from the actual conditions of operation. If the foreign matter adsorbed does not appear in the form of a monatomic layer — and it seldom does — it finally forms what in thermodynamic terms is called a separate phase, i.e. a multilayer, which in turn has to be investigated. Optical methods can be particularly useful here. For example, light-absorption measurements yield information on the chemical composition of the multilayer. Such measurements are already possible in the near infra-red, since germanium and silicon are transparent to radiation in this spectral range.

The incandescent lamp is Philips' oldest product. Its filament presents an example of a whole series of extreme demands with respect to the materials employed. Wien's Displacement Law requires the highest possible temperatures in order to shift the radiation maximum well towards the visible part of the spectrum while, at the same time, reducing the radiation in the infra-red. Materials are thus wanted that combine an extremely high melting point with a low vapour pressure and which in addition radiate selectively within the visible part of the spectrum. The two first requirements limited the choice to such an extent that for more than fifty years tungsten was almost the only material employed. It is possible, however, to return the evaporated material to the filament by means of a regenerative cycle. For this purpose suitable transport media are found in the halogen family, of which fluorine has particularly interesting possibilities [2]. The use of an appropriate process cycle might compensate the evaporation almost entirely, thus providing a wider choice of materials for the filament.

The return of the evaporated metal to the filament is essentially a chemical transport reaction. These reactions are also of interest in the production of thin films and of pure materials in polycrystalline or single-crystal form [3].

Other interesting subjects of fundamental research are phenomena that cannot completely be explained in terms of classical physics. They include many interactions between electrons, light and sound quanta, particularly those that occur under extreme physical conditions such as very strong magnetic or electric fields, extremely low temperatures, etc., and which come under the collective heading of "transport phenomena" [4]. It would be going too far to deal here with the many and various methods of experimental physics which are employed in the study of such transport phenomena. They include e.g. optical experiments and measurements of the anisotropy of electrical conductivity. Mention may also be made of a little known method, viz, the measurement of the propagation properties of ultrasonic vibrations (from 50 c/s to 3 Gc/s) in solids. These measurements yield data, for example, on the mobility of electrons. Other interaction phenomena, e.g. with lattice imperfections, can also be studied in this way. An example of practical importance of this field of research is the ultrasonic amplifier operating on the travelling wave principle.

Optical measurements on metals and semiconductors are on the one hand a means of obtaining information on their physical properties, and on the other hand they have a close relation with the practical field of light production. Most light sources not only radiate in the visible part of the spectrum but also to an appreciable extent in the infra-red, which represents a loss of energy. Since certain substances exhibit selective reflection, in the sense that they pass visible light without severe absorption loss whereas the infra-red is almost completely reflected, it is possible to increase the luminous efficiency of light sources by coating lamp bulbs with a thin layer of such substances. This principle has been advantageously applied in sodium lamps [5].



Part of one of the fluorine laboratories at the Aachen laboratory.

In this context we should mention also our investigation of *P-N* luminescence, for this again is concerned with the production of light. A special study has been made of gallium phosphide, which was the first known material to give a quantum yield of practical interest at room temperature (about 1 per cent). In addition, GaP and other III-V compounds are photoconductors. By combining electroluminescent GaP with photo-conducting materials novel types of circuit were obtained which led to a new kind of circuitry, known by the name of "opto-electronics". As regards applications this interesting field is still in the development stage; some examples are mentioned in this issue [6]. In the present state of the art, *P-N* luminescence cannot yet be employed for lighting purposes, although it might conceivably be used even now for luminous indicator devices.

As an example of a substance showing an unexpected new effect, mention may be made of barium titanate, which has long been used as a dielectric in capacitors, and which, moreover, has interesting piezo- and ferroelectric properties. The new effect consists in a very abrupt increase, within a small temperature interval, of

[1] H. U. Harten, R. Memming and G. Schwandt, Investigations on the germanium-electrolyte interface; page 127 of this number.

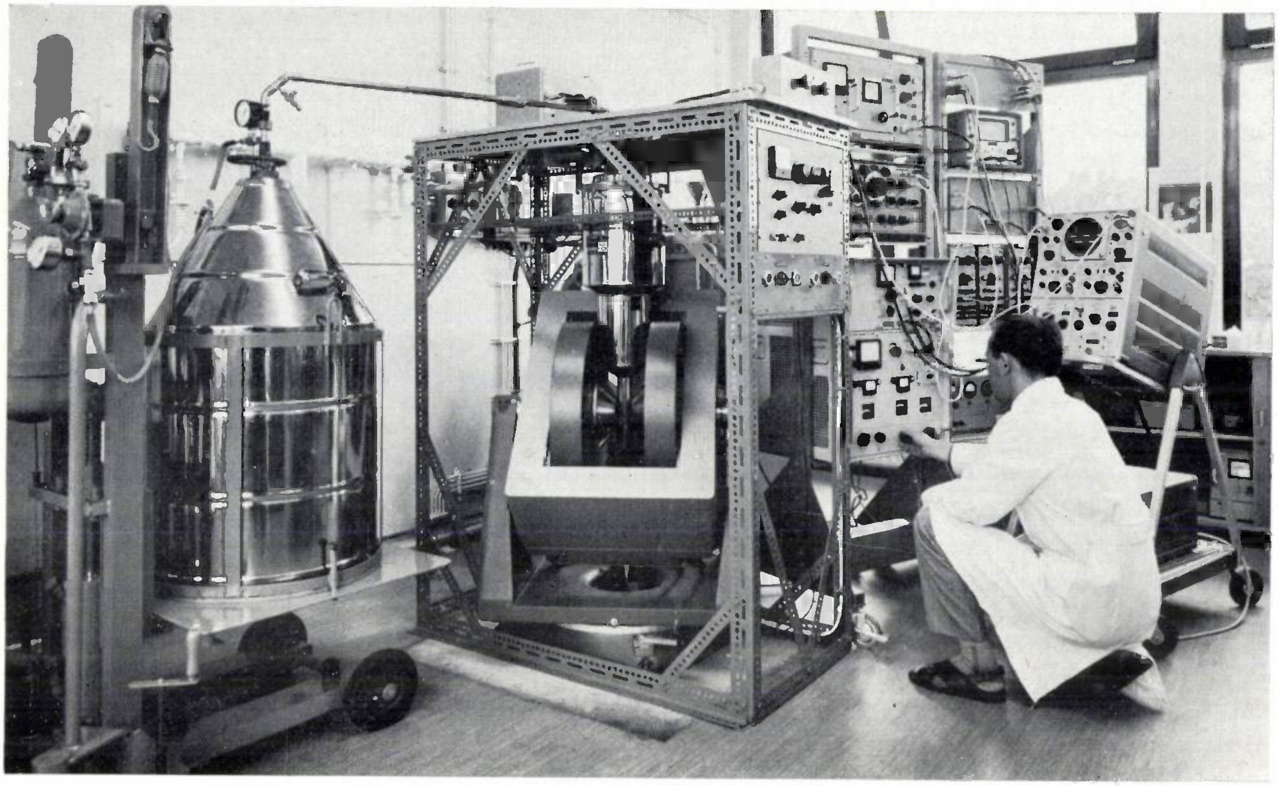
[2] J. Schröder, Examples from fluorine chemistry and possible industrial applications; page 111 of this number.

[3] A. Rabenau, Chemical transport reactions; page 117 of this number.

[4] K. J. Schmidt-Tiedemann, Transport phenomena in solids; page 99 of this number.

[5] R. Groth and E. Kauer, Thermal insulation of sodium lamps; page 105 of this number.

[6] H. G. Grimmeiss, W. Kischio and H. Scholz, Gallium phosphide light sources and photocells; page 136 of this number.



Experimental equipment in the Hamburg laboratory for investigating the interaction of electron currents with sound waves in solids.

the electrical resistivity which has been reduced by doping. Within this transition range the resistivity of barium titanate shows a temperature coefficient which is the highest yet known. Apart from investigations into the possible uses of this effect some effort is being spent on the development of a more exact theory of the mechanism involved [7].

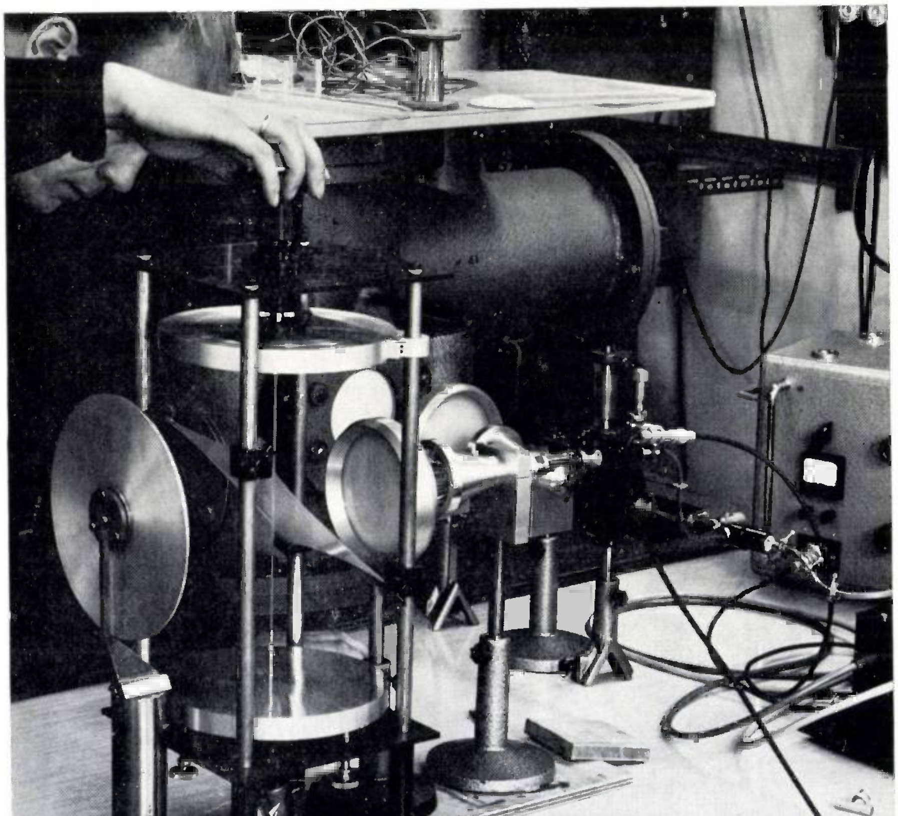
Systems-oriented research

Some of the investigations mentioned in the foregoing have been concerned with new or improved materials and devices, or with improved methods of producing them. Investigations of this kind may give rise to fields of applications previously unknown. In contrast to this, the field of research which we have grouped under the heading "systems-oriented research" is concerned, in contrast, with turning new technical principles and possibilities to practical use in the various types of system. The term "system" in this connection refers equally to relatively small apparatus, such as electronic measuring instruments, as to extremely complex electronic equipment. Examples of the latter category include microwave transmission systems, industrial measuring and control equipment, and electronic data-processing systems for administrative, commercial, technical or scientific purposes. The essential characteristic of this category of research, then, is that it deals primarily with the actual practical problem that has to be solved by the system. A thorough study of these practical problems can be regarded as a first step towards the formulation of the research programme. The next step is to examine whether, and in how far, the demands using the given analysis of the system and the available components can be met in optimum fashion and in accordance with the latest technical know-how.

Using the well-known ferromagnetic materials, ferroxcube and ferroxdure as basic material, a series of investigations were carried out in the higher frequency ranges, which are of fundamental importance, for example, for the reception of metre waves and for radar applications. The demand for a "small" VHF aerial (i.e. shorter than the wavelengths to be received) led to the development of ferrite aerials (magnetic aerials). In the microwave range, ferrites can be used to produce electrically controllable phase shifts between signals in various parts of a system. This prompted investigations into the possibility of magnetically rotating the directional pattern of a microwave aerial. Closely bound up with this work are investigations of the behaviour of ferrite phase-shifters under high-power conditions where non-linear effects start to play a significant role. Another interesting application of ferrites is for stabilizing the output power of a klystron, using an electrically-controlled ferrite attenuator.

In the range of millimetre waves, ferrites gradually reach the limit of their usefulness and moreover it becomes necessary to look for essentially new ways and means of generating and propagating waves; in this realm too a great deal of important research remains to be done. The work undertaken in this connection has included the propagation of millimetre waves along dielectric transmission lines, the development of a millimetre wave frequency standard, and the generation of frequencies above 300 Gc/s by frequency multiplication using the non-linear field emission effect An interesting development in the course of this work was a superconducting resonator with an extremely high figure of merit.

Another example of systems-oriented research is our activity in the field of input and output equipment for data-processing systems. In the last two decades, electronic data processing has undergone extraordinary development and has gained a foothold in the most diverse fields. Since, despite the considerable differences as regards application, the computations and correlations carried out with the data are limited to relatively few basic arithmetic operations, such as addition and multiplication, a computer can in principle be designed which is capable of handling a variety of problems. In order to employ such a multi-purpose computer for certain special tasks, however, it is necessary to have peripheral equipment which collects, sorts and checks the data and presents it to the computer in a machine-readable form together with the computer programme. At the output side the results have to be delivered in the desired form, that is to say printed or in the form of graphs. It is obvious that, depending on the field of application, a wide range of peripheral equipment is needed for carrying out these tasks. The general aspects and trends in the development of peripheral equipment are dealt with in one of the articles in this issue [8].



Apparatus developed in the Hamburg laboratory for the transmission and measurement of millimetre waves.

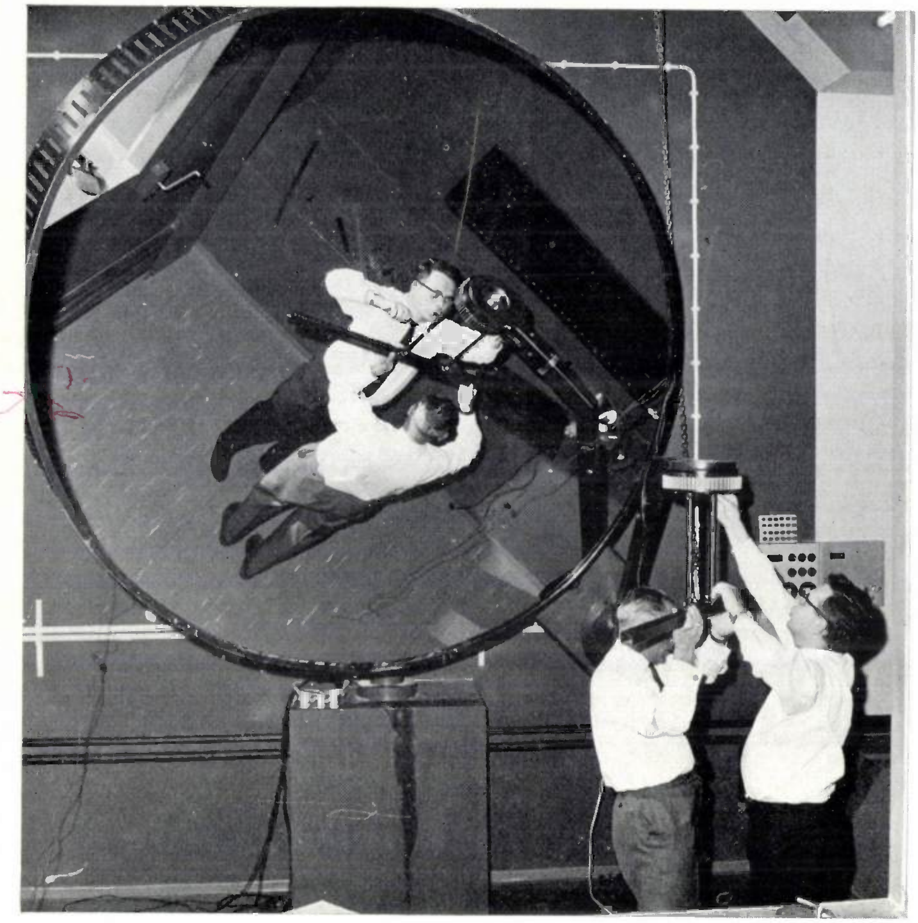
The investigation of optimum coding systems for error-free data transmission or the avoidance of reading-in errors, is a problem of a mathematical nature; on the other hand, the design of equipment to achieve a given residual error probability on the basis of this mathematical knowledge is a problem of circuitry [9]. The input of the data is usually effected by means of a punched tape. This is read, for example, by a photoelectric device. With a view to increasing the flow of data, efforts are being made to increase the input speed, always provided, however, that the tape can be stopped at any time before the next character is reached. This leads to interesting electro-mechanical problems, which are being investigated.

Similar problems exist at the output side of a data processing system. To increase the flow of data using a punched tape at the output, attempts are being made to increase the punching speed. In order to achieve substantially higher punching speeds than the inertia

[7] E. Andrich and K. H. Hardtl, Investigations on $BaTiO_3$ semiconductors; page 119 of this number.

[8] G. Haas, Problems and trends in the development of peripheral equipment for computers; page 148 of this number.

[9] G. Renelt and J. Schröder, Data checking during input and transmission by means of one or two check characters; page 156 of this number.



Rotatable mirror in the lighting engineering department of the Aachen laboratory, used for measuring the distribution of light from luminaires.

of mechanical punching equipment allows, alternative physical principles can be adopted. One such alternative is punching by means of a shock wave produced by a spark discharge. The punching is then achieved within a few microseconds.

As a rule the output data is delivered in normal print, frequently with several copies. To increase print-out speeds, use can be made of electrostatic electrography based on a mosaic print principle, and the properties required of the electrogram carrier for this purpose have been the subject of research in the laboratory. An entirely different method of printing, again using shock-wave energy, is dealt with in the previously cited article [8].

This peripheral equipment involves numerous problems of electronic circuitry. For instance, for the purposes of experimenting with input and output apparatus a novel type of pulse generator had to be developed, which is capable of producing the various pulse patterns required for driving and testing the individual peripheral devices.

In the sector of control engineering there are again numerous other problems that call for fundamental scientific study in the research laboratory. Unlike data processing, process control is a branch of engineering which already has a long history. There is no reason to believe, however, that it has yet reached its final stage of development. On the contrary, the demands regarding automation, accuracy and reliability, and the fact that in this field too digital process control computers are increasingly gaining importance, has produced a whole series of extremely interesting problems which are not encountered at all in analogue process control systems. This is not to say, however, that analogue principles are to be regarded as out of date. Especially for small and relatively simple systems, and where price is a more important consideration than speed, analogue methods still retain their usefulness, and research in both directions is pursued in our laboratories.

The analysis of various industrial processes has resulted in a system philosophy based on a new system of digital building blocks briefly described in this issue [10]. The fundamental differences between the primary demands of data processing on the one hand and of process control on the other give rise to different approaches, the results of which are not always readily compatible. Incidentally, here again, it was found that research directed towards a specific end sometimes leads quite unexpectedly to results applicable to entirely different fields of application. A surprising by-product of the above-mentioned investigations was a digital tuning instrument, which provides an extremely simple method of tuning a piano or organ [11].

Just as in data processing, peripheral devices play a very important part in process control. They include on the one hand pick-ups which convert the physical quantities to be measured into an electrical analogue or digital signal, and on the other hand the control instruments which intervene in a process when fed such electrical signals. Some examples should also be given of the investigations carried out in this field in our laboratories. The development of semiconductors has led to new types of strain gauge, as used for the measurement of mechanical forces. Research on possible applications of springs clamped at one end has resulted in a method of modulating infra-red radiation as well as microwaves. This type of work requires the coordinated investigation of mechanical, electrical and physical problems. This was seen very distinctly, for example, in the development of a method for continuously measuring the density of fluids by means of mechanical vibrations. Also included in this category of instruments are a zero-point thermostat, which makes elegant use of the Peltier effect, and a light probe for contactless displacement measurements in the micron range. Research in all these cases was concentrated on the discovery of new measuring principles.

As regards our investigations on actuators mention is made here of a "synchronized induction motor", where electronic know-how led to interesting new possibilities in the field of electrical machines. This motor has the unique property of possessing a load-independent control range at any speed of revolution between zero and the nominal value. Other advances in the development of small electric motors are due to the recent developments of ferrites which, because of their high resistivity and their mechanical properties are superior to conventional metallic magnetic materials even at low frequencies, in special cases, in spite of their lower permeability and saturation magnetization. It has proved possible, for example, to build a self-starting synchronous motor with no commutator using a rotor consisting of a permanent magnetic ferrite. This motor provides interesting possibilities for electric shavers and record players [12].

One part of the Aachen laboratory of Philips Zentrallaboratorium GmbH has now been equipped as a lighting engineering laboratory. Its task will be to study the principles underlying the optimum practical application of light sources and luminaires (lighting fittings). Prominence is given in this work to the interactions between light and the conditions of vision, and to their effect on the design of lighting fittings. In the field of light measurement special devices are being developed for studying the behaviour of lamps, luminaires and structural materials.

[10] D. Gossel, G. Kaps and W. Schott, A new system of digital circuit blocks for industrial measuring and control equipment; page 164 of this number.

[11] D. Gossel, Generation of musical intervals by a digital method; page 170 of this number.

[12] R. Thees, Small electric motors; page 143 of this number.

Transport phenomena in solids

K. J. Schmidt-Tiedemann

537.311.33

Introduction

By "transport phenomena" the physicist understands the relation between flows and forces in the widest sense. Simple examples are the conduction of electricity across a potential drop, the conduction of heat as a result of a temperature difference, or diffusion due to a concentration gradient.

Every semiconducting device usually utilizes more than one of these transport mechanisms, and the work of the development scientist can in fact be defined as the task of determining and realizing that combination of transport phenomena which, because of their natural relationships, best meet the technical objective aimed at. This objective can be reached the more easily and elegantly the greater the extent of our knowledge of transport phenomena. This, then, is one of the reasons for the major importance of fundamental scientific research in the field of solid-state physics.

The method by which the answer to a complicated technical problem can, as it were, "automatically" be found by the application of an appropriate natural law will be explained below with two examples.

Let us first consider the problem, frequently encountered in electronic measurement techniques, of measuring an electric current in a very wide range with the same relative accuracy everywhere. A device that can do this, known as a "logarithmic amplifier", can be designed with the aid of the Boltzmann distribution law. This law guarantees (within certain limits) an exact logarithmic relation between e.g. the current and voltage of a diode. The technical problem of ex-

Dr. K. J. Schmidt-Tiedemann is a research worker at the Hamburg laboratory of Philips Zentrallaboratorium GmbH.

our laboratories. The development of semiconductors has led to new types of strain gauge, as used for the measurement of mechanical forces. Research on possible applications of springs clamped at one end has resulted in a method of modulating infra-red radiation as well as microwaves. This type of work requires the coordinated investigation of mechanical, electrical and physical problems. This was seen very distinctly, for example, in the development of a method for continuously measuring the density of fluids by means of mechanical vibrations. Also included in this category of instruments are a zero-point thermostat, which makes elegant use of the Peltier effect, and a light probe for contactless displacement measurements in the micron range. Research in all these cases was concentrated on the discovery of new measuring principles.

As regards our investigations on actuators mention is made here of a "synchronized induction motor", where electronic know-how led to interesting new possibilities in the field of electrical machines. This motor has the unique property of possessing a load-independent control range at any speed of revolution between zero and the nominal value. Other advances in the development of small electric motors are due to the recent developments of ferrites which, because of their high resistivity and their mechanical properties are superior to conventional metallic magnetic materials even at low frequencies, in special cases, in spite of their lower permeability and saturation magnetization. It has proved possible, for example, to build a self-starting synchronous motor with no commutator using a rotor consisting of a permanent magnetic ferrite. This motor provides interesting possibilities for electric shavers and record players [12].

One part of the Aachen laboratory of Philips Zentrallaboratorium GmbH has now been equipped as a lighting engineering laboratory. Its task will be to study the principles underlying the optimum practical application of light sources and luminaires (lighting fittings). Prominence is given in this work to the interactions between light and the conditions of vision, and to their effect on the design of lighting fittings. In the field of light measurement special devices are being developed for studying the behaviour of lamps, luminaires and structural materials.

[10] D. Gossel, G. Kaps and W. Schott, A new system of digital circuit blocks for industrial measuring and control equipment; page 164 of this number.

[11] D. Gossel, Generation of musical intervals by a digital method; page 170 of this number.

[12] R. Thees, Small electric motors; page 143 of this number.

Transport phenomena in solids

K. J. Schmidt-Tiedemann

537.311.33

Introduction

By "transport phenomena" the physicist understands the relation between flows and forces in the widest sense. Simple examples are the conduction of electricity across a potential drop, the conduction of heat as a result of a temperature difference, or diffusion due to a concentration gradient.

Every semiconducting device usually utilizes more than one of these transport mechanisms, and the work of the development scientist can in fact be defined as the task of determining and realizing that combination of transport phenomena which, because of their natural relationships, best meet the technical objective aimed at. This objective can be reached the more easily and elegantly the greater the extent of our knowledge of transport phenomena. This, then, is one of the reasons for the major importance of fundamental scientific research in the field of solid-state physics.

The method by which the answer to a complicated technical problem can, as it were, "automatically" be found by the application of an appropriate natural law will be explained below with two examples.

Let us first consider the problem, frequently encountered in electronic measurement techniques, of measuring an electric current in a very wide range with the same relative accuracy everywhere. A device that can do this, known as a "logarithmic amplifier", can be designed with the aid of the Boltzmann distribution law. This law guarantees (within certain limits) an exact logarithmic relation between e.g. the current and voltage of a diode. The technical problem of ex-

Dr. K. J. Schmidt-Tiedemann is a research worker at the Hamburg laboratory of Philips Zentrallaboratorium GmbH.

cluding unwanted effects (such as leakage currents, etc.), can be solved very elegantly by means of a suitable negative feedback arrangement [1]. In this way, for example, the measurement can be given a relative accuracy of 1% in a current range from one to ten-thousand-million (1 : 10^{10}). The relation between the measured current and the output signal involves, apart from the temperature (which should be appropriately stabilized), only the two physical constants e and k, the elementary charge and the Boltzmann constant. If, on the other hand, the current can be kept constant, the output signal depends in an exactly known way upon the temperature, and we have an electronic thermometer.

As a second example we shall consider an infra-red radiation detector. Semiconductors whose electrical resistance changes under irradiation generally have a wide spectral sensitivity range. If a magnetic field is applied at the same time, however, under certain conditions the sensitivity is concentrated in a small part of the spectrum. By varying the magnetic field it is possible to tune the detector, like a radio receiver, to a particular wavelength. This instrument [2], which may also be of technical interest in connection with the transmission of signals by means of light waves, is again based on a "fundamental physical effect". What happens is that, in the magnetic field, the allowed energy states of the conduction electrons are bunched together in a complicated way, which can only be understood in quantum-mechanical terms, to form what are known as Landau levels.

In the following attempt to provide a survey of typical transport phenomena in semiconductors (or, more generally, in solids) it will be necessary, in view of the enormous scope of the subject, to make a selection. To preserve some uniformity we shall confine ourselves to effects in which mechanical deformations of the solid (i.e. tensile, compressive or shear stresses) play an essential part. Even with this limitation, we shall still be unable to mention many highly interesting and valuable results. To make the subject matter easier to understand we have dispensed for the greater part with mathematical derivations and exact details. A more thorough treatment of the problems concerned will be found in the original papers to which reference will be made in the text.

Static elastic deformation

The electrical resistivity of certain semiconductors changes in an exceptionally marked way when the substance is subjected to a uniaxial tensile or compressive stress [3]. The relative changes in N-type silicon, for example, can be as much as 1% per 100 kg/cm^2, which is about one-hundred times higher than comparable values for metals. To explain this effect the first consideration is that, owing to interaction with the crystal lattice, the "effective mass" of the conduction electrons — which relates the force exerted by an external field to the acceleration — can differ considerably from the mass m_0 of a free electron. In silicon there are indeed several "kinds" of conduction electrons, which exist in various energy states, that is to say in sublevels or "valleys" as they are called. For example, with an electric field along a four-fold axis of symmetry two-thirds of the electrons have a mass of 0.2 m_0 while the remaining third have a mass of 0.9 m_0. If a compressive stress is exerted along the same crystal axis, the energy of the valleys is reduced so that electrons "overflow" into states of higher mass. Their acceleration thereby decreases, and hence the electrical resistivity increases. These effects find a practical application first and foremost in strain gauges, although the use of brittle semiconductor wafers instead of ductile metal wires presents a number of technological problems. The pressure dependence of the voltage-current characteristic of semiconductor diodes, which is based on similar effects, allows the construction of highly sensitive miniature microphones [4].

For the purposes of fundamental research the measurement of the "piezo-resistance" effect described enables useful predictions to be made regarding the structure of the energy states of the charge-carriers (band structure), especially on their anisotropy. The piezo-resistance effect is also influenced by those processes which slow down the charge-carriers accelerated by the external field (scattering processes characterized by relaxation times) and which therefore partly determine the electrical resistivity. Measurements of this nature [5] made it possible, for example, to decide experimentally between two different hypothetical models of the electron transition between two sublevels under the influence of lattice vibrations ("lattice intervalley scattering").

If, instead of a DC electric field, an alternating field of sufficiently high frequency is used, the influence of the scattering processes can be neglected when the electron undergoes numerous oscillations between two collisions. The quantity to be measured then is not the conductivity but the susceptibility. Since the free movements of the charge-carriers between two collisions take place in a time of the order of 10^{-13} s, the above-mentioned condition implies the use of optical frequencies, and hence involves investigating the refractive index under the influence of elastic stresses. These photoelastic effects, as they are sometimes called, show apart from the contribution of the 10^{23} cm^{-3} bound electrons, a contribution from the conduction electrons which, compared with the valence electrons, possess a polarizability (in the frequency range of interest) which is about 10^4 times greater. These experiments [6] yield particularly useful information on the connection between elastic stresses and the shift of valley energies ("deformation potentials").

In a magnetic field the charged particles describe, in the simplest case, a helical path about the direction of the field. Their angular velocity depends only on the magnetic field strength and on their mass. Projected on to an axis perpendicular to the magnetic field their motion shows a harmonic variation, rather resembling the swing of a pendulum. When an alternating electric field is applied, energy can thus be imparted to this motion, which shows a resonance effect when its frequency is near what is called the "cyclotron resonance frequency". At resonance the electron system absorbs a maximum of energy from the field. Provided the disturbance caused by scattering processes is weak enough, an absorption line is thus observed as the magnetic field is varied (at constant frequency), the position of the absorption line being directly related to the mass of the particles.

"Cyclotron resonance" provides at the present time the most accurate and detailed method of determining the effective mass of charge carriers [7], although the condition of a sufficiently long mean free path limits its use to extremely pure crystals and low temperatures. In this way, for example, the masses of the conduction electrons in germanium and silicon, including their anisotropy, have been determined accurately to within a few per cent. On the other hand, in the same substances the masses of the defect electrons in the valence band could only be found very approximately. Even though the conditions were chosen so as to minimize "collision broadening", the lines remained weak and blurred, particularly in silicon. The physical causes are to be found in the fact that the valence band consists of closely adjacent energy levels which are so deformed by mutual interaction that various charge carriers in the same band correspond to different resonant frequencies. A way out of this dilemma was found by Hensel and Feher [8], who performed cyclotron resonance measurements on elastically deformed silicon. The deformation has the effect of separating the originally adjacent bands far enough for the interaction between them to become negligible. The result is a clearly distinct resonance line, similar to that of the conduction electrons. By theoretical studies along the same lines [9] the whole line pattern in dependence on the elastic stress has been evaluated in terms of the deformation potentials and the coupling parameters between the bands.

Heat is conducted through a crystal — from the microscopic point of view — because more thermal lattice vibrations pass from the point of higher temperature to the lower than vice versa. The energy flow involved is directly related to the observed flow of heat. Thermal conductivity is a measure of the extent to which lattice waves are scattered by imperfections in the crystal. In germanium, for example, some 2×10^{16} cm^{-3} atoms of antimony or arsenic, were incorporated in the lattice as donor impurities [10]. At a temperature of 2.5 °K the As addition was found to have reduced the thermal conductivity by roughly a half, whereas the doping with Sb reduced it to one-sixth. This difference could at first be explained by the interpretation that the scattering stems from electrons bound to the foreign atoms in the ground state. The scattering probability, which depends on the difference in energy (or chemical shift) between singlet and triplet states, was thought to be greater for antimony. Clear confirmation of the scattering model had not been forthcoming, however, until the application of static elastic deformation. In view of the structure of the donor ground state, heat conduction in a substance doped with antimony should theoretically show marked dependence on the elastic deformation, but only if the deformation were applied along a body diagonal of the cubic crystal. For arsenic the effects should be an order of magnitude smaller, and the same applies to antimony in the case of deformation along the cube edge. The experiments in fact demonstrated that in the first case the thermal conductivity rose by a factor of 3 under a tensile stress of 100 kg/cm^2, whereas in all other cases changes of only about 10% were measured. If the doping with antimony atoms was increased to 10^{17} cm^{-3}, this piezothermal effect disappeared. This is a further argument in support of scattering from localized electrons, for it is known from other investigations that at these densities of impurities the individual donor ground states merge to form a single impurity band.

Sound propagation in solids

A simple example of a dynamic elastic deformation is a longitudinal sound wave, whereby a sequence of

[1] J. F. Gibbons and H. S. Horn, Digest of tech. papers, Intern. Solid State Circuits Conf., Univ. of Pennsylvania, Philadelphia, 1963.

[2] M. A. C. S. Brown and M. F. Kimmitt, Brit. Comm. and Electronics **10**, 608, 1963.

[3] R. W. Keyes, Solid State Physics **11**, 149, 1960.

[4] W. Rindner, J. appl. Phys. **33**, 2479, 1962.

[5] J. E. Aubrey, W. Gubler, T. Henningsen and S. H. Koenig, Phys. Rev. **130**, 1667, 1963.

[6] K. J. Schmidt-Tiedemann, Phys. Rev. Letters **7**, 372, 1961 B. G. Weigel, Philips: Unsere Forschung in Deutschland, pp. 37-39, Philips ZL GmbH, Aachen-Hamburg 1964.

[7] B. Lax and J. G. Mavroides, Solid State Physics **11**, 261, 1960.

[8] J. C. Hensel and G. Feher, Phys. Rev. **129**, 1041, 1963

[9] H. Hasegawa, Phys. Rev. **129**, 1029, 1963 (No. 3).

[10] R. J. Sladek, Rep. Internat. Conf. on the physics of semiconductors, Exeter 1962, p. 35.

regions of compression and expansion pass through the crystal with the speed of sound. In accordance with the picture discussed above, where the electron energy is influenced by mechanical deformation, the potential energy is influenced by mechanical deformation, the energy of a conduction electron is also seen to have a wave form in space and to move with the speed of sound. Let us imagine, for example, a corrugated washing board. Since the electrons try to find or maintain the energetically most favourable situation in the "valleys" of the washing board, some of them are carried along by the sound wave (wave-particle drag). The result is an "acousto-electric current". This "ideal" electron distribution is continuously disturbed by collisions which tend to destroy the drift velocity of the electrons. This has the effect of damping the sound wave. A primary condition here is that the mean free path of the electrons should be small compared with the wavelength of the sound, so that within every valley a thermal distribution can form. If we now apply in the direction of the wave propagation a DC electric field which imparts to the electrons a drift velocity equal to the speed of sound, this damping effect disappears. The electrons, by virtue of the energy supplied by the electric field, can now travel along the valleys without any additional acceleration from the sound field. If the electric field is further increased, so that the drift velocity exceeds the speed of sound, an extraordinarily interesting effect is observed. The electric field now gives up energy to the sound field, in other words the sound amplitude is increased. Using this principle Hutson, McFee and White [11] achieved sound amplifications in cadmium sulphide up to 54 dB/cm at a frequency of 45 Mc/s. Various effects found in bismuth (e.g. the anomalous magnetoresistance effect, Esaki [12]) are also presumably attributable to strong acousto-electric interactions.

Among the semiconducting substances lead-telluride, which has recently been extensively studied, lends itself particularly well to such investigations in view of its high charge-carrier mobility.

The cyclotron resonance method mentioned above finds only limited application to metals. The skin effect prevents the necessary high-frequency field from penetrating any deeper than about 10^{-5} cm into the crystal, and resonance phenomena can only be observed in a magnetic field oriented parallel to the surface of the crystal (Azbel-Kaner effect) [7]. Since, on the other hand, a sound wave travelling through a crystal is accompanied by an alternating field, this can be utilized for the excitation of cyclotron resonance lines. This is referred to as "magneto-acoustic resonance". The most easily observable variant of these phenomena is "spatial cyclotron resonance" which occurs when the diameter of the electron path is roughly equal to the sound wavelength.

Lattice vibrations

If the crystal lattice is traversed by a wave train which has such a high frequency that its wavelength is small compared with the mean free path of the electrons, the electron distribution cannot relax to thermal equilibrium. The energy of every electron is slightly increased in, for example, the compression region and slightly decreased in the expansion region, and it is assumed that after the passage of a finite wave train the energy state of the electrons is the same as before.

This model, according to which the energy of the electrons follows adiabatically the fluctuations of the lattice constants, is only correct within certain limitations. It happens — more or less frequently — that the electron exchanges energy with the lattice vibration, and in doing so changes its state. These changes of state constitute the "scattering processes" responsible for electrical resistivity; in this case the term used is "lattice scattering". In the quantum-mechanical interpretation this exchange of energy cannot take place continuously but only in discrete jumps with a quantum of energy $h\nu$, where h is Planck's constant and ν is the lattice vibration frequency. These energy quanta, as quasi-particles called "phonons", are a convenient concept for lattice waves or vibrations, and often characterize them more appropriately than the relative positions of the atoms in the crystal. An electron "emits a phonon" when it gives up part of its energy to a lattice vibration and thus increases the latter's amplitude. In the converse case the electron absorbs a phonon.

The theory of phonons and their interaction with each other and with charge carriers, is one of the main pillars of the physics of transport phenomena in semiconductors, and is therefore one of the most investigated fields of research. We shall refer here to only two results. With the aid of neutron scattering, Brockhouse and co-workers were able to determine the phonon spectra of the principal semiconductors germanium and silicon [13], that is to say the relation between the phonon energy and the wavelength of the given lattice vibration. (This method is not limited to semiconductors, however [14].) The observation of Shockley *et al.* [15] that the current in a germanium crystal saturates with increasing voltage, and the marked anisotropy of the conductivity in high electric fields [16] could be interpreted quantitatively by three different kinds of electron-phonon interaction processes (acoustic, optical and "intervalley" phonons). These "hot" electrons [17] exchange energy with the crystal lattice mainly through the emission of what are termed optical phonons, whose energy at a given lattice vibration wavelength is several hundred times greater than that of the comparable "acoustic" phonons, which correspond to ordinary sound waves. The theoretical analysis of the experiments described yielded numerical values for the coupling constant between electrons and optical phonons, a quantity which had previously been very difficult to determine.

Polarons

The model of an essentially free conduction electron which only now and then is scattered by a phonon, is no longer satisfactory in semiconductors where there is strong electron-phonon interaction. Here the electron continuously emits and absorbs phonons which cover it like a mantle on its way through the crystal (dressed particle). The new quasi-particle (electron plus surrounding phonons) is known as a "polaron". Whereas the case of weak electron-phonon interaction can be reduced by quantum-mechanical perturbation theory to a one-body problem, the polaron represents a true many-body problem, the solution of which calls for field theory methods [18]. In classical terms one can imagine that the electron, by electrostriction due to its electric field, elastically deforms its environment in the crystal. The electron together with the state of strain it produces in the crystal, forms an entity — the polaron — which can only be moved as a whole in any transport of charge. Research work on this transport mechanism, which has probably been established in the silver halides [19], is at present too much in a state of flux to allow any brief summary of results.

The opposite extreme to the hypothesis of the "free" conduction electron is found when the electron is nearly always localized at a particular lattice site. The impact of one or more phonons can give a weakly bound electron sufficient energy to jump to the next lattice site. This kind of electrical conductivity, described as a "hopping" mechanism or "small polaron", occurs for example in oxides of the transition metals which have been appropriately doped with foreign atoms of a different valence. At low temperatures the mobile charge carriers collect around the foreign atoms and form rotatable mechanical dipoles, for here again the charge carrier is accompanied by an elastically deformed zone. These dipoles try to orient themselves in the field of a sound wave and thus cause a certain damping, which has a pronounced maximum when the frequency of the sound wave corresponds to the hopping frequency. From damping measurements of this kind performed on nickel oxide doped with lithium, Van Houten [20] determined the activation energy and the frequency constant of these processes.

Superconductivity

One of the most remarkable transport phenomena is superconductivity, i.e. the property, shown by many metals, of becoming exceptionally conductive below a certain "transition temperature", at which the electrical resistance falls suddenly to zero. Although from the physicist's point of view more fundamental importance seems to attach to the absolute diamagnetism (Meissner effect) which occurs at the same time (the lines of magnetic induction being expelled from inside the superconductor at this temperature) the existence of "persistent currents", bound up with the disappearance of resistance, is of considerable interest for technical applications. A current induced in a toroid of these materials continues to circulate indefinitely after the magnetic field has been removed. None of the measurements so far carried out [21] give any indication of the decrease of such currents with time. Since the resistance is probably exactly zero, these materials can make ideal devices for the conduction of electrical energy, for building electromagnets, etc. Of course the applications are severely limited by the low transition temperatures of the 450 or more superconductors known at the present time [22]. Niobium-tin has the highest transition temperature yet measured, i.e. 18 °K or about —255 °C. A satisfactory explanation of this effect, discovered in 1911 by Kamerlingh Onnes [23], was given in 1957 by Bardeen, Cooper and Schrieffer [24], who built up their theory on the basis of a whole series of partial solutions. After some apparent digression, their explanation leads us back to our broad theme of the elastic deformations of the crystal lattice. Superconductivity is shown to be the consequence of an attractive force between the electrons which exceeds the normal Coulomb repulsion (between like charges). This attractive force has its origin in the exchange of pho-

[11] A. R. Hutson, J. H. McFee and D. L. White, Phys. Rev. Letters **7**, 237, 1961.
[12] L. Esaki, Phys. Rev. Letters **8**, 4, 1962.
[13] B. N. Brockhouse and P. K. Iyengar, Phys. Rev. **111**, 747, 1958, and B. N. Brockhouse, Phys.Rev. Letters **2**, 256, 1959.
[14] A. D. B. Woods, B. N. Brockhouse, R. A. Cowley and W. Cochran, Phys. Rev. **131**, 1025, 1963.
[15] E. J. Ryder and W. Shockley, Phys. Rev. **81**, 139, 1951.
[16] W. Sasaki, M. Shibuya and K. Mizuguchi, J. Phys. Soc. Japan **13**, 456, 1958.
[17] H. G. Reik, Festkörperprobleme **1**, 89, 1962, and K. J. Schmidt-Tiedemann, Festkörperprobleme **1**, 122, 1962.
[18] R. P. Feynman, R. W. Hellwarth, C. K. Iddings and P. M. Platzman, Phys. Rev. **127**, 1004, 1962.
[19] F. Garcia-Moliner, Phys. Rev. **130**, 2290, 1963.
[20] S. van Houten, Rep. Internat. Conf. on the physics of semiconductors, Exeter 1962, p. 197.
[21] B. Serin, Hdb. d. Physik, Vol. 15, p. 210, Springer, Berlin 1956.
[22] American Institute of Physics Handbook, 2nd ed., pp. 9-112 ff., McGraw-Hill, New York 1963.
[23] H. Kamerlingh Onnes, Comm. Phys. Lab. Leiden, No. 119, 120, 122, 191.1
[24] J. Bardeen, L. N. Cooper and J. R. Schrieffer, Phys. Rev. **108**, 1175, 1957.

nons between the electrons. It had already been proposed by Fröhlich [25] that lattice vibrations might be involved in the production of superconductivity, and the observed dependence of the transition temperature on the atomic mass (isotope effect) seemed to offer experimental confirmation [26]. The simple hypothesis — that an electron emits a phonon, which is again absorbed by another electron — must nevertheless be modified in two respects. First, the force is attractive only in the case of procesess where the energy exchanged between the electrons is smaller than the energy of the activating phonon. In classical theory, then, the law of the conservation of energy would completely rule out the emission of such a phonon. Quantum mechanical theory states that for a transition consisting of two steps the energy equation need only be satisfied in respect of the whole process but not for each of the two steps, always provided they follow each other fast enough (to be more exact, are mutually coherent). The theory therefore speaks of the exchange of a "virtual" quantum or phonon. Secondly, it is necessary to treat these electrons as a sort of combined state, a "Cooper pair" as they are called. It was only by thus extending the idea of Fröhlich that Bardeen and co-workers were able to bring the theory into agreement with experimental observation.

There is another reason why superconductivity fascinates the physicist. One ordinarily expects that the behaviour of a macroscopic body will obey the laws of classical physics, because the quantum-mechanical properties should statistically be averaged out in view of the large number of atoms. A superconducting ring, however, represents a quantum state of macroscopic dimensions and constitutes — apart from superfluid helium — the only known configuration possessing this property. Direct experimental proof of this was recently given by Deaver and Fairbank [27] and independently by Doll and Näbauer [28] by measuring the quantum jumps of the magnetic flux in a superconducting ring. The magnitude of the flux quantum shows that the superconducting "charge carriers" are to be interpreted as electron pairs and not as single electrons [29]. A system of superconducting rings with frozen-in magnetic flux can be employed as a digital storage element [30] in which — unlike other systems based on superconductivity — the stored information is not destroyed if the power is temporarily switched off (cold memory).

[25] H. Fröhlich, Phys. Rev. 79, 845, 1950.

[26] E. Maxwell, Phys. Rev. 78, 477, 1950, also C. A. Reynolds, B. Serin, W. H. Wright and L. B. Nesbitt, Phys. Rev. 78, 487, 1950.

[27] B. S. Deaver, Jr., and W. M. Fairbank, Phys. Rev. Letters 7, 43, 1961.

[28] R. Doll and M. Näbauer, Phys. Rev. Letters 7, 51, 1961.

[29] N. Byers and C. N. Yang, Phys. Rev. Letters 7, 46, 1961, also L. Onsager, Phys. Rev. Letters 7, 50, 1961.

[30] D. J. Dumin and J. F. Gibbons, J. appl. Phys. 34, 1566, 1963.

Summary. The author gives a survey of those transport processes in semiconductors (electrical and thermal conductivity) in which elastic deformations of the solid by mechanical means (tensile, compressive or shear stresses) play an essential part. It is shown that the piezo-resistance effect, the photoelastic effects and special applications of "cyclotron resonance" all involve the interaction between conduction electrons and crystal lattice, whereas heat conduction involves the scattering of lattice waves from crystal imperfections. Acousto-electric interactions and magneto-acoustic resonance are discussed. After explaining the concepts phonon and polaron, the author concludes with a discussion of superconductivity, showing that this also can be interpreted in terms of phonon exchange between electron pairs, agreeing with the general picture of elastic deformations of the crystal lattice.

Thermal insulation of sodium lamps

R. Groth and E. Kauer

536.21:621.327.532

Introduction

In spite of the considerable progress made in electric lamps and lighting during recent years, conventional light sources convert only a small fraction of their energy input into light. The rest is dissipated as heat. The energy efficiency is thus small: about 2% for an incandescent lamp and about 15% for the sodium lamp — the most efficient of all conventional sources. Many attempts have been made to reduce these undesirable thermal losses. One of the possibilities will be discussed — the use of a selective-reflecting coating on the envelope of a sodium lamp.

The idea of regulating the spectral distribution of the radiation from a source by means of suitable reflection-filters is by no means new. In a number of early patents, the idea usually takes the form of a thin metal coating, preferably gold or copper, on the glass envelope of the lamp. The coating has to have the property of reflecting the infra-red radiation back to the incandescent filament while remaining virtually transparent to visible light. For the same filament temperature, the power consumption will then be less and the efficiency higher. The term efficiency will be used in the following to mean the lumen efficiency, that is, the ratio of the total light flux (in lumens) to the power consumption (in watts).

It is perhaps at first sight not entirely evident that an infra-red-reflecting filter can also improve the efficiency of a sodium lamp; the latter is, strictly speaking, not a thermal radiator in the usual sense but a low-pressure discharge lamp. For a proper understanding of the effect of such filters it is necessary to look a little further into the physical principles of the sodium lamp.

The energy balance of a sodium lamp

A 140 W sodium lamp has a U-shaped discharge tube mounted in a cylindrical glass envelope. The discharge tube has an external diameter of 17 mm and a total length of 850 mm (*fig. 1*). It is filled with neon plus a little argon — total pressure about 10 torr — and further contains a quantity of free sodium. When the lamp is cold the sodium vapour pressure is very low so that, after switching on, the lamp first emits chiefly the well known red spectrum of neon. As the lamp heats up more of the sodium evaporates until the pressure of sodium vapour is equal to the saturation vapour pressure corresponding to the temperature of the tube. The lamp radiates most light when the partial pressure of sodium is some thousandths of a torr (i.e. some microns Hg); this corresponds to a discharge tube temperature (at the coldest place) of about 270 °C. If the temperature is raised further by increasing the discharge current, the light flux diminishes because of a greater self-absorption due to the increased sodium vapour pressure. The optimum working temperature will clearly depend on the design of the lamp.

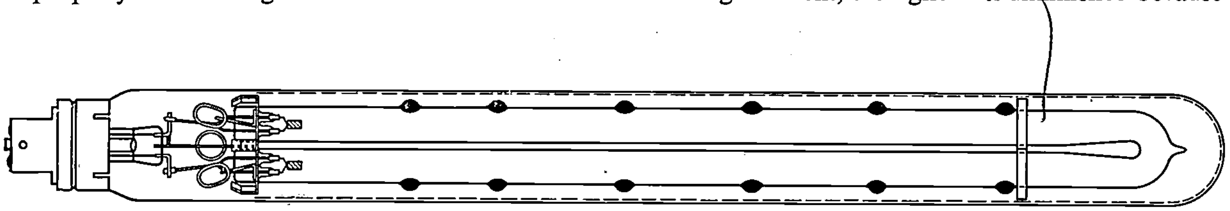


Fig. 1. Schematic diagram of a modern sodium lamp. The black spots in the discharge tube represent condensed droplets of free sodium. The dotted line on the inside of the envelope is the reflection filter that reflects back the infra-red radiation from the discharge tube.

When thermal equilibrium has been set up the rate at which electric energy is fed to the lamp must clearly be equal to the power dissipated as radiation and via conduction and convection. In modern types of sodium lamps the space between the discharge tube and the envelope is evacuated; this eliminates convection losses and greatly reduces conduction losses. The losses in the form of thermal radiation (the discharge tube is at 270 °C) are not reduced however. The tube radiates almost as a black body and for a 140 W sodium lamp about 100 W is emitted as radiation.

Fig. 2 shows the spectral distribution of the energy radiated from the discharge tube (T = 540 °K) into surroundings at a temperature of 300 °K. The *continuum* is radiated exclusively by the surface of the

Dr. R. Groth and Dr. E. Kauer (deputy director) are research workers at the Aachen laboratory of Philips Zentrallaboratorium GmbH.

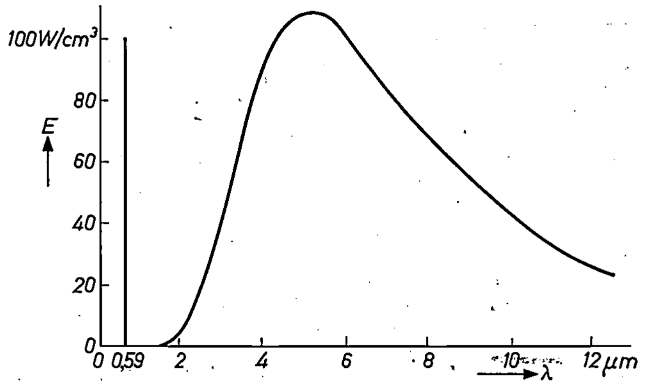


Fig. 2. Spectral distribution $E(\lambda)$ of the energy radiated by a black body at 540 °K into surroundings at 300 °K. The discharge tube of a sodium lamp radiates approximately thesamespectrum. The position of the sodium D lines is indicated by a line at 0.59 μm (neither width nor height of this line is intended as a measure of the intensities of the D-lines).

discharge tube. The maximum lies in the infra-red at a wavelength of about 5.5 μm.

To reduce the radiation losses in the infra-red we have the following possibilities:

1) The discharge tube is surrounded by a filter that absorbs infra-red radiation but which is transparent to sodium light. This method has been described by Van de Weijer [1].

2) The discharge tube is surrounded by a filter that reflects infra-red radiation but is again transparent to sodium light. Such a filter would be ideal if it passed the sodium light unattenuated and reflected all the infra-red of wavelength greater than about 4 μm. A filter with these properties would reflect the thermal radiation back on to the discharge tube without itself showing any rise in temperature. Such a filter may be coated on the interior surface of the outer envelope, or on the discharge tube.

Heat-reflection filters for sodium lamps

No filters exist that entirely satisfy the above-mentioned requirements. A high transparency for sodium light was first attained with the development of interference filters, built up from alternating dielectric multilayers. These have the disadvantage, however, that they reflect well in the infra-red only over a relatively small wavelength range. Moreover lamps with such filters would be too expensive for normal use.

A reasonable approximation to the required properties can also be achieved with a filter consisting of a single thin metallic layer. Copper and gold are both suitable in this respect. As an example *fig. 3* shows the transmission and reflection of a number of thin gold films on glass as a function of the wavelength. It is seen that the reflection coefficient R approaches 90% even for the near infra-red, whilst the transmission D for the sodium D lines remains larger than 50%.

Evidently there exists an optimum layer thickness: if too thick, the layer will transmit only a little light while if too thin the layer, though absorbing little light, will not reflect well in the infra-red. In order to estimate the improvement in lamp efficiency to be expected from the layer it is necessary to know the spectral transmission and reflection curves for layers of various thicknesses. In principle it is possible to calculate these curves from the optical constants of the material, viz the refractive index n and the extinction coefficient K [2]. The calculations are unreliable, however, because for thin layers of less than 20-30 nm (200-300 Å) the values of the optical constants may differ from those for the bulk material. For the estimation of the change in lamp efficiency it is therefore necessary to use experimental data instead of the calculated curves. The optimum thickness for a gold layer is then estimated at about 15 nm. Measurements on a series of test lamps coated with gold films of various thicknesses have confirmed this estimate. A film thickness of 16 nm increased the lamp efficiency from 100 lm/W to 130 lm/W (these figures refer to a 140 W sodium lamp). A disadvantage was that the total light flux from the lamp decreased by about 30%. With the temperature of the lamp held the same, the decrease in light flux is due to the smaller number of excited atoms per unit volume (less current must be fed to the lamp) and to the reflection and absorption of visible light in the gold film (useful light is thus converted into heat).

To determine the efficiency that would be attained with an *ideal* filter and the discharge tubes at present in use, no reflection layers were applied but instead the discharge tube was heated by thermal radiation from outside. The effect, as far as the discharge tube is concerned, is identical to that of a reflection filter; moreover, the effective reflection coefficient is readily variable. The method has the advantage that all the measurements can be performed on the one discharge tube.

The results of this investigation are summarized in *fig. 4*. The efficiency is plotted as a function of the power consumption of the discharge. Each curve refers

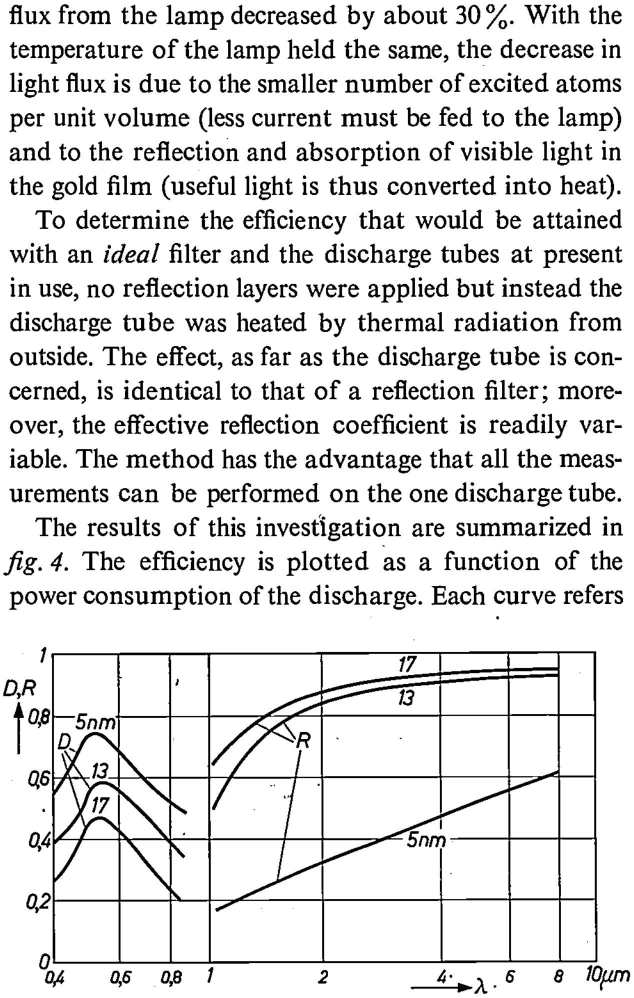


Fig. 3. Transmission D and reflection R of gold films of various thicknesses as functions of the wavelength λ.

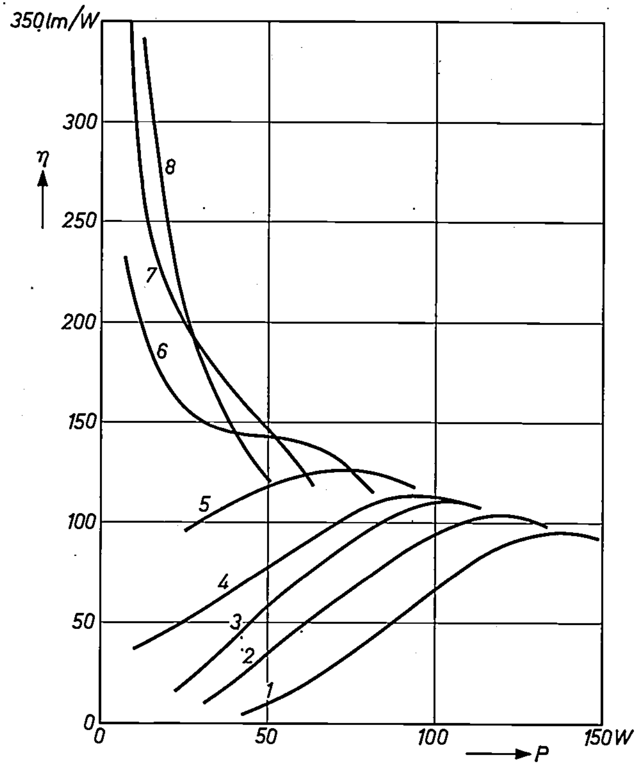


Fig. 4. The efficiency η of the discharge tube of a 140 W sodium lamp as a function of the power consumption P for various values of the thermal insulation. The insulation increases from curve *1* to curve *8*.

to a given effective heat insulation, i.e. to a constant flux of heat radiation from outside. The latter increases with increasing curve number; curve *1* refers to the case when additional heating from outside is absent. No attempt has been made to correlate each effective insulation with a reflection coefficient, because the conduction heat losses, which are *not* affected by a reflection filter, are not accurately known. The measurements, however, cover a range that includes all practicable values of the effective heat insulation; curve *7*, for example, represents the case of a gold film 16 nm thick. From fig. 4 it may be concluded that with a sufficient heat insulation, sodium lamps should be able to reach an efficiency of more than 300 lm/W. At the same time, it is evident that this rise in efficiency is necessarily accompanied by a drop in the light flux. This remains true even when the filter is ideal, i.e. transmits the sodium light unattenuated. If, with the higher efficiency, it is required that the light flux does not change substantially, then the conditions governing the discharge have to be re-adjusted to match the improved heat insulation.

Since a heat radiation filter inevitably involves an attenuation of the visible light emitted it is important to try and reduce such attenuation to the minimum. With filters of gold or other metals the reflection coefficient for sodium light can be greatly reduced by coating the metal film with a dielectric layer. This measure is analogous to the well-known "blooming" of the surfaces of optical components to reduce reflections.

As an example, *fig. 5* shows the transmission and reflection of a 13 nm thick gold film with and without an anti-reflection layer — in this case a ZnS film about 32 nm thick. This dielectric layer increases the transmission of sodium light from 57% to 77% without reducing the reflection coefficient in the relevant infrared region. Such improved gold filters may be expected to improve lamp efficiency by 30%.

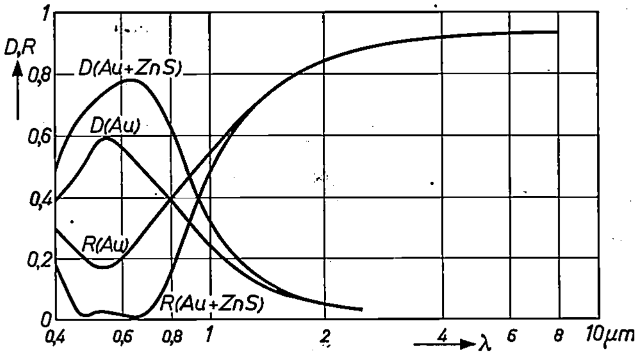


Fig. 5. Spectral transmission D and reflection R of a 13 nm thick gold film. Au: gold film alone. Au + ZnS: gold film coated with anti-reflection layer of ZnS.

Further details will not be discussed here except to mention that to minimize reflection the refractive index of the dielectric coating has to be higher as the reflection coefficient of the metal film increases [2]. Substances suitable for anti-reflection coatings are those with high refractive index such as ZnS ($n = 2.3$), $PbCl_2$ ($n = 2.15$), Bi_2O_3 ($n = 1.91$), etc.

Metals other than gold can be used as the basis of reflection filters. *Fig. 6* shows the optical properties of a

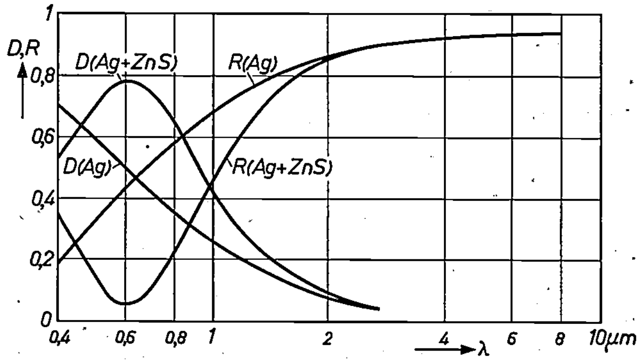


Fig. 6. Spectral transmission D and reflection R of a silver film, c. 10 nm thick. Ag: silver film alone. Ag+ZnS: silver film coated with ZnS anti-reflection layer 0.11 μm thick.

[1] M. H. A. van de Weijer, Philips tech. Rev. 23, 246, 1961/62.

[2] See e.g. H. Mayer, Physik dünner Schichten, Teil I, Stuttgart 1950; or O. S. Heavens, Optical properties of thin solid films, Butterworth, London 1955.

silver film with and without an anti-reflection coating of zinc sulphide. The anti-reflection coating increases the transmission for sodium light from 50% to 76%. As in the case of the gold film, the infra-red reflection is left unimpaired.

Although the transmission of thin metal films with anti-reflection coatings may be as high as 80%, the question remains whether it is possible to make filters that do not absorb at all but still have a high reflection coefficient in the infra-red. Since the high infra-red reflection of metals is a consequence of their high electric conductivity, i.e. of the presence of free charge carriers, it is a natural step to consider whether strongly-doped semi-conductors might not also be used.

Free charge carriers — such as electrons in semiconductors — can have a marked effect on the optical constants of a medium. Suppose that a medium has a dielectric constant ε_g in the *absence* of free charge carriers. It is assumed that, in the frequency range under consideration, ε_g is real so that in this range there is no absorption (other than that now to be described). If we now bring free charge carriers into the material by suitable doping then, after Drude[3] the following dispersion relations apply:

$$n^2 - K^2 = \varepsilon_g - \frac{Ne^2}{\varepsilon_0 m^*(\omega^2 + \gamma^2)},$$

$$2nK = \frac{\gamma Ne^2}{\varepsilon_0 m^* \omega(\omega^2 + \gamma^2)}.$$

The value of damping factor γ in these relations depends on the D.C. mobility μ of the charge carriers: $\gamma = e/\mu m^*$. The quantity N is the concentration of free charge carriers, e is the charge on the electron, m^* the effective mass of the charge carriers and ε_0 the dielectric constant of free space.

The two expressions may be re-written, introducing the so-called plasma frequency

$$\omega_p = \left(\frac{Ne^2}{\varepsilon_0 \varepsilon_g m^*} - \gamma^2\right)^{\frac{1}{2}}.$$

This is defined by putting $n^2 - K^2 = 0$ in the first of the above dispersion relations. The formulae are then much simplified, with ω/ω_p as the independent variable and γ/ω_p as material parameter:

$$n^2 - K^2 = \varepsilon_g \left[1 - \frac{1 + \left(\frac{\gamma}{\omega_p}\right)^2}{\left(\frac{\omega}{\omega_p}\right)^2 + \left(\frac{\gamma}{\omega_p}\right)^2}\right],$$

$$2nK = \varepsilon_g \frac{\frac{\gamma}{\omega_p}\left[1 + \left(\frac{\gamma}{\omega_p}\right)^2\right]}{\frac{\omega}{\omega_p}\left[\left(\frac{\omega}{\omega_p}\right)^2 + \left(\frac{\gamma}{\omega_p}\right)^2\right]}.$$

From these expressions the values of n and K can be calculated and then the reflection coefficient R can be calculated from

$$R = [(n-1)^2 + K^2]/[(n+1)^2 + K^2].$$

The results of such calculations are plotted in *fig. 7*. It is seen that the steepness of cut-off depends on the ratio γ/ω_p. To obtain a filter with a sufficiently sharp cut-off the ratio γ/ω_p should be as small as possible.

The position of the cut-off wavelength is mainly determined by the plasma frequency ω_p, provided the film thickness has been correctly chosen

For sodium lamps the cut-off is required to lie between 1 and 3 μm so that the required ω_p is substantially fixed. Hence to make γ/ω_p small, γ itself must be small. This means that we require a medium in which the free charge carriers have the largest possible product of mobility and effective mass. *Fig. 8* shows the calculated transmission and reflection curves of an imaginary filter with the following not unrealistic values for the material parameters: $\varepsilon_g = 10$, $N = 1.13 \times 10^{22}$ cm^{-3}, $\mu = 100$ cm^2/Vs, $m^* = 0.5\ m_0$ (m_0 is the rest mass of the electron) and $d = 0.03$ µm. It is seen that for the right film thickness an adequately steep cut-off is possible between the visible and infrared.

For a semi-conductor to be transparent to sodium light, its band gap must be greater than 2.2 eV. Many semiconductors satisfy this condition.

In addition we require a high concentration of free charge carriers: N must be at least 10^{20} cm^{-3}. It is not so easy to achieve this concentration because most semiconductors cannot be doped so strongly. There are, however, a number of oxides such as CdO, Tl_2O_3, In_2O_3 and SnO_2 in which concentrations higher than 10^{20} cm^{-3} can be achieved.

Of these oxides, tin oxide (SnO_2) in the form of a thin film on glass has already found a number of applications. When correctly prepared SnO_2 films have a good electric conductivity and are very transparent to visible light. They are particularly useful as transparent electrodes for electroluminescent devices and as transparent electric heater coatings on glass.

Tin oxide films are usually applied by spraying. A mixture of $SnCl_4$ with an organic solvent such as ethanol or butyl acetate is sprayed from a nozzle on to the object, the latter being heated to a certain temperature in an oven. If the temperature is high enough the $SnCl_4$ is transformed into SnO_2 that forms as a thin layer on the glass plate or other substrate. A very high conductivity film is obtained when some $SbCl_3$ [4] or HF [5] is added to the above mixture. Such films have a high infrared reflection whilst retaining a high transparency to visible light [6]. Measurements on such

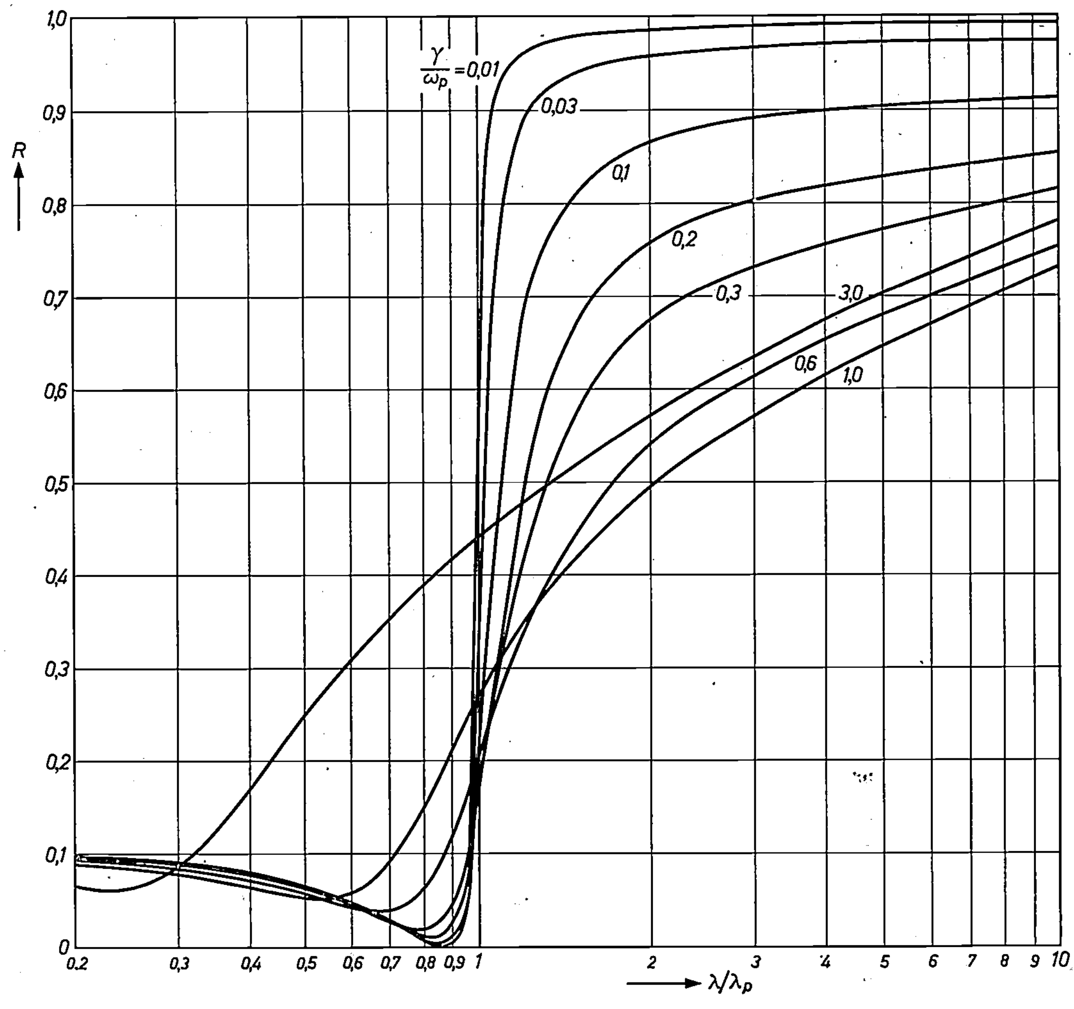


Fig. 7. Reflection coefficient for a medium containing free charge carriers. The material is defined by the relative dielectric constant of the medium ε_g = 4, and by the plasma wavelength $\lambda_p = 2\pi c/\omega_p$ (for the definition of ω_p, the plasma frequency, see text).

films have shown that for $\lambda > 4$ μm (where the reflection is high) calculations of the reflection from the electric properties of the film on the basis of Drude's theory are a very good approximation [7]. The highest concentration of free charge carriers was obtained by doping the spray with fluorine. It would appear that the limiting solubility of fluorine in the SnO_2 lattice is reached at a concentration of 6×10^{20} cm^{-3}. This concentration is however high enough to give a very satisfactory cut-off between sodium light and the infra-red. It is of course necessary that the charge carriers are sufficiently mobile. In contrast to the effective mass, which depends effectively only on the material of the film, the mobility may be strongly dependent on the microstructure of the film. In the literature the mobil-

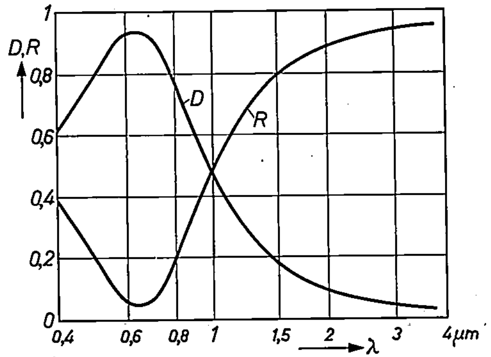


Fig. 8. Spectral transmission D and reflection R of a hypothetical filter with the following material parameters: ε_g = 10; N = 1.13×10^{22} cm^{-3}; μ = 100 cm^2/Vs; m^* = 0.5 m_0; d = 0.03 μm.

[3] P. Drude, Phys. Z. **1**, 161, 1900.

[4] J. M. Mochel, U.S. Patent No. 2564707.

[5] W. O. Lytle and A. E. Jonge, U.S. Patent No. 2566346.

[6] V. K. Miloslavskii and S. P. Lyashenko, Optics and Spectroscopy **8**, 455, 1960.

[7] R. Groth, E. Kauer and P. C. van der Linden, Z. Naturf. **17a**, 789, 1962.

ity in sprayed layers of tin oxide is quoted as lying between 3 and 15 cm²/Vs [6][8][9]. A further investigation showed that the properties of these films are markedly dependent on the temperature at which the mixture is sprayed. Reproducible results were obtained when care was taken that the spray had the same temperature as the substrate. With a spray system in which the spray mixture was pre-heated as it traversed a long oven, it was possible to investigate the effect of the spraying temperature on the properties of the film.

Fig. 9 gives the result of this investigation and shows how the concentration N of the free charge carriers and their mobility μ (at room temperature) vary as functions of the spraying temperature T_S. The concentration N varies only little with T_S but the mobility increases considerably as T_S is increased, reaching a value of 20 cm²/Vs for T_S = 475 °C. As was to be expected the infra-red reflection was found to increase as the mobility increased, reaching a value of 80%. The highest possible spraying temperature should therefore be used. The temperature must, however, remain somewhat below 500 °C as otherwise the film becomes milky as a result of an undesirable reaction between the spray liquid and the glass substrate.

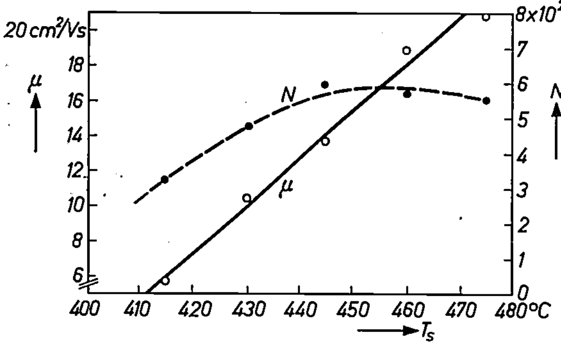


Fig. 9. Concentration N and mobility μ (at room temperature) o charge carriers in SnO₂ films on glass as functions of the spraying temperature T_S.

Furthermore to obtain an optimum filter the film thickness is very important. The film should be so thick that the infra-red reflection of the bulk material is closely approached. To estimate the minimum film thickness required, the reflection coefficient for λ = 5.5 μm has been calculated as a function of the film thickness. For this purpose we assume a concentration of free charge carriers and a mobility as obtained for the most favourable spraying temperature (N = 6 × 10²⁰ cm⁻³, μ = 20 cm²/Vs).

The results (*fig. 10*) show that the reflection increases with increasing film thickness, approaching the bulk value for a thickness of 0.3 μm. For this order of thickness, the transmission in the visible region is largely determined by interference. The film thickness must therefore further be chosen so that it corresponds to the wavelength of the sodium D lines. In fig. 10 the thicknesses that satisfy this condition are marked by circles. The thicknesses 0.16, 0.32 and 0.48 μm refer respectively to interference of the zeroth, first and second orders. A film thickness of 0.32 μm turns out to be the most effective; a thicker layer hardly gives any improvement whilst the zero-order thickness gives a 10% smaller infra-red reflection.

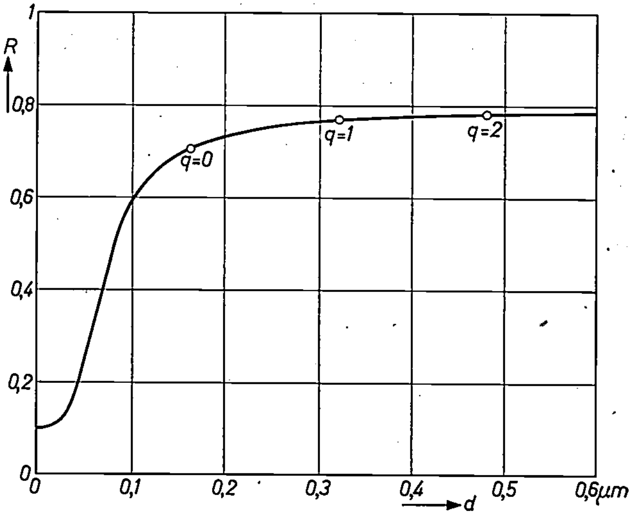


Fig. 10. Calculated reflection coefficient of SnO₂ films on glass for λ = 5.5 μm as a function of the film thickness d. The concentration N of free charge carriers is taken as 6 × 10²⁰ cm⁻³ and the mobility μ as 20 cm²/Vs.

Fig. 11 gives finally a graph of the spectral transmission and reflection of such an optimum film of 0.32 μm thickness. The infra-red reflection reaches 80% whilst the transmission for sodium light is 89%. This latter value is only slightly less than that for an uncoated glass plate (92%). These SnO₂ films can therefore be regarded as practically absorption-free. Measurements on a number of films prepared under identical conditions showed that the spread in reflection and transmission coefficients is only about 1%.

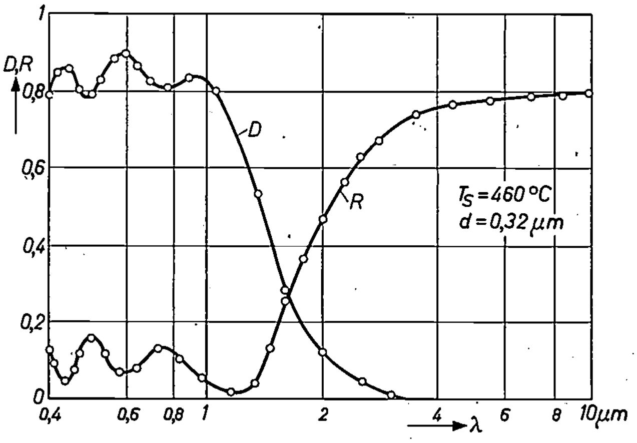


Fig. 11. Spectral transmission D and reflection R of an optimum SnO₂ film of thickness 0.32 μm.

Test lamps (type SO 140 W) whose existing heat absorption screens (glass tubes coaxial with the envelope) were coated internally with these optimum SnO_2 films, were found to have an efficiency of 135 lm/W; normal lamps of this type have an efficiency of 100 lm/W. The light flux was reduced to about 70%. The efficiency could be increased still further by coating the filter on the inside of the envelope and omitting the absorption screens; with a good infra-red filter the absorption screen hardly gets warmed and therefore only causes undesirable and unnecessary reflection of the sodium light. It should therefore be possible to reach an efficiency of at least 140 lm/W for this 140 W lamp. Still better results may be expected if the U-form discharge tube were replaced by a straight tube with a concentric reflector. The combined result of all the above measures should be such as to increase the efficiency of sodium lamps to more than 150 lm/W.

[8] I. Imai, J. Phys. Soc. Japan **15**, 937, 1960.
[9] A. Fischer, Z. Naturf. **9**a, 508, 1954.

Summary. In sodium lamps energy is wasted as thermal radiation from the discharge tube. By means of a selective filter that reflects the infra-red radiation but transmits the sodium light unattenuated, the efficiency of the lamp can be considerably increased. The required properties can be realized in filters whose properties depend on the optical effects of free charge carriers. Certain metals such as gold and copper, as their reddish colour suggests, are good reflectors of infra-red radiation. The transmission of sodium light through very thin films of these metals on glass can be increased by coating them with anti-reflection dielectric layers. Semiconductors with a high concentration of charge carriers also form filters suitable for this purpose. For example, with strongly doped SnO_2 films on glass, filters are obtained that absorb practically no light at all yet strongly reflect the thermal radiation. With the aid of such filters the efficiency of sodium lamps of the type SO 140 W has been raised from about 100 lm/W to 135 lm/W. The possibilities for further improvement are also discussed.

Examples from fluorine chemistry and possible industrial applications

J. Schröder

546.16

The element fluorine is the first member of group 7 of the periodic table, the halogens. Its exposed position in the system of the elements already gives an indication of its extraordinary properties. The extreme affinity of fluorine for nearly all other elements is one of the reasons why its isolation, investigation and experimental handling was originally fraught with great difficulty. Later, however, the number of known fluorine compounds rapidly increased, and our present knowledge of fluorine chemistry makes it seem likely that more inorganic fluorine compounds will be made than those of any other element.

The significance of fluorine chemistry in modern science and industry is based on the special properties of fluorine. Before giving a number of examples of our work in the field of fluorine chemistry, we shall therefore briefly review the preparation, handling and some characteristic properties of fluorine.

The preparation and handling of fluorine and fluorine compounds

Pure fluorine is prepared now as it always was, by electrolysis. However, the method used is varied in many ways according to the purpose for which it is intended. The original hydrofluoric acid used by Moissan has been replaced by potassium hydrofluoride melts with various hydrofluoric acid contents (KF.4HF to KF.HF), the corresponding melting point and working temperature varying from 80 to 240 °C. Copper, monel, silicon-free iron or carbon steel is used for the electrolysis vessel and the cathode; carbon or nickel is used as anode material. The fluorine is purified by passing it over dry sodium fluoride, which adsorbs the hydrogen fluoride carried off from the generator along with the fluorine. A much purer product is obtained if the fluorine is then passed through a trap cooled with liquid oxygen.

It is best to generate the fluorine on the spot, as needed for the experiments. Small amounts of fluorine can be stored in condensed form in quartz, nickel or

Dr. J. Schröder is a research worker at the Aachen laboratory of Philips Zentrallaboratorium GmbH.

Test lamps (type SO 140 W) whose existing heat absorption screens (glass tubes coaxial with the envelope) were coated internally with these optimum SnO_2 films, were found to have an efficiency of 135 lm/W; normal lamps of this type have an efficiency of 100 lm/W. The light flux was reduced to about 70%. The efficiency could be increased still further by coating the filter on the inside of the envelope and omitting the absorption screens; with a good infra-red filter the absorption screen hardly gets warmed and therefore only causes undesirable and unnecessary reflection of the sodium light. It should therefore be possible to reach an efficiency of at least 140 lm/W for this 140 W lamp. Still better results may be expected if the U-form discharge tube were replaced by a straight tube with a concentric reflector. The combined result of all the above measures should be such as to increase the efficiency of sodium lamps to more than 150 lm/W.

[8] I. Imai, J. Phys. Soc. Japan **15**, 937, 1960.
[9] A. Fischer, Z. Naturf. **9**a, 508, 1954.

Summary. In sodium lamps energy is wasted as thermal radiation from the discharge tube. By means of a selective filter that reflects the infra-red radiation but transmits the sodium light unattenuated, the efficiency of the lamp can be considerably increased. The required properties can be realized in filters whose properties depend on the optical effects of free charge carriers. Certain metals such as gold and copper, as their reddish colour suggests, are good reflectors of infra-red radiation. The transmission of sodium light through very thin films of these metals on glass can be increased by coating them with anti-reflection dielectric layers. Semiconductors with a high concentration of charge carriers also form filters suitable for this purpose. For example, with strongly doped SnO_2 films on glass, filters are obtained that absorb practically no light at all yet strongly reflect the thermal radiation. With the aid of such filters the efficiency of sodium lamps of the type SO 140 W has been raised from about 100 lm/W to 135 lm/W. The possibilities for further improvement are also discussed.

Examples from fluorine chemistry and possible industrial applications

J. Schröder

546.16

The element fluorine is the first member of group 7 of the periodic table, the halogens. Its exposed position in the system of the elements already gives an indication of its extraordinary properties. The extreme affinity of fluorine for nearly all other elements is one of the reasons why its isolation, investigation and experimental handling was originally fraught with great difficulty. Later, however, the number of known fluorine compounds rapidly increased, and our present knowledge of fluorine chemistry makes it seem likely that more inorganic fluorine compounds will be made than those of any other element.

The significance of fluorine chemistry in modern science and industry is based on the special properties of fluorine. Before giving a number of examples of our work in the field of fluorine chemistry, we shall therefore briefly review the preparation, handling and some characteristic properties of fluorine.

The preparation and handling of fluorine and fluorine compounds

Pure fluorine is prepared now as it always was, by electrolysis. However, the method used is varied in many ways according to the purpose for which it is intended. The original hydrofluoric acid used by Moissan has been replaced by potassium hydrofluoride melts with various hydrofluoric acid contents (KF.4HF to KF.HF), the corresponding melting point and working temperature varying from 80 to 240 °C. Copper, monel, silicon-free iron or carbon steel is used for the electrolysis vessel and the cathode; carbon or nickel is used as anode material. The fluorine is purified by passing it over dry sodium fluoride, which adsorbs the hydrogen fluoride carried off from the generator along with the fluorine. A much purer product is obtained if the fluorine is then passed through a trap cooled with liquid oxygen.

It is best to generate the fluorine on the spot, as needed for the experiments. Small amounts of fluorine can be stored in condensed form in quartz, nickel or

Dr. J. Schröder is a research worker at the Aachen laboratory of Philips Zentrallaboratorium GmbH.

copper vessels cooled by liquid nitrogen. The storage of fluorine under pressure is however dangerous, and should only be done with special precautions.

Fluorine *compounds* can also be used instead of fluorine for carrying out reactions. All compounds which easily give off or take up fluorine on oxidation or reduction are suitable here. Such compounds are the fluorides of elements which can occur in various valencies and which can thus bind varying amounts of fluorine. Particularly important fluorination agents of this kind are CoF_3/CoF_2 and AgF_2/AgF. Other fluorination agents are compounds which on thermal dissociation give fluorine and inert decomposition products or by-products which can easily be removed. For example, under normal conditions NF_3 is a completely inert, easily handled gas, which dissociates into fluorine and nitrogen at higher temperatures. The fluorination of many compounds, in particular organic ones, can also be carried out *electrochemically* without the use of elementary fluorine, by dissolving the substance in hydrofluoric acid and electrolysing.

The most important methods of preparing fluorine compounds can be divided into four groups:

1) Preparations in aqueous solution, e.g. the reaction of oxides, hydroxides and carbonates with hydrofluoric acid. This is the most common method of preparing binary and complex metal fluorides of normal valency. The reaction products usually contain water of crystallization.
2) The treatment of halides, oxides and many other compounds with 100% hydrogen fluoride. The reaction must usually be carried out at elevated temperatures, and gives anhydrous fluorides of normal valency.
3) The direct reaction of elements or compounds with fluorine. Because of the strong oxidizing power of fluorine, this method normally leads to high, and sometimes anomalous, valencies, e.g. AgF_2, CrF_5, MnF_5, WF_6, SF_6, IF_7.
4) The thermal treatment of two or more co-precipitated or mixed fluorides. This method is mainly used for the preparation of complex compounds.

A special experimental technique must be used for handling fluorine, as practically all materials normally used in the laboratory are attacked by fluorine. Glass reacts with fluorine and many fluorine compounds, giving SiF_4. Organic materials like rubber, polyethylene, tap grease, etc. are attacked by fluorine, often violently. Many metals react with fluorine even at room temperature.

Quartz glass is not attacked by fluorine under normal conditions, but at high temperatures or in the presence of water or organic compounds it is strongly attacked. Sintered corundum is very resistant, and calcium-fluoride ceramics can be used up to 900 °C. Apart from platinum a number of metals such as nickel, copper and aluminium are resistant to fluorine, because they become covered with a protective layer of fluoride. Silver-soldered connections must be protected by thick nickel-plating. Organic materials suitable for vessels and seals are polymeric tetrafluoroethylene and monochlorotrifluoroethylene.. Low-polymer monochlorotrifluoroethylene is a fluorine-resistant grease for taps, ground joints, etc. In *Table I* are given for a number of important materials the temperatures up to which they are not noticeably attacked by fluorine.

Table I. A number of fluorine-resistant materials. *T* is the limiting temperature up to which no noticeable attack occurs.

Material	*T* °C	Material	*T* °C
Platinum	300	Quartz	100
Copper	350	Polytetra-	
Monel	400	fluoroethylene	100
Magnesium	400	Corundum	700
Aluminium	450	CaF_2 ceramics	900
Nickel	500		

Properties of fluorine and fluorine compounds

The difference between the physico-chemical properties of fluorine and those of the other halogens and the neighbouring chalkogens is extremely great. For example, fluorine reacts with nearly all elements and compounds at relatively low temperatures. Most metals, sulphur, boron, carbon, silicon and even asbestos react violently with fluorine at room temperature. Fluorine can displace the oxygen and halogen in halides and oxides, on gentle warming. Hydrogen and hydrides react particularly violently with fluorine. Fluorine even forms stable compounds with inert gases.

The great reactivity of fluorine is due on the one hand to the fact that fluorine forms the strongest known single bonds with other elements. By way of example, *Table II* gives the dissociation energies of HX compounds (X signifies halogen) and the mean bond energies of CX and SiX molecules.

Table II. Dissociation energies and mean bond energies at 291 °K in kcal/mole of fluorides compared with other halides.

Dissociation energies		Mean bond energies			
H-F	136	C-F	110.0	Si-F	132
H-Cl	102.1	C-Cl	67.0	Si-Cl	86
H-Br	85.9	C-Br	54.5	Si-Br	69

The other factor determining the great reactivity is the surprisingly low dissociation energy of the molecule, viz 37.4 kcal/mole, compaıed to 118 kcal/mole for O_2, or 57.2 kcal/mole for Cl_2. The dissociation of F_2 already begins to be appreciable at 400 °C, and its degree of dissociation is greater than that of Cl_2, Br_2 and O_2 at comparable temperatures. A further reason for the aggressive behaviour of fluorine may also be the high diffusion rate of the relatively small molecules and atoms.

Some fluorine compounds have a polar character. For example, the low-valency fluorine compounds of the metals are more strongly polar than the other halides. The lattice energies of the fluorides are higher, and the properties of the salt-like fluorides resemble those of the oxides more than those of the other halides. The fluoride ion can replace oxygen or the isoelectronic OH group in many structures. While e.g. $AlCl_3$ is a largely covalent compound, subliming as low as 180 °C, AlF_3 is a non-volatile salt (melting point 1290 °C). The bond in the HF molecule is also more strongly polar than that in the other hydrogen halides. The proton cannot penetrate far into the electron shell of the fluorine, giving a high dipole moment (1.91 debye). While the other halogen acids are monomeric and very volatile (boiling point of HCl = —83.7 °C), hydrogen fluoride (b.p. = 19.5 °C) is polymerized by means of hydrogen bonds to chains and rings of low volatility. The F-H-F bond is the strongest known hydrogen bridge. The enthalpy of formation for the reaction $HF + F^- \rightarrow HF_2^-$ is about 50 kcal/mole, and for the reaction $2\,HF \rightarrow H_2F_2$ about 20 kcal/mole.

The fluorides of the non-metals and the high-valency transition metals, on the other hand, are typical covalent compounds. The dividing line between the salt-like and volatile fluorine compounds runs diagonally through the periodic table from beryllium via aluminium, titanium and tin to bismuth. The molecules are the more volatile the more completely the central atom is surrounded by fluorine: thus the volatility increases with increasing valency. For example, WF_6 (molecular weight 297.92; b.p. = 17.5 °C) is the heaviest known gas, and the fluorocarbons are more volatile than the hydrocarbons of the same molecular weight. The screening effect of fluorine is so strong that in molecules well surrounded by fluorine the weak intermolecular forces are comparable to those in the noble gases (see *Table III*).

Table III. Boiling point in °C of fluorides compared to that of chlorides, hydrocarbons and noble gases of comparable molecular weight.

	SF_6	C_2F_6	$C_{10}H_{22}$	CCl_4	Xe
Molecular weight	146	138	142.3	153.8	131.3
Boiling point	—64	—78.3	+147	+76.8	—112

Some fluorides, e.g. SF_6, NF_3, CF_4 and higher homologues, also resemble the noble gases in their chemical behaviour, and are completely inert up to moderate temperatures.

A particular feature of fluorine compounds is their very high heat of formation and thermal stability compared to all other compounds. These properties of fluorine compounds also play an important role in the applications described below.

The production of light by fluorine reactions in combustion-type flash lamps

Up till the present, the production of light in combustion-type flash lamps is based exclusively on the combustion of substances with oxygen. Apart from some other metals, the substance burnt is mainly zirconium. Of the properties demanded of a good flash lamp, we shall mention only three of the most important here. First, it should produce as much light as possible in a short time (a few milliseconds). Since normal photographic films are designed to be used for daylight, a second important demand is that the colour temperature of the flash should be as near as possible to that of daylight. The lamp should also be small, handy and safe; i.e. it must naturally not explode during the flash.

In principle, a combustion-type flash lamp is a thermal radiator. The heat of reaction of the combustion process is used to heat the burning or already burnt substance and thus to make it radiate. It is known that the emission of visible radiation increases strongly with increasing temperature of the radiating body. The luminous efficiency of the total radiation from a black body reaches a maximum at 6500 °K, i.e. at the temperature of the sun's surface. The temperature of the substances undergoing combustion should therefore be as high as possible, in order to give a high luminous output and a high colour temperature. This means that the combustion reaction should yield as much energy as possible per unit amount of substance in a given time. For reactions at normal temperature and constant pressure this energy is equal to the enthalpy of formation ΔH, i.e. the energy change of the system on reaction of the stoichiometric number of moles. In the reactions, temperatures are reached at which the reaction products are no longer stable. A limit to further increase of the temperature is thus set by the dissociation. In order to get high temperatures, the substances should be as stable as possible. This stability is determined by the free energy of formation, which can be calculated from the enthalpy of formation and other thermodynamic data. In *fig. 1* the free energies of formation ΔG in kcal/g-atom of fluorine or oxygen for various fluorides and oxides are plotted against the temperature. For e.g. ZrO_2 and ZrF_4, the free energies are for the reactions $\frac{1}{2}Zr + \frac{1}{2}O_2 \rightarrow \frac{1}{2}ZrO_2$; $\frac{1}{4}Zr + \frac{1}{2}F_2 \rightarrow \frac{1}{4}ZrF_4$. If the energies are referred to equal amounts of the metal (Zr), the free energies for the fluorides are twice as high.

Comparison of these graphs shows that the reac-

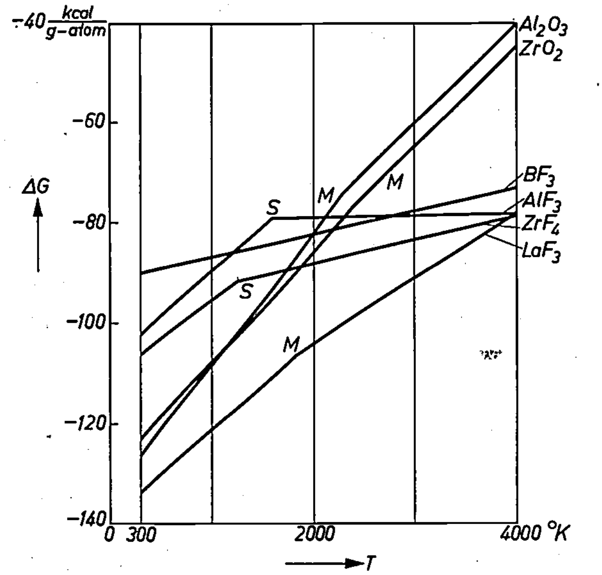


Fig. 1. Free energies of formation ΔG per gram-atom of oxygen for certain oxides, compared with the free energy of formation per gram-atom of fluorine for some fluorides in the temperature range from 300 to 4000 °K. M is melting point, S is sublimation point.

tions with fluorine can occur at higher temperatures than those with oxygen. In order to get a high colour temperature as well as a large heat of reaction, it is therefore advisable to carry out the combustion with fluorine instead of with oxygen, as has been done so far.

Since elementary fluorine is difficult to handle, it is better to use gaseous fluorine compounds with fluorine contents as high as possible, a low thermal stability, and under normal conditions a lower chemical reactivity than fluorine (e.g. OF_2, $\Delta H = +7.6$ kcal/mole, or NF_3, $\Delta H = -27$ kcal/mole).

The energy balance of the combustion is favoured not only by the ease of thermal dissociation of such fluorides but also by the fact that they contain more than two bonds per molecule, and therefore release more free valencies for reaction than molecular fluorine with the same volume of gas and the same pressure. In OF_2 two additional oxygen valencies are involved. The volume of the lamp can thus be made smaller for a given number of equivalents reacted. By way of example, we give below the reaction equations for the combustion of zirconium with equal volumes of O_2, NF_3 and OF_2, together with the corresponding enthalpies of formation:

$$2\,O_2 + 2\,Zr \rightarrow 2\,ZrO_2 \qquad -516\ \text{kcal},$$
$$2\,NF_3 + 1\tfrac{1}{2}\,Zr \rightarrow 1\tfrac{1}{2}\,ZrF_4 + N_2 \qquad -630\ \text{kcal},$$
$$2\,OF_2 + 2\,Zr \rightarrow ZrF_4 + ZrO_2 \qquad -729\ \text{kcal}.$$

Some results obtained so far from our experiments with fluorine reactions in flash lamps may be summarized as follows [1]:

1) Both fluorine and certain fluorine compounds (e.g. NF_3 and OF_2) produce light on reaction with finely divided metals (e.g. Zr).
2) The colour temperature of the light emitted in fluorine reactions is higher than that in corresponding oxygen reactions.
3) The amount of light produced from the reaction of a given amount of substance is in some cases greater than with combustions using oxygen.

Transport reactions for stabilization of filaments in electric incandescent lamps

The luminous efficiency and the brightness of an incandescent lamp increase strongly with the temperature of the filament. Only substances with a high melting point can therefore be used as filament materials, e.g. tungsten (m.p. = 3370 °C), carbon (m.p. = about 3750 °C) and tantalum carbide (m.p. = 3880 °C). But the vapour pressures of these substances are also strongly temperature dependent. Thus at high temperatures the lamp bulb blackens and finally destruction of the hot filament is caused by its evaporation.

The filament does not evaporate uniformly. No matter how precisely the filament is made, the temperature of certain parts of the filament will be higher than the average value. The reason for this is not yet quite clear; but all such "hot spots", no matter what their precise nature, can be regarded in principle as constrictions of the filament. The decrease in the cross-sectional area and the resulting increase in the electrical resistance leads to an increase in the temperature at these spots. This causes the rate of evaporation to increase, so that the constriction becomes even narrower. Since the total resistance of the filament is hardly affected by these little spots, it is easy to understand that the wire always burns out at one of these hotter spots.

The appearance of such spots, and in fact the burning out of the hot filament and the blackening of the bulb by vaporization can in principle be prevented by chemical transport reactions with fluorine [2].

Fluorine is the only element to react at room temperature with tungsten and carbon, giving the gaseous fluorides WF_6 and CF_4. When fluorine is present in the gas filling of the lamp, the evaporated filament material reacts to give gaseous fluoride, which undergoes thermal dissociation at the hot filament, depositing the filament material there again. This cycle can be maintained at any bulb temperature, and completely pre-

[1] Both the idea and the investigation of the use of fluorine reactions in flash lamps are very largely due to the valuable cooperation of L. M. Nijland, Lighting Division, Philips, Eindhoven. A more detailed description of the investigation will appear in due course in this journal.

[2] The basic principles of chemical transport reactions are described in detail in the book: H. Schäfer, Chemische Transportreaktionen, Verlag Chemie, Weinheim, 1962.

[3] W. Schilling, Elektrotechn. Z. B 13, 485-487, 1961.

vents the blackening of the bulb by evaporated filament material (see *fig. 2*).

The dissociation of tungsten fluoride and the deposition of tungsten only occur at very high temperatures in the immediate vicinity of the hot filament. With other halogens, in particular iodine which is now used in the well-known iodine lamp, the deposition occurs at relatively low temperatures in the gas filling, so that a cloud of atomic tungsten is formed round the filament [3]. While this already has a positive effect, the situation is basically more favourable with fluorine, as we shall now briefly show.

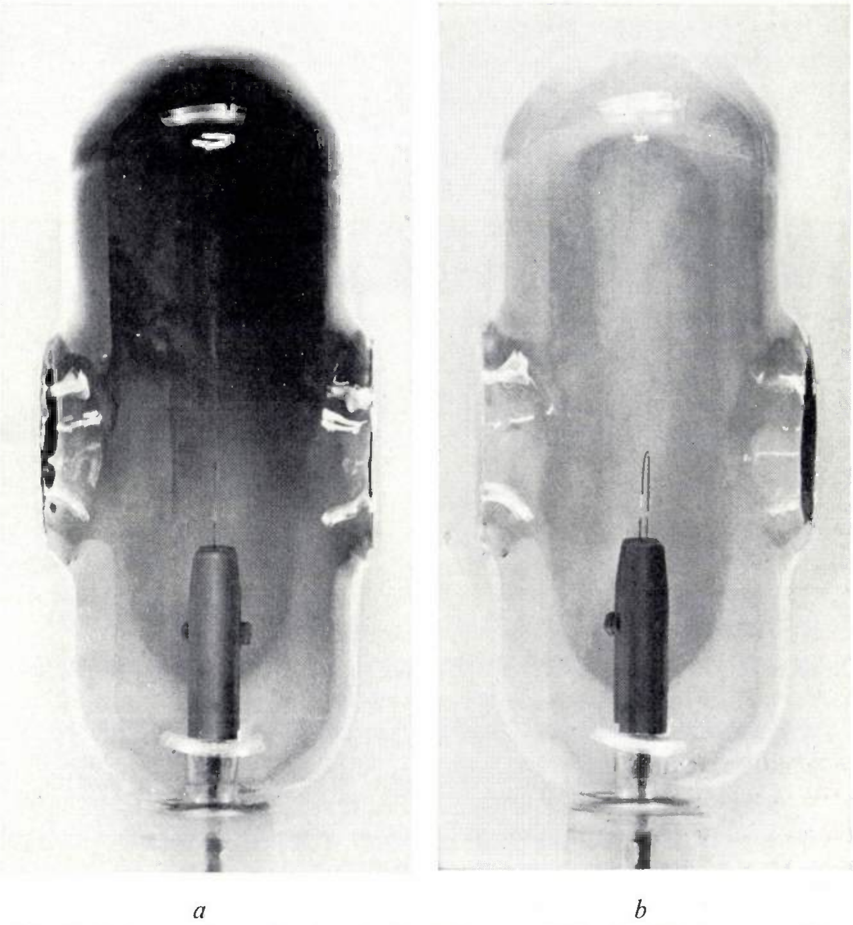


Fig. 2. *a*) Tungsten lamp after burning for 10 hours at 3000 °C in 500 torr argon. The strong blackening is due to evaporated tungsten.
b) The same lamp as in (*a*), after having burnt for 15 minutes at 3100 °C in 500 torr argon + 2 torr WF_6. The blackening has completely disappeared.

The kinetic processes occurring in the lamp are complicated and difficult to describe quantitatively, the complication being that very steep temperature and concentration gradients are present and give rise to convection currents. For qualitative considerations we can however assume that a thin layer of gas remains in contact with the filament, and thus has practically the same temperature as the latter. In this case, we can speak to a first approximation of an equilibrium process, and write the following expression for the equilibrium constant of the dissociation reaction $WF_6 \rightleftarrows W + 6\ F$:

$$\frac{P_W \cdot (P_F)^6}{P_{WF_6}} = K_p .$$

According to Van 't Hoff, the variation of the equilibrium constant with temperature is given by:

$$\frac{d \ln K_p}{dT} = \frac{\Delta H}{RT^2},$$

where ΔH is the enthalpy of formation and R the gas constant. According to the Clausius-Clapeyron equation, the temperature dependence of the vapour pressure in the evaporation equilibrium may be described by an equation of the same form, with ΔH replaced by the enthalpy of evaporation Q:

$$\frac{d \ln P}{dT} = \frac{Q}{RT^2} .$$

Fig. 3 shows the temperature variation of the values of log K_p and log P calculated by Ulich's approxima-

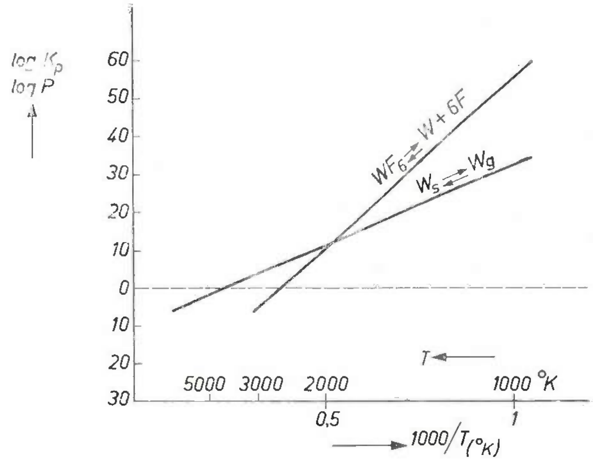


Fig. 3. Temperature dependence of the dissociation equilibrium $WF_6 \rightleftarrows W + 6\ F_6$ and the evaporation equilibrium $W_{solid} \rightleftarrows W_{gas}$. K_p = equilibrium constant of WF_6 dissociation, P = tungsten pressure, T = absolute temperature.

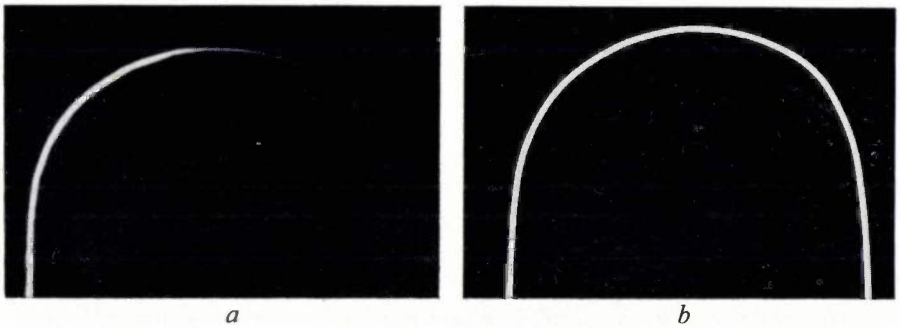


Fig. 4. *a*) Photo of a tungsten filament directly after the switching on of the lamp. The left-hand side of the wire loop is thinner than the right-hand side. The temperature of the wire is 3100 °C on the left and 2800 °C on the right.
b) Photo of same tungsten filament as in (*a*), but after burning for 10 minutes in 500 torr argon + 2 torr WF_6. The originally non-uniform geometry and temperature of the wire have been completely equalized by chemical transport via WF_6.

tion. The following conclusions may be drawn from this diagram:

1) In the temperature range in question for the burning temperature of tungsten lamps, WF_6 just begins to dissociate to an appreciable extent.
2) The dissociation of WF_6 varies more strongly with the temperature than does the vapour pressure of the tungsten. As a result, a temperature increase at some point on the filament will increase the deposition of tungsten due to dissociation faster than the evaporation. Tungsten will thus be transported to the hot spot until its temperature again becomes equal to the mean value for the filament.

Unlike the other halides, which are largely dissociated at much lower temperatures and where the temperature dependence of the dissociation is less than that of the vapour pressure of the corresponding filament material, the fluoride can thus serve to homogenize and keep constant the temperature and the form of the filament.

The experiments described below will help to illustrate and confirm this regenerative transport.

A small length of a tungsten filament was electrolytically etched. This reduction in the cross-section produced an artificial hot spot, whose temperature was higher than that of the rest of the filament. When the lamp was burnt with a gas filling containing a few torr of WF_6 or NF_3, the geometry and temperature of the whole length of the wire became completely equalized after a short burning time. *Fig. 4a* shows a photo of a filament etched on the left side, immediately after the lamp was switched on, and fig. 4*b* shows the same filament after a short burning time in a WF_6 atmosphere at 3100 °C. *Fig. 5a* shows a microscopic view of the junction between the etched and the unetched wire, and fig. 5*b* shows the same spot after brief burning in a WF_6 atmosphere.

Summarizing, we may state that transport reactions involving fluorine hinder the burning out of hot filaments as a result of evaporation, and the blackening of the bulb of the lamp by the evaporated filament material. The fluorine incandescent lamp can therefore be used at filament temperatures not far below the melting point. Compared with the present attainable values of luminous efficiency and brightness, this increase in the burning temperature gives a considerable improvement of these factors. It should however be stressed that the results obtained so far all refer to laboratory experiments; there are still many problems to be solved before the process can be applied to commercial lamps.

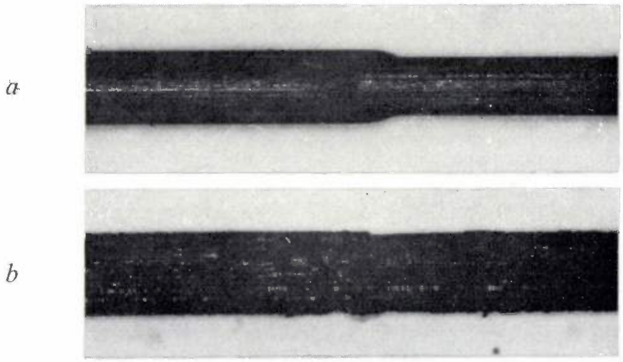


Fig. 5. *a*) Tungsten wire whose right-hand side has been made much thinner by etching.
b) The same spot on the tungsten wire after burning for 15 minutes at 3000 °C in 500 torr argon + 2 torr WF_6.

Summary. The significance of fluorine chemistry for modern science and industry is based on the special properties of fluorine and its compounds. Fluorine is the most reactive of all elements. Even at relatively low temperatures, it reacts with nearly all other elements and compounds. A special experimental technique is needed to handle this very aggressive element and its compounds.

This article gives a brief introduction to fluorine chemistry. After a short description of the preparation, handling and some important properties of fluorine and its compounds, two of the author's own contributions to fluorine chemistry are described. The first contribution refers to fluorine reactions, which owing to their great reaction energies, are suitable for the production of actinic light in "combustion-type flash lamps". In the second piece of work, the reactivity of fluorine and the stability of fluorides are made use of to carry out chemical transport reactions at very high temperatures in incandescent lamps.

Chemical transport reactions

A. Rabenau

536.422:548.55

A "chemical transport reaction" may be defined as the transference of a condensed phase $A_{(f)}$ (which may be a solid or liquid) through a gaseous phase by means of a chemical reaction in which gases or vapours $B_{(g)}$, $C_{(g)}$, $D_{(g)}$ ... are involved:

$$A_{(f)} + B_{(g)} \rightleftarrows C_{(g)} + D_{(g)} + \dots .$$

This definition of the process shows that it differs essentially from sublimation and distillation. A chemical transport reaction is necessarily reversible; a concentration gradient is induced e.g. by means of a temperature gradient which reverses the reaction and causes substance A to precipitate out of the gas phase. Such reactions have been applied industrially for many years. One example is the Mond process for purifying nickel, using carbon monoxide as the transport agent:

$$Ni_{(f)} + 4CO_{(g)} \rightleftarrows Ni(CO)_{4(g)},\ 50 \rightarrow 190\ °C.$$

Another is Van Arkel, De Boer and Fast's method for producing pure metals:

$$Zr_{(f)} + 2I_{2(g)} \rightleftarrows ZrI_{4(g)},\ 280 \rightarrow 1450\ °C.$$

Once Schäfer [1] had worked out the basic principles of chemical transport reactions and indicated their many potential applications, a whole series of systematic investigations was undertaken in various quarters. In the electronic industry, the interest of investigators has been mainly concentrated on the growth of single crystals and the production of thin films, particularly epitaxial films. The present article is concerned with the first of these.

The chemical transport process differs from sublimation in that the vapour pressure of the condensed phase is negligibly small in the range of temperature within which transport takes place. Its great advantage, then, is that one can work at much lower temperatures than would be feasible in a sublimation process. Thus it is possible to work with substances which, for a high enough sublimation pressure to be achieved, would have to be raised to temperatures so high that the problem of crucible material and contamination would become critical. Again, it is possible to transport compounds whose dissociation pressures, even below the melting point, are of magnitudes that cannot be controlled with straightforward laboratory equipment. In many cases the ability to work at low temperatures allows phase transformations to be avoided that would otherwise lead to destruction of the single crystal.

There is a drawback in that the transport agent may be incorporated into the lattice of the crystal that is grown by the transport reaction. However, there is some choice of transport agents and it will usually be possible to find one which will not affect the special properties, e.g. conductivity, of the crystal in which it is incorporated. Alternatively, a transport agent may be available which, due to its relatively large atomic size, will only be incorporated in negligibly small quantities.

One approach to the choice of transport agents and operating temperatures can be made by a thermodynamic study of the system under consideration. Questions such as these and the rate of transport are dealt with at length in Schäfer's monograph [1].

To those concerned with growing single crystals, the size of the resulting crystal is often a major consideration. The usual transport process, carried out in a stationary system with a constant temperature gradient, will only in favourable conditions yield crystals whose dimensions are of the order of one centimetre or so. Seeding probabilities under the prevailing conditions of supersaturation are such that it is more likely to produce a large number of very small crystals. In recent years a method has been developed in this laboratory which in many cases yields crystals of appreciable size, despite these difficulties. The problem, in the first instance, was to grow gallium phosphide crystals. GaP melts at about 1470 °C under a phosphorus pressure of about 30 atm. There was therefore no question of adopting the conventional method of growing from the melt. The obvious step was to see whether the chemical transport process could be employed. It was found that the compound could be transported with iodine as transport agent through a temperature gradient ranging from 1000 °C down to 900 °C. The result, however, was a mass of microcrystals. Maak [2] conceived the idea of combining the transport process with the technique of seed selection familiar from the conventional method of growing from the melt. The collaboration of numerous other workers finally led to the development of a process which effectively solved our problem, and which, in view of its potential usefulness in similar applications we shall now describe.

Dr. A. Rabenau (deputy director), Aachen laboratory of Philips Zentrallaboratorium GmbH.

[1] H. Schäfer, Chemische Transportreaktionen, Verlag Chemie, Weinheim 1962.

[2] I. Maak, unpublished communication.

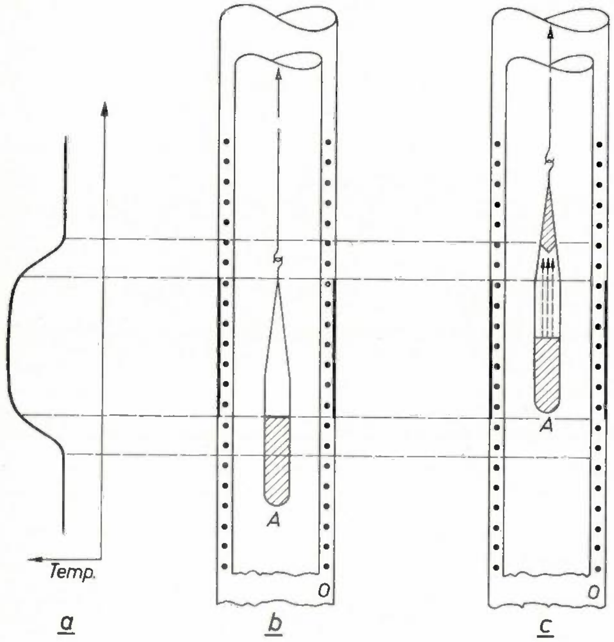


Fig. 1. The principle of crystal growth using a chemical transport reaction. *a*) curve of temperature as a function of height in the furnace, *b*) situation at the start of the pulling operation, *c*) a later stage of the transport process. Capsule *A* is suspended in a furnace *O* consisting of two concentric quartz glass tubes, the heating element being wound round the smaller of the two. A section of the inside surface of the outer quartz glass tube carries a vacuum-deposited gold film which reflects radiant heat, and so creates a high-temperature zone (see diagram *a*) in this part of the furnace.

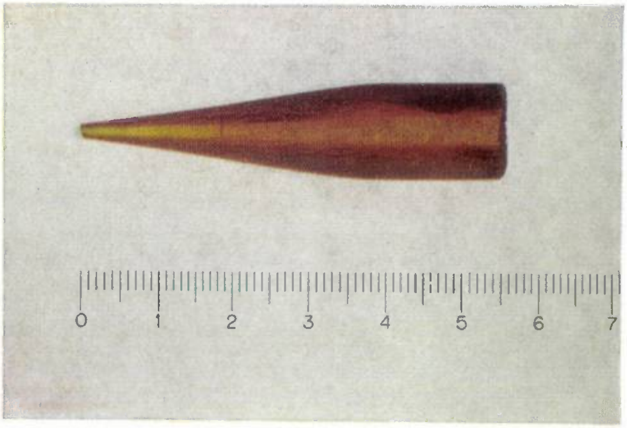


Fig. 2. Longitudinal section of a GaP single crystal produced in a transport-pulling experiment.

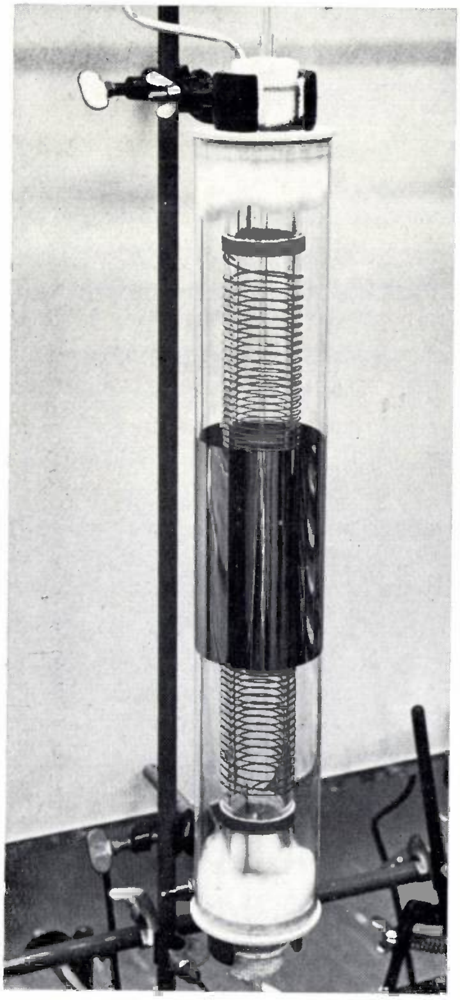


Fig. 3

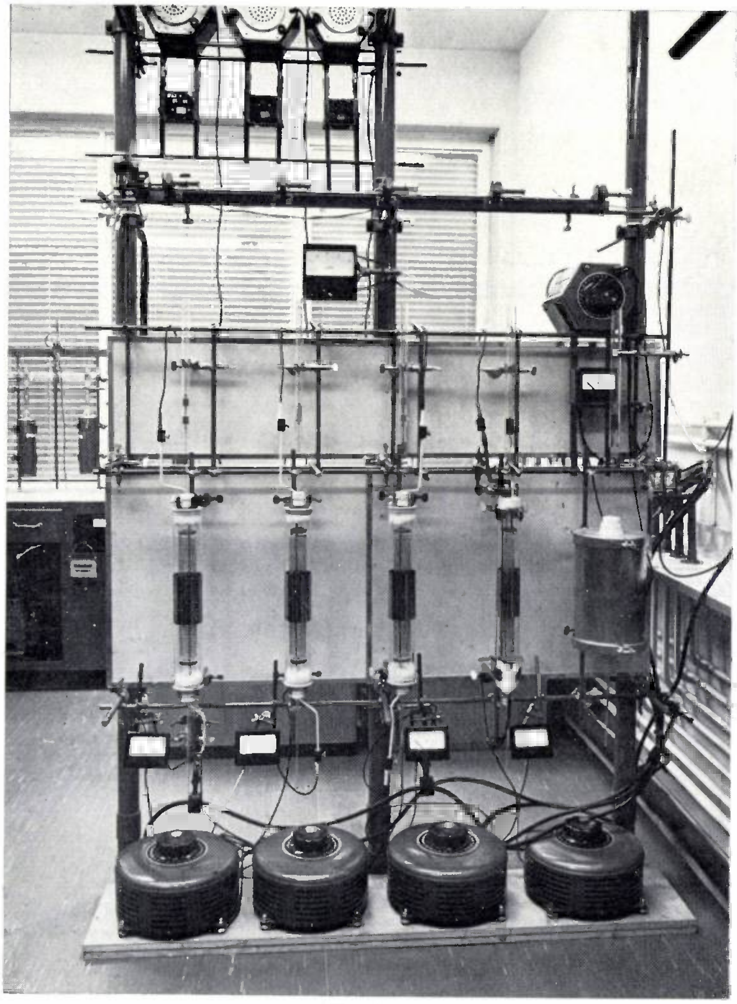


Fig. 4

Fig. 3. Crystal-pulling furnace working on the principle described. The heater coil and gold film can be clearly seen. The film transmits some light in the visible range of the spectrum, and so the capsule can be observed during the pulling process.
Fig. 4. A battery of four crystal-pulling furnaces, together with the associated power supplies and pulling gear.

The transport-pulling process

The principle is illustrated schematically in *fig. 1.* The substance to be transported is enclosed, together with the transport agent, in a quartz capsule, the upper part of which is tapered to a point. The tapered end enters the zone of highest temperature first. This cleans the inner surface of the capsule by the chemical transport of any crystallites deposited on the glass during the initial heating which would otherwise cause spurious nucleation. As the pulling operation progresses the lower end of the capsule enters the maximum temperature zone, and at the same time the pointed tip passes into a zone of lower temperature; at a certain point the transported substance starts to deposit. The shape of the upper part of the capsule is such as to favour seed selection, and consequently the number of crystals that grow laterally across the tip of the capsule is only limited. With luck there will only be one such crystal. It has been possible in this way to produce crystals having a length of several centimetres; an example may be seen in *fig. 2.* GaAs as well as GaP has so far been transported in this manner, and mixed crystals containing the two compounds in fixed proportions have also been obtained.

The general design of the crystal-pulling furnace may be seen in *fig. 3.* It consists of a heating element wound on a quartz tube which is surrounded by a second concentric tube. A section of the inside surface of the outer tube carries a film of gold about 0.05 μm thick, produced by vacuum deposition. The film is translucent to visible light but it has a high infra-red reflectance, thus limiting the radiation of heat outside the system and so producing a localized high-temperature zone.

A number of these furnaces have been combined into the single unit shown in *fig. 4.* The power supplies may be seen in the foreground of the photograph; the pulling mechanism is mounted above the furnaces.

Summary. An account is given of the application of chemical transport reactions to the growth of single crystals and a comparison is made of their advantages and disadvantages with respect to sublimation methods. A "transport-pulling" method has been developed for growing large crystals; it consists in raising a tapered quartz capsule through a furnace in which a certain temperature gradient prevails, the conditions being such as to give rise to a chemical transport reaction. The pointed end of the capsule has a shape that favours seed selection. The equipment and method used for growing gallium phosphide crystals are described by way of example.

Investigations on $BaTiO_3$ semiconductors

E. Andrich and K. H. Härdtl

621.315.592.4

Introduction

In the last 20 years the compound $BaTiO_3$ has been the subject of numerous publications. The great majority of the investigations have been concerned with the dielectric properties of barium titanate, e.g. its ferroelectric behaviour. Since the occurrence of electrical conductivity is a hindrance in dielectric investigations, efforts were made to keep the resistance as high as possible.

It now appears that the resistivity of $BaTiO_3$ can be lowered to about 10 ohm-cm by doping with La, Nb, Sb and other elements. Experiments on such semiconducting $BaTiO_3$ samples have also yielded quite remarkable results in regard to their semiconducting properties. In *fig. 1* the resistivity of semiconducting $BaTiO_3$, and that of $(Ba,Sr)TiO_3$ and $(Ba,Pb)\,TiO_3$ mixed crystals is plotted as a function of temperature. It can be seen that there is a steep and very pronounced increase of resistivity near the ferro-electric Curie temperature T_C. The coincidence of the Curie point with the tem-

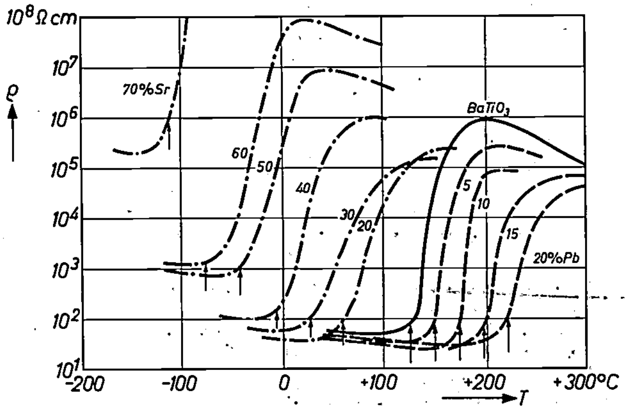


Fig. 1. Temperature dependence of the resistivity ρ of semiconducting $BaTiO_3$, and of $(Ba,Sr)TiO_3$ and $(Ba,Pb)TiO_3$ mixed crystals. Depending on composition, the Curie temperature T_C lies between −100 °C and +220 °C.

Dipl. Phys. E. Andrich and Dr. K. H. Härdtl are research workers at the Aachen laboratory of Philips Zentrallaboratorium GmbH.

The transport-pulling process

The principle is illustrated schematically in *fig. 1.* The substance to be transported is enclosed, together with the transport agent, in a quartz capsule, the upper part of which is tapered to a point. The tapered end enters the zone of highest temperature first. This cleans the inner surface of the capsule by the chemical transport of any crystallites deposited on the glass during the initial heating which would otherwise cause spurious nucleation. As the pulling operation progresses the lower end of the capsule enters the maximum temperature zone, and at the same time the pointed tip passes into a zone of lower temperature; at a certain point the transported substance starts to deposit. The shape of the upper part of the capsule is such as to favour seed selection, and consequently the number of crystals that grow laterally across the tip of the capsule is only limited. With luck there will only be one such crystal. It has been possible in this way to produce crystals having a length of several centimetres; an example may be seen in *fig. 2.* GaAs as well as GaP has so far been transported in this manner, and mixed crystals containing the two compounds in fixed proportions have also been obtained.

The general design of the crystal-pulling furnace may be seen in *fig. 3.* It consists of a heating element wound on a quartz tube which is surrounded by a second concentric tube. A section of the inside surface of the outer tube carries a film of gold about 0.05 μm thick, produced by vacuum deposition. The film is translucent to visible light but it has a high infra-red reflectance, thus limiting the radiation of heat outside the system and so producing a localized high-temperature zone.

A number of these furnaces have been combined into the single unit shown in *fig. 4.* The power supplies may be seen in the foreground of the photograph; the pulling mechanism is mounted above the furnaces.

Summary. An account is given of the application of chemical transport reactions to the growth of single crystals and a comparison is made of their advantages and disadvantages with respect to sublimation methods. A "transport-pulling" method has been developed for growing large crystals; it consists in raising a tapered quartz capsule through a furnace in which a certain temperature gradient prevails, the conditions being such as to give rise to a chemical transport reaction. The pointed end of the capsule has a shape that favours seed selection. The equipment and method used for growing gallium phosphide crystals are described by way of example.

Investigations on $BaTiO_3$ semiconductors

E. Andrich and K. H. Härdtl

621.315.592.4

Introduction

In the last 20 years the compound $BaTiO_3$ has been the subject of numerous publications. The great majority of the investigations have been concerned with the dielectric properties of barium titanate, e.g. its ferroelectric behaviour. Since the occurrence of electrical conductivity is a hindrance in dielectric investigations, efforts were made to keep the resistance as high as possible.

It now appears that the resistivity of $BaTiO_3$ can be lowered to about 10 ohm-cm by doping with La, Nb, Sb and other elements. Experiments on such semiconducting $BaTiO_3$ samples have also yielded quite remarkable results in regard to their semiconducting properties. In *fig. 1* the resistivity of semiconducting $BaTiO_3$, and that of $(Ba,Sr)TiO_3$ and $(Ba,Pb) TiO_3$ mixed crystals is plotted as a function of temperature. It can be seen that there is a steep and very pronounced increase of resistivity near the ferro-electric Curie temperature T_C. The coincidence of the Curie point with the tem-

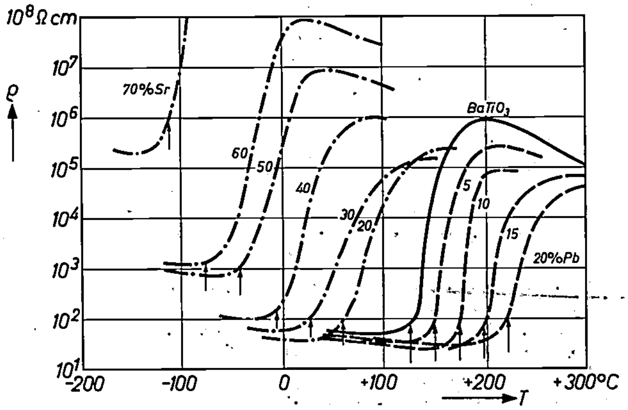


Fig. 1. Temperature dependence of the resistivity ρ of semiconducting $BaTiO_3$, and of $(Ba,Sr)TiO_3$ and $(Ba,Pb)TiO_3$ mixed crystals. Depending on composition, the Curie temperature T_C lies between −100 °C and +220 °C.

Dipl. Phys. E. Andrich and Dr. K. H. Härdtl are research workers at the Aachen laboratory of Philips Zentrallaboratorium GmbH.

perature at which the resistivity shows a steep rise immediately suggests a correlation between the semiconducting properties and the dielectric properties of the substance. Investigations into the nature of this correlation are of considerable interest. Moreover, the high positive temperature coefficient of the resistance (maximum 60% per degree) offers interesting prospects for the use of $BaTiO_3$ semiconductors as switching elements. A special advantage in this connection is that the range with maximum resistance-temperature coefficients can be shifted between —90 °C and +400 °C by varying the composition of the mixed crystals [1,2].

In the following we shall consider the physical causes of this increase of resistivity and discuss some applications of semiconducting $BaTiO_3$ as switching elements.

Physical investigations

As can be seen from fig. 1, the electrical resistivity of semiconducting $BaTiO_3$ rises by several powers of ten in a relatively narrow temperature interval in the vicinity of the Curie point [3,4]. An initial clue to the reason for this extraordinary behaviour is the fact that the effect is only found with ceramic material. No such increase of resistivity has been found in experiments on semiconducting $BaTiO_3$ single crystals [5]. This immediately suggests that the resistivity increase is due to the grain boundaries in the polycrystalline ceramic, which form regions of higher resistance and, at least above T_C, govern the total series resistance of grain volume and boundary.

Fig. 2 shows the one-dimensional model of a polycrystalline conductor of this nature, with a sequence of grains *K* interrupted at the grain boundaries *G* by thin boundary layers of high resistance. As the grains are so small, direct proof of the existence of these boundary layers cannot be obtained by the usual methods of measuring resistivity, but it can be deduced from the following experiment [6].

When a current *I* flows through a series of grains as represented by the model in fig. 2, the marked difference in resistivity between "high-ohmic" grain boundary and "low-ohmic" grain volume will result in a stepwise voltage drop $V(x)$. The grain boundaries, therefore, are regions of high electric field strength. In a one-dimensional model, as in *fig. 3*, this strong electric field spreads out where the grain boundaries meet the surface of the material. As a result of this strong, inhomogeneous stray field, dielectric particles experience a force that pulls them towards the grain boundary. If, therefore, the surface of a polycrystalline conductor, through which a current flows, is coated with a suspension of dielectric particles in a non-conducting fluid, one would expect the particles to show a preference to accumulate at the grain boundaries, provided the dielectric constant of the particles is large with respect to that of the fluid. This behaviour should be observable under the microscope.

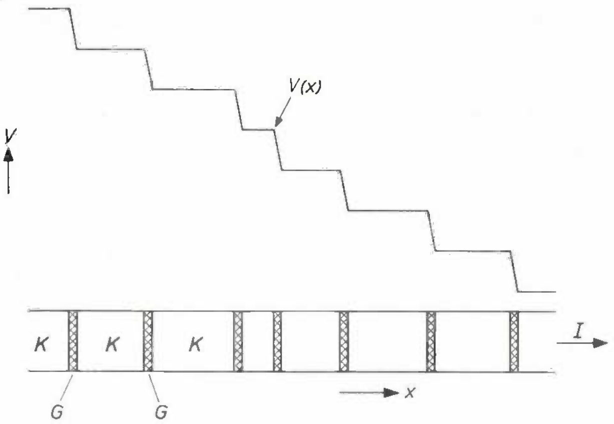


Fig. 2. One-dimensional model of a polycrystalline conductor consisting of grains *K* and barrier layers of high resistance at the grain boundaries *G*. When a current *I* flows through the conductor, the voltage drop $V(x)$ follows a stepped curve.

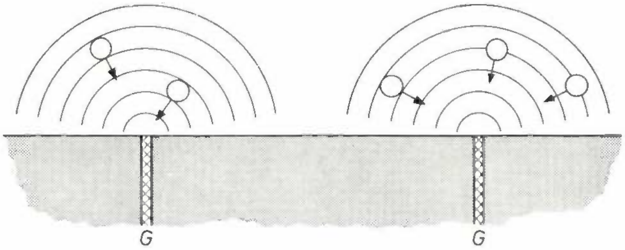


Fig. 3. Principle of making the grain boundaries *G* visible by means of a suspension of dielectric particles under the influence of the electric field at the grain boundaries.

Fig. 4a shows a slice of a ceramic, semiconducting $BaTiO_3$ specimen. The grain boundaries are clearly visible. Fig. 4*b* depicts the surface after coating with a suspension of TiO_2 powder in silicon oil and applying a field in the direction marked by the arrow. The accumulation of particles at the grain boundaries can be seen distinctly.

[1] O. Saburi, J. Amer. Cer. Soc. **44**, 54, 1961.
[2] B. Frank, unpublished.
[3] P. W. Haayman, R. W. Dam and H. A. Klasens, Federal German Patent No. 929350, 1955.
[4] O. Saburi, J. Phys. Soc. Japan **14**, 1159, 1959.
[5] G. Goodman, J. Amer. Cer. Soc. **46**, 48, 1963.
[6] P. Gerthsen and K. H. Härdtl, Z. Naturf. **18a**, 423, 1963.
[7] For a summary of the subject and a review of the literature, see: J. Volger, Progress in Semiconductors **4**, 207, 1960 (Heywood, London).
[8] W. Heywang, Solid State Electronics **3**, 51, 1961.

This experiment demonstrates the occurrence of grain boundaries of very high resistivity above T_C. They represent an inhomogeneous resistance, which is apparent from the following effects:

1) The resistance is frequency dependent. *Fig. 5* shows the real part of the resistivity as a function of temperature and frequency. The decrease of resistivity with frequency can be explained with the aid of the equivalent circuit in *fig. 6*, representing the series arrangement of a grain and a grain boundary [7]. At zero frequency the total resistance is given by the sum of the grain boundary resistance R_G and grain volume resistance R_K. As the frequency increases, the high resistance R_G is shunted by the grain boundary capacitance C_G, so that at high frequencies the total resistance approaches the bulk resistance R_K.

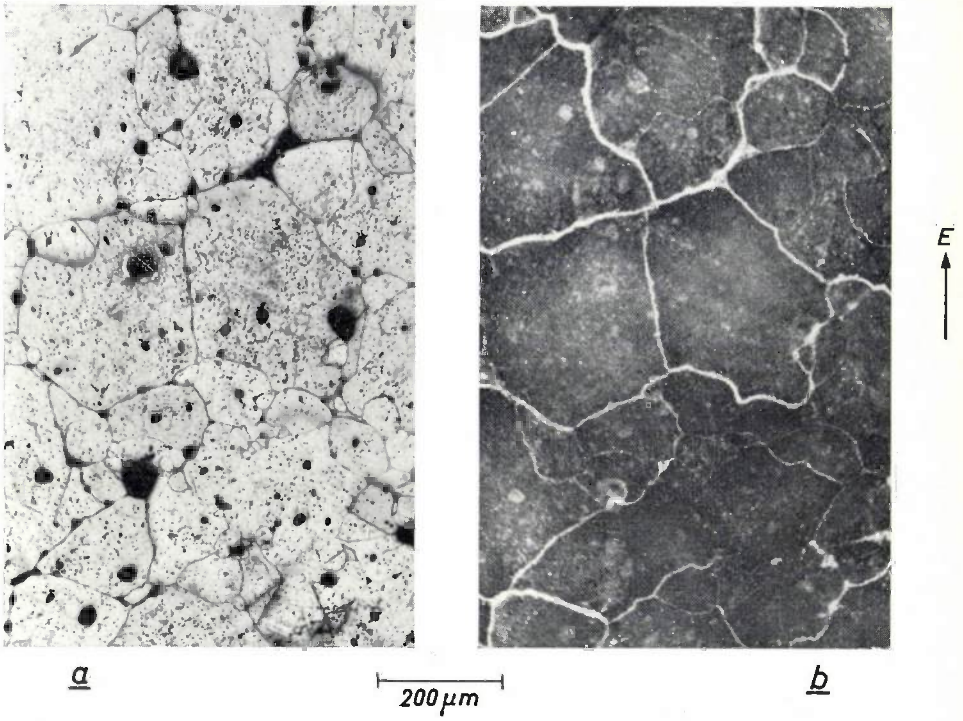
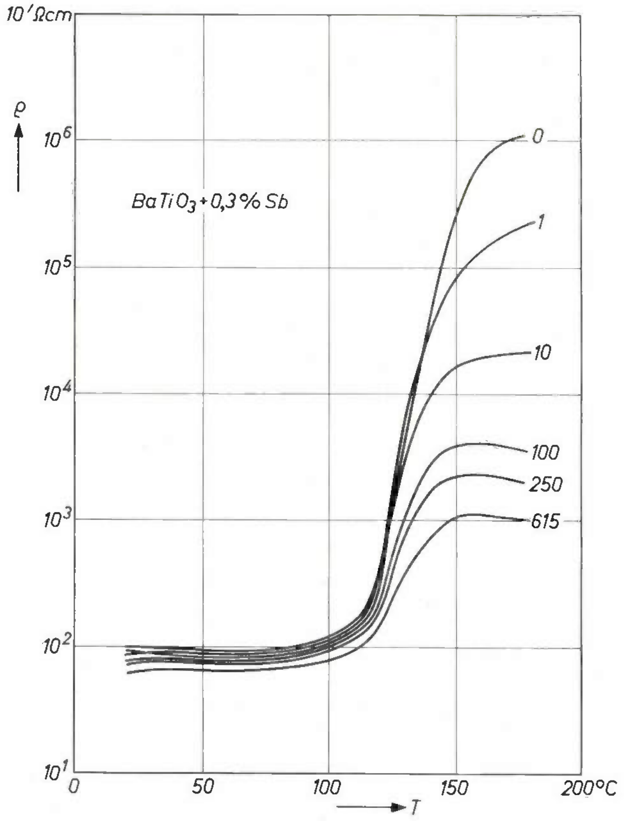


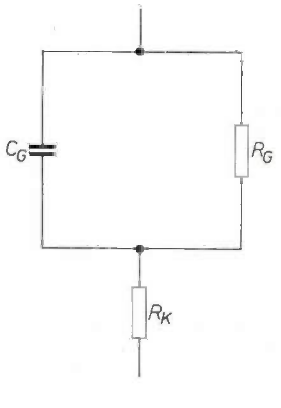
Fig. 4. *a*) Slice of polycrystalline BaTiO$_3$. *b*) The same slice after coating the layer with a suspension of TiO$_2$ particles in silicon oil and applying an electric field E in the direction of the arrow. The voltage per grain boundary is about 1 volt, the temperature 170 °C.



← Fig. 5. Real part of the resistivity of BaTiO$_3$ + 0.3% Sb as a function of temperature at various frequencies (in kc/s).

2) Non-linear current-voltage characteristics. *Fig. 7* shows the current density j as a function of the voltage per grain boundary V at various temperatures. Under heavy loading the temperature was kept constant by using a pulse technique for the measurements. At low voltages the characteristic curves are linear. At higher voltages, however, characteristics are found which have the form $j \propto V^3$, resembling the characteristics of voltage-dependent resistors (VDR) of SiC. Such deviations from Ohm's law are typical of non-linear resistors.

The cause of the high grain-boundary resistivities remains an open question. Heywang [8] suggests that the high-resistance boundary layers are built up from depletion layers, formed by electrons passing from the conduction band in electronic surface levels S to the



→ Fig. 6. Equivalent circuit for grains and grain boundaries. R_G = grain boundary resistance. C_G = grain boundary capacitance, R_K = grain volume resistance.

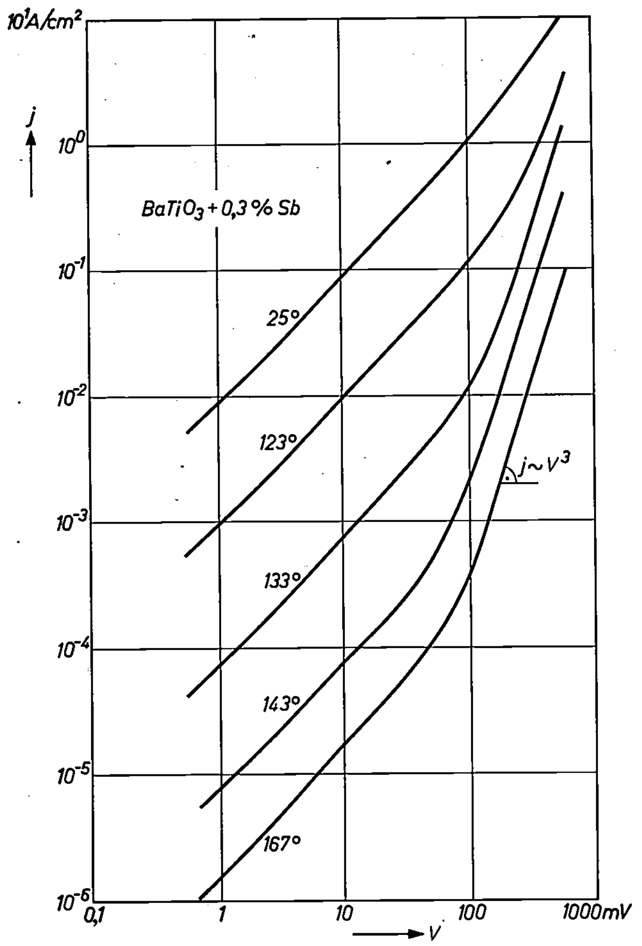


Fig. 7. Current-voltage characteristics of $BaTiO_3$ + 0.3 % Sb The current density j is plotted as a function of V, the voltage per grain boundary.

grain boundary. This gives rise to a high-resistance zone of thickness $2L$ which is free of charge-carriers, and which is left with an uncompensated positive space charge in the form of positively charged donors with a density n_D. This space charge leads to a potential "bulge" which, at the grain boundary, reaches a maximum given by:

$$\varphi_0 = \frac{e\, n_D\, L^2}{2\varepsilon\, \varepsilon_0} . \quad \ldots\ldots \quad (1)$$

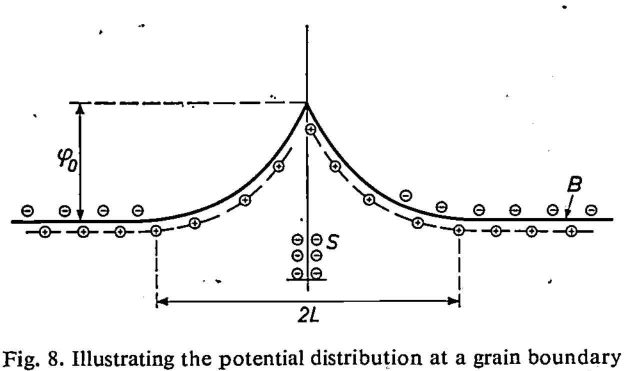


Fig. 8. Illustrating the potential distribution at a grain boundary of $BaTiO_3$. B = conduction band edge, S = electronic surface levels.

For a current to flow the electrons have to overcome this potential barrier by thermal movement. The resistivity is therefore given by:

$$\rho \propto \exp\,(\varphi_0/kT)\,, \quad \ldots\ldots \quad (2)$$

which indicates that the resistivity is closely dependent on the height of the potential barrier.

To explain the temperature behaviour of the resistivity as represented in fig. 1, the semiconducting properties just described have to be correlated with the known dielectric properties of $BaTiO_3$. This is done with the aid of equation (1), which gives the height φ_0 of the potential barrier governing the resistance, as a function of the dielectric constant ε [8]. Since the dielectric constant of $BaTiO_3$ above T_C decreases in accordance with the Curie-Weiss law, $\varepsilon = K/(T - T_C)$, it follows that φ_0/kT increases with temperature and the resistivity increases as given by equation (2). Upon a further rise of temperature, however, the number of electrons in the surface states S will decrease as a result of thermal ionization, leading to a lower φ_0 and thus lowering the resistivity. This model gives at least a reasonable qualitative description of the temperature behaviour of the resistivity [8].

Characteristics and technical applications

Semiconducting titanates have become known by various names. Here we shall use the most commonly used term "PTC resistor" (positive temperature coefficient) or PTC thermistor.

The essential properties of the PTC resistor are described by the following characteristic curves:

a) the resistivity-temperature curve, or R-T characteristic (fig. 1),
b) the current-voltage characteristic, and
c) the current-time characteristic.

We shall now consider these characteristics in more detail, with particular attention to the latter two.

Another important characteristic property is the heat dissipation to outside media. This depends on the geometry, on the lead-in wires and on the thermal conductivity of the material. In the working range of these thermistors the heat dissipation can be regarded as a constant. Its value indicates the electrical power required to maintain the steady-state temperature in the semiconductor in undisturbed air at 1° above the ambient temperature. This dissipation constant D is between 10 and 20 mW/degree in typical examples.

Static current-voltage characteristic of semiconducting $BaTiO_3$

When a PTC resistor is electrically loaded, its temperature rises through the generation of Joule heat. For practical applications it is therefore useful to plot the static current-voltage characteristic, i.e. the curve obtained when the steady-state temperature is reached for each measuring point. Semiconducting BaTiO$_3$ with a Curie temperature of 120 °C is particularly suited for switching purposes. An ideal characteristic for this semiconductor, in which the resistivity remained independent of temperature between 20 and 120 °C and rose abruptly by several powers of ten in a narrow temperature interval near 120 °C, if transferred to the static current-voltage diagram would result in the characteristic shown in *fig. 9.*

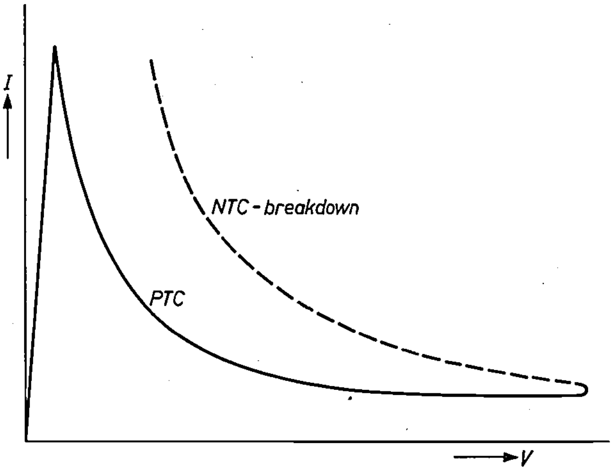


Fig. 9. Current-voltage characteristic of a BaTiO$_3$ positive-temperature coefficient (PTC) resistor.

The curve first rises linearly to a point corresponding to the power at which the resistor is heated from the ambient temperature T_u to 120 °C. The current maximum

$$I_{\max} = \frac{D(120 - T_u)}{V} \quad . \; . \; . \; . \quad (3)$$

has now been reached, and a further voltage increase turns the characteristic into a hyperbola:

$$VI = D(120 - T_u) = \text{constant} \quad . \; . \; . \quad (4)$$

The reason for this is that, when the applied voltage is increased in this range, the PTC resistor becomes so highly resistive that it opposes any further power consumption, and hence any further rise in temperature.

Since the increase of resistance does not in reality take place at exactly 120 °C but between 110 and 180 °C, the descending characteristic differs from the hyperbolic form. At the end of the current-voltage characteristic the dissipation is somewhat higher, $D(180 - T_u)$, than in the neighbourhood of the current maximum, $D(110 - T_u)$. Moreover, as can be seen from *fig. 10*, one obtains from eqs. (3) and (4) for various ambient temperatures a set of roughly hyperbolic curves starting from the initial ohmic portion of the characteristic.

The PTC behaviour is limited to a particular temperature range. Above that range the resistors again show an NTC characteristic, so that at $V_{\max}$ thermal breakdown occurs, as indicated by the dashed curve in fig. 9.

The maximum permissible voltage can be calculated if the maximum resistance $R_{\max}$ and the dissipation $D(180 - T_u)$ are known. For $R_{\max}$ under load, however, one can no longer take 10^4 times the value of the cold resistance $R_{\min}$ but only 10^2 to 10^3 times that value. The dielectric strength is therefore not as high as one might expect from the *R-T* characteristic. This is easily understandable, however, if we refer to the model of the barrier layers between the grains. If there are about 100 grain boundaries in the path of the current and the voltage across the ceramic semiconductor is 100 V, there will be a potential drop of 1 volt at each barrier layer. A barrier layer of this kind is probably less than 1 micron thick, so that the field strength there would be close to the disruptive electric field strength for non-conducting BaTiO$_3$. Consequently, the resistance values found at lower voltages can no longer be used. This model also implies that fine-grained ceramic BaTiO$_3$ semiconductors should have a higher dielectric strength than those with a coarser grain. PTC resistors of BaTiO$_3$ can already be made that have a cold resistivity of 100 ohm-cm and are capable of withstanding 600 V per cm. Experience indicates that the maximum voltage that can be applied is given by:

$$V^2_{\max} = 10^2 R_{\min} D(180 - T_u). \quad . \; . \; . \quad (5)$$

The equivalent circuit representing the essential properties of the PTC resistor therefore consists of three elements in parallel; a PTC resistor following its *R-T* characteristic, a voltage-dependent resistor that limits the dielectric strength, and a capacitor (with a capacitance up to several nanofarads) which constitutes the capacitive shunt at higher frequencies (see fig. 6).

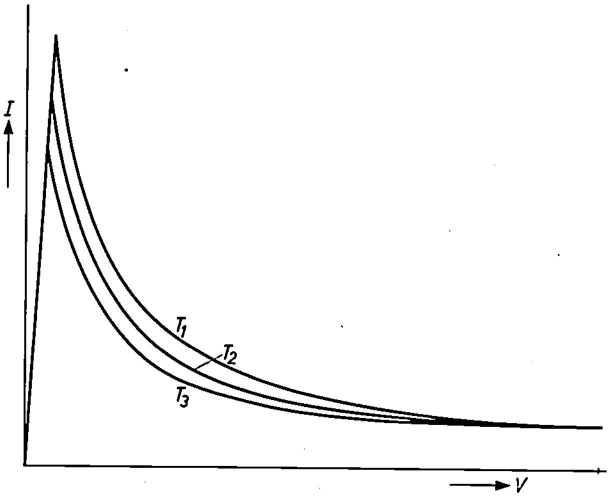


Fig. 10. Temperature dependence of the *V-I* characteristic of a PTC resistor; $T_3 > T_2 > T_1$.

The current-time characteristic

Since the current-voltage characteristic in fig. 9 results from the temperature behaviour of the PTC resistor, it can only be obtained by waiting for the static current value to settle for every value of applied voltage. Only when this static value is reached is there equilibrium between the power $V^2/R(T)$ converted in the semiconductor and the dissipated power $D(T - T_u)$. The greater the initial electric power, the faster will the PTC resistor be heated up to high resistance values. *Fig. 11* shows the current-time characteristics for various starting currents, set by using appropriate series resistances. Similar characteristics are obtained in charging-up a capacitor. For this reason a PTC resistor used in special circuits at low frequencies (<1 c/s) can act as a very high capacitance. In this case, of course, the electrical energy is not stored but dissipated as heat. The marked influence of the ambient temperature and the difference between heating up and cooling down times must also be taken into account. The comparison with the capacitor thus arises from the dynamic current-voltage characteristic of this semiconductor at very low frequencies, and has nothing to do with the real capacitive shunting of the barrier layers at high frequencies.

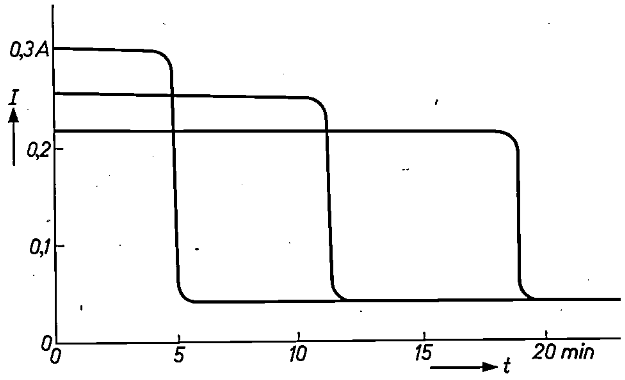


Fig. 11. Current-time characteristics of a $BaTiO_3$ semiconductor load for various initial currents.

Similar considerations apply to the NTC resistor, which at low frequencies may be compared with an inductance, its characteristics being the converse of the PTC characteristics. By combining these two resistors, circuits with interesting functions can be produced.

Applications relating to the temperature characteristic

PTC resistors can be used in many technical applications where NTC resistors are employed as temperature dependent devices. New applications will mainly be found, however, where the special characteristics of the PTC resistor come into their own. The following are some examples:

a) The positive temperature coefficient can be useful in "fail-safe" circuits, i.e. as switching elements that still operate in the event of the failure of a temperature control device, e.g. a contact break (R = ∞).

b) The steepness of the *R-T* characteristic allows either a more sensitive measurement or less amplification of the measured value.

c) The region in which the resistance is virtually independent of temperature (from 20 to 100 °C for $BaTiO_3$) can be useful for controlling electronic amplifiers. It may sometimes be required, for example, that the operating point of a driving circuits hould not be affected by certain fluctuations of the ambient temperature, but only if a preset temperature is exceeded. Semiconducting titanate, with its abrupt change of resistivity near the Curie temperature, can meet such requirements.

Applications relating to the current-voltage characteristic

Fig. 12 shows the current-voltage characteristic of a PTC resistor and the load line of a resistive load connected in series with it, at a voltage V_a. Because of the non-linearity of the PTC characteristic, three operating points are possible: P_1 and P_2 are stable operating points, and P_3 is unstable. Point P_3 is important for the generation of oscillations. The instability can be derived from *fig. 13*. In P_3 the thermistor has a resistance value corresponding to a load line through P_3 and the zero point. During voltage fluctuations or variations in heat dissipation, however, the resistance may provisionally assume a different value x. This gives the point of intersection P_x where the voltage division is such that the PTC resistor must acquire the value y, and so on. Thus, where there is a small deviation from P_3, the operating point shifts in the direction of P_1 or P_2.

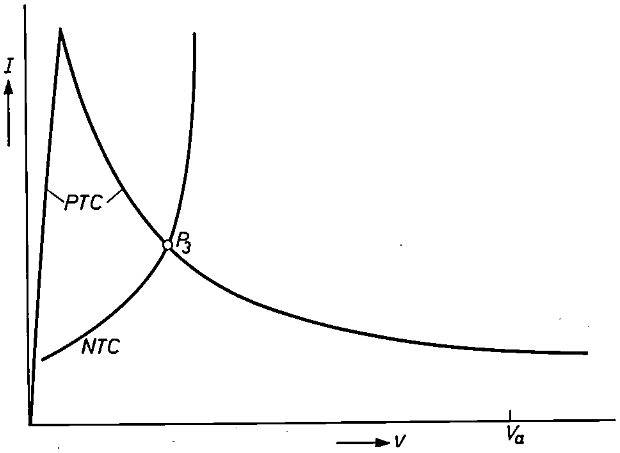


Fig. 12. Operating points P_1, P_2, P_3 of a PTC resistor and ohmic connected in series.

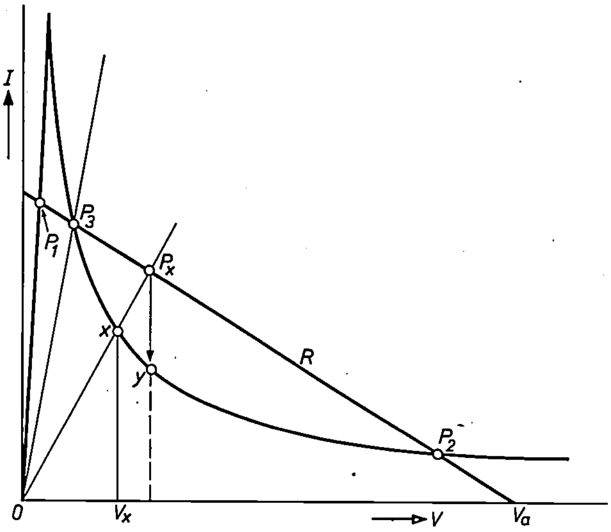


Fig. 13. Illustrating the instability of point P_3 in the characteristic of fig. 12.

When the voltage V_a is applied to a series circuit of this kind, the first operating point established is P_1. There are now three possible ways of shifting this point to P_2, the high-ohmic value of the PTC resistor:

1) By increasing V_a (*fig. 14a*),
2) By raising the ambient temperature (fig. 14*b*),
3) By reducing the load resistance (fig. 14*c*).

As regards fig. 14*a*, we can take as an example small motors required to operate at 110 and 220 V. Such motors are often rated only for 110 V, and if connected to 220 V they must be protected by a series resistance to drop half the voltage. Automatic adjustment to the mains voltage in such a case can be effected by a PTC thermistor shunted across the series resistor. At 110 V the thermistor should operate at P_1, i.e. its resistance should be low, while at 220 V its operating point should shift to P_2, thus no longer shunting the series resistance.

As an example to fig. 14*b* we can take a small electric motor which, enclosed in its small housing, can gradually get too hot under heavy loading. A PTC resistor connected in series and in thermal contact with the motor would prevent overheating. Owing to the thermal contact, the heat dissipation of the PTC resistor deteriorates as the motor gets hotter; the *V-I* characteristic changes (dashed curve in fig. 14*b*) and there is then only one stable operating point left, in the high-resistance portion of the PTC characteristic. At that point only one-tenth of the initial current flows, and moreover almost the entire applied voltage is dropped across the PTC resistor, which has now become a high resistance.

An example to fig. 14*c* is a load that changes its resistance during operation. Cases in point are electronic valves or an electric motor whose impedance decreases under a braking load. The resultant high current can lead to serious damage due to overloading. A suitably dimensioned PTC resistor in series with the motor is heated up by the excessive current, and at the operating point P_2 reduces the power in the load to less than one-hundredth.

In an automatic grid bias circuit a PTC resistor in the cathode lead of a valve can provide effective overload protection. Since it becomes a high resistance when the current is excessive, it drops correspondingly more voltage, so that the cathode goes more positive and cuts off the valve. In some cases it is even possible to do without the bypass capacitor, since a capacitive shunt is already present at the barrier layer in the semiconductor.

The behaviour of a *series* arrangement of two temperature-dependent resistors under load is easily seen from a diagram as in *fig. 15*, which relates to two PTC resistors in series. There are again three operating points, the middle one being unstable for reasons similar to those considered in connection with fig. 13. Only one of the two resistors is thus heated up to high resistance values. The other, owing to the low residual current,

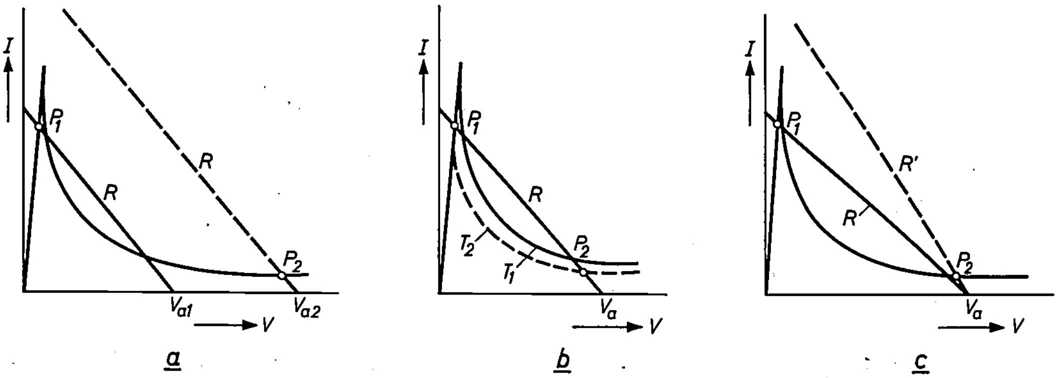


Fig. 14. In a series arrangement of PTC resistor and load the operating point P_1 can be shifted to P_2: *a*) by increasing the voltage, *b*) by raising the temperature (T_2>T_1), *c*) by reducing the load resistance.

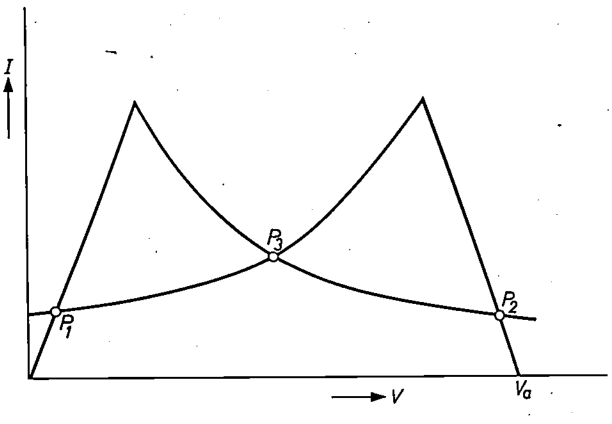


Fig. 15. Illustrating the behaviour of two PTC resistors connected in series.

cannot get hot. This demonstrates that it is pointless to connect several PTC resistors in series with a view to operation at higher voltages. In such a case a uniform voltage distribution would have to be enforced by a parallel arrangement of stabilizing resistors (VDR).

It is, however, possible to operate PTC resistors *in parallel* under load. They are uniformly heated and, when graphically represented, the corresponding current values can be added for every voltage value.

The behaviour of PTC resistors in series, as just described, can be utilized for the generation of relaxation oscillations of very low frequency; see *fig. 16*. The NTC resistors in this bridge circuit are too highly resistive to heat up at $V_a/2$. Since only one PTC resistor in the other branch can have a high resistance value, this resistor and the NTC resistor parallel to it carry nearly the entire voltage V_a. This may suffice to heat the NTC resistor to the thermal breakdown point. The steeply rising current through this resistor then bypasses the hot PTC resistor and now heats up the cold one, and so on. The result is a self-oscillating multivibrator with a frequency lower than 0.1 c/s.

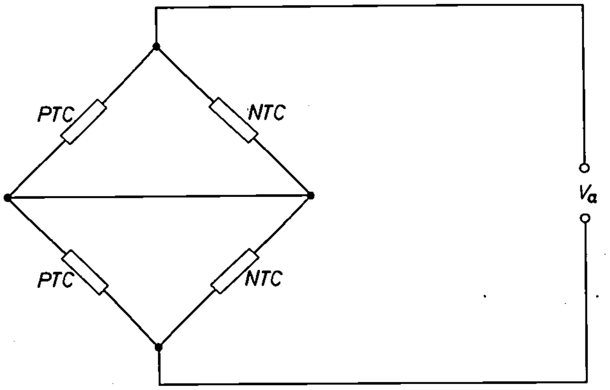


Fig. 16. Bridge circuit consisting of pairs of PTC and NTC resistors, operating as a multivibrator.

A self-excited oscillator can also be produced by connecting a low-ohmic PTC resistor in series with a high-ohmic NTC resistor. *Fig. 17* shows that the two descending characteristics have a point of intersection. When this point is reached, it may be unstable at certain values of V_a; it then corresponds to point P_3 in figs. 12 and 13. In such a case no other operating point is possible, and thus the circuit oscillates. At the very low oscillating frequencies of this arrangement, less than 1 c/s, the PTC resistor acts as a capacitance and the NTC resistor as an inductance. The role of the generator is thereby taken over by the negative resistance of one of the circuit elements.

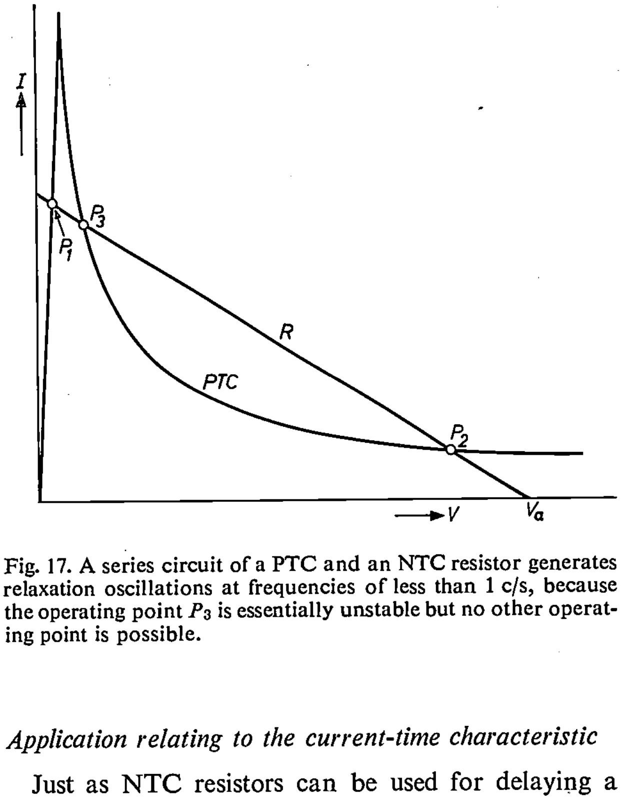


Fig. 17. A series circuit of a PTC and an NTC resistor generates relaxation oscillations at frequencies of less than 1 c/s, because the operating point P_3 is essentially unstable but no other operating point is possible.

Application relating to the current-time characteristic

Just as NTC resistors can be used for delaying a switching-on process, PTC resistors can be used for slowing down a switching-off process. To protect the windings of fast electric motors used in domestic appliances a maximum operating period of a few minutes is often specified. A suitable PTC resistor in series with the load automatically switches off the motor at the preset time. Under increasing braking load the temperature of the resistor rises faster as the motor winding becomes hotter. The switching time of the PTC resistor also depends, of course, on the ambient temperature, in the same way as the heating of the winding. This dependence in the present case is thus an advantage.

Direct switching with PTC resistors is in general only useful in the case of a power consumption under 100 W. Where higher powers are involved, it is better to use a relay. Furthermore, it should be remembered that only the switch-over from P_1 to P_2 (in fig. 12) takes place with any degree of sensitivity. Switching back to P_1, however, presents difficulties. For this purpose the load line should intersect the PTC characteristic at only one point, which should be in the linear part of the characteristic. This is possible only at low voltages V_a (fig. 14*a*), at fairly low temperatures (fig. 14*b*) or when the load resistance becomes very high (fig. 14*c*).

Summary. Barium titanate has hitherto mainly been known and employed technically for its dielectric properties. By suitable doping, however, semiconducting $BaTiO_3$ can be produced which, in the form of a polycrystalline ceramic, exhibits a non-linear resistance-temperature characteristic that rises steeply at the ferro-electric Curie point. The position of the ferro-electric Curie point can be chosen between —90 °C and +400 °C by altering the mixed-crystal composition. The steep increase in resistivity can be explained in terms of the energy band model from the existence of grain boundaries of very high resistance in the polycrystalline ceramic. PTC (positive temperature coefficient resistors) made from semiconducting $BaTiO_3$ ceramic are interesting for a wide variety of applications, e.g. for overload protection, for automatically limiting the operating period of small electric motors, or as a simple means of producing relaxation oscillations of very low frequency.

Investigations on the germanium-electrolyte interface

H. U. Harten, R. Memming and G. Schwandt

541.183:546.289

It is necessary to have sufficient information about the bulk properties of a crystal before starting to investigate the properties of its surface. These experiments are difficult because "surface effects" may be greatly influenced by very small amounts of water vapour, oxygen or even by grease from the fingers; briefly, by any kind of impurity.

Up to the present the best single crystals can be made from germanium or silicon. Therefore these two semiconductors are mainly used for investigations of the surface [1]. Moreover these investigations are not only of scientific but also of technical interest: the aging of transistors for example can be connected with processes occurring at the surface.

Since impurities have a strong influence on the surface properties it was attempted to get the surface as clean as possible. Efforts in this direction were made by cleaning the crystal under extremely high vacuum. In this way it was hoped to obtain surfaces as illustrated schematically in *fig. 1a*: the atoms of the outermost lattice plane have no longer any neighbours on one side and their free valences reach into space. If foreign atoms are present there, then they will be absorbed or chemisorbed (fig. 1*b*) — but also if they are absent the

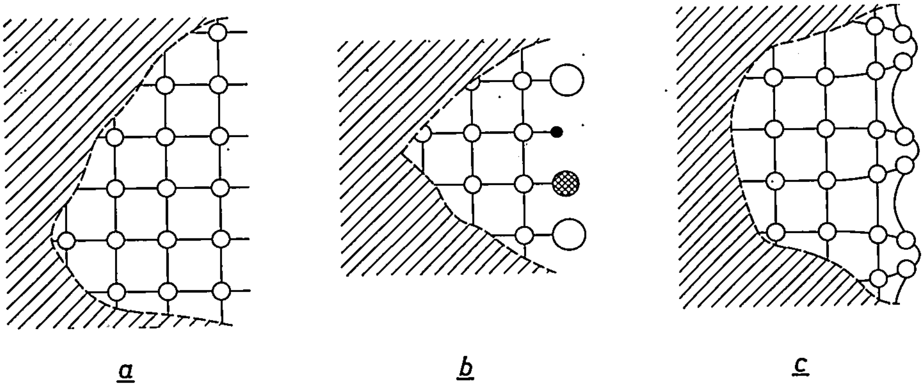


Fig. 1. Schematic representation of the surface of a solid.
a) A "clean" surface. *b*) A surface where foreign atoms have been adsorbed. *c*) A clean surface where the "free" valences are saturated by mutual bondings.

Dr. R. Memming and Dr. G. Schwandt are research workers at the Hamburg laboratory of Philips Zentrallaboratorium GmbH. Prof. Dr. H. U. Harten, a former member of the same laboratory, is now professor of experimental physics at the University of Göttingen.

[1] Recent survey articles: G. Heiland, Fortschr. Physik **9**, 393 1961; H. Flietner, Physica status solidi **2**, 221, 1962; see also the papers by G. Heiland and H. U. Harten in Festkörperprobleme III (edited by F. Sauter), Vieweg, Brunswick 1964.

the switch-over from P_1 to P_2 (in fig. 12) takes place with any degree of sensitivity. Switching back to P_1, however, presents difficulties. For this purpose the load line should intersect the PTC characteristic at only one point, which should be in the linear part of the characteristic. This is possible only at low voltages V_a (fig. 14*a*), at fairly low temperatures (fig. 14*b*) or when the load resistance becomes very high (fig. 14*c*).

Summary. Barium titanate has hitherto mainly been known and employed technically for its dielectric properties. By suitable doping, however, semiconducting $BaTiO_3$ can be produced which, in the form of a polycrystalline ceramic, exhibits a non-linear resistance-temperature characteristic that rises steeply at the ferro-electric Curie point. The position of the ferro-electric Curie point can be chosen between —90 °C and +400 °C by altering the mixed-crystal composition. The steep increase in resistivity can be explained in terms of the energy band model from the existence of grain boundaries of very high resistance in the polycrystalline ceramic. PTC (positive ,temperature coefficient resistors) made from semiconducting $BaTiO_3$ ceramic are interesting for a wide variety of applications, e.g. for overload protection, for automatically limiting the operating period of small electric motors, or as a simple means of producing relaxation oscillations of very low frequency.

Investigations on the germanium-electrolyte interface

H. U. Harten, R. Memming and G. Schwandt

541.183:546.289

It is necessary to have sufficient information about the bulk properties of a crystal before starting to investigate the properties of its surface. These experiments are difficult because "surface effects" may be greatly influenced by very small amounts of water vapour, oxygen or even by grease from the fingers; briefly, by any kind of impurity.

Up to the present the best single crystals can be made from germanium or silicon. Therefore these two semiconductors are mainly used for investigations of the surface [1]. Moreover these investigations are not only of scientific but also of technical interest: the aging of transistors for example can be connected with processes occurring at the surface.

Since impurities have a strong influence on the surface properties it was attempted to get the surface as clean as possible. Efforts in this direction were made by cleaning the crystal under extremely high vacuum. In this way it was hoped to obtain surfaces as illustrated schematically in *fig. 1a*: the atoms of the outermost lattice plane have no longer any neighbours on one side and their free valences reach into space. If foreign atoms are present there, then they will be absorbed or chemisorbed (fig. 1*b*) — but also if they are absent the

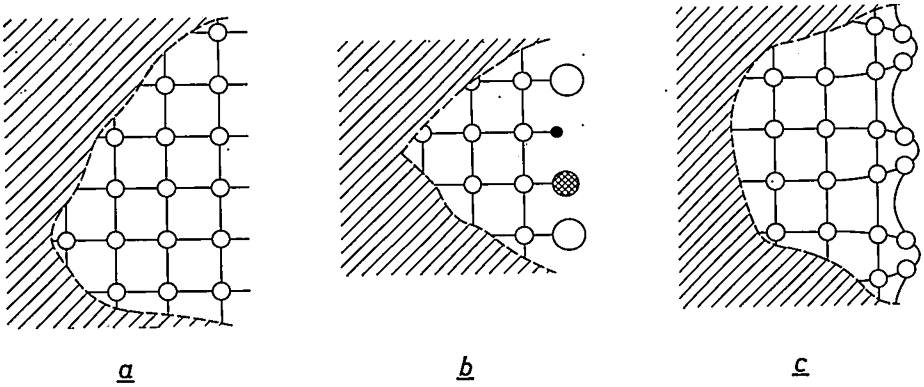


Fig. 1. Schematic representation of the surface of a solid.
a) A "clean" surface. *b*) A surface where foreign atoms have been adsorbed. *c*) A clean surface where the "free" valences are saturated by mutual bondings.

Dr. R. Memming and Dr. G. Schwandt are research workers at the Hamburg laboratory of Philips Zentrallaboratorium GmbH. Prof. Dr. H. U. Harten, a former member of the same laboratory, is now professor of experimental physics at the University of Göttingen.

[1] Recent survey articles: G. Heiland, Fortschr. Physik **9**, 393 1961; H. Flietner, Physica status solidi **2**, 221, 1962; see also the papers by G. Heiland and H. U. Harten in Festkörperprobleme III (edited by F. Sauter), Vieweg, Brunswick 1964.

free valences do not in general remain free; they are saturated by mutual bonding. The normal valence directions are not "suitable" for this mutual bonding, and the lattice is deformed at the surface (fig. 1*c*), i.e. the "clean" surface is not free from disturbances as was originally expected [2].

From our point of view only one consequence of this deformation is of importance: electrons which are mobile in the bulk of the crystal can be quite strongly bound in the distorted lattice planes. In the band model this situation leads to the appearance of additional energy levels for electrons at the surface — so called "surface states" — within the forbidden gap (*fig. 2*).

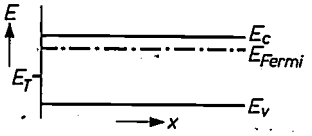


Fig. 2. Schematic representation of surface states in the band model. The ordinate represents the electron energy, and the abscissa the space coordinate. E_c is the lower edge of the conduction band, E_{Fermi} the Fermi level, E_v the top of the valence band, E_T the level of a surface state.

One can distinguish between different types of surface state, which among other things differ in the distance between their energy levels and the band edges. In this picture the "simplest" surface would be a surface with no surface states at all. On a suggestion of Brattain and Boddy we will call such a surface "perfect" [3]. Brattain and Boddy were also the first to make perfect surfaces — however, only "under water", i.e. in contact with an aqueous electrolyte.

Before trying to determine whether a surface is perfect, one must first know how surface properties are influenced by the presence of surface states. If their energy level lies below the Fermi level as indicated in fig. 2, then they are occupied by electrons. These electrons originating in the bulk itself form a negative "surface charge" denoted by Q_F. Since neutrality must be conserved throughout the crystal a positive charge is built up just below the surface. Its depth is about one micron for germanium. The existence of such a space charge layer leads to "band bending" in the band model (*fig. 3a*). The corresponding potential drop across this layer is denoted by U_R. Integrating over the whole region the total space charge may also be described by a surface charge. In our example the charge is positive, and the density per unit area will be denoted by Q_R. Since all electrons trapped in the surface states originate in the space charge region then:

$$Q_R + Q_F = 0.$$

Fig. 3*b* shows the atomic model of this situation. Finally in fig. 3*c* the distribution of the electrostatic potential is shown in a rather simplified way. The complications which undoubtedly arise as a result of the existence of atomic dipole layers at the surface will be ignored here.

If we now change the band bending by some external means then the charge Q_R in the space charge region will also be changed. Furthermore it is conceivable that by these means levels of surface states will be raised above the Fermi level — their distances from band edges being fixed. The surface states must then release their electron, i.e. Q_F is also changed. Finally, if it is possible to measure the change of the two charges separately, one could not only prove the existence of surface states but also determine the position of their energy levels within the band gap.

There are a number of possibilities for this. One possibility is to measure the capacity of a germanium electrode which is in contact with an electrolyte. The basic experimental arrangement is illustrated in *fig. 4*. The germanium electrode is dipped into a glass cell filled with water, in which e.g. 0.1 M KNO_3 is dissolved, in order to give it a certain conductivity, and to which a buffer is added in order to keep the *p*H at a constant value. The electrode is connected by means of a variable d.c. voltage source *U* to a second electrode (counter electrode) of platinum *Pt*, which must have a large surface area (platinized platinum gauze). Close to the germanium surface is placed a standard (e.g. calomel)

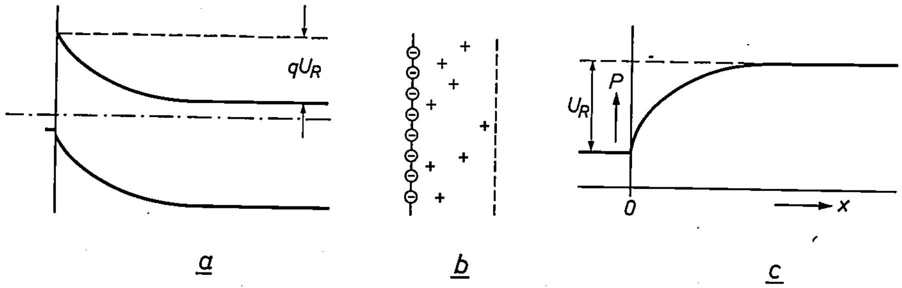


Fig. 3. *a*) Band bending as a result of surface states. *q* is elementary charge, U_R the voltage over the space charge layer. *b*) Atomic representation of the space charge region. *c*) The distribution of the electrostatic potential *P* of a clean surface (a simplified picture).

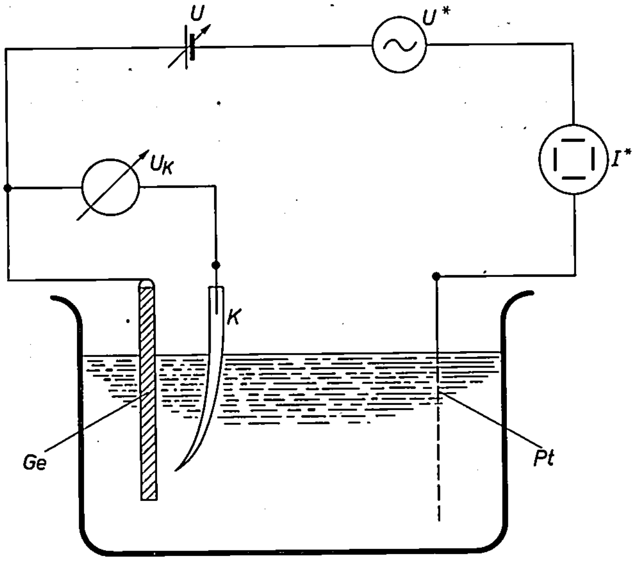


Fig. 4. Schematic representation of the experimental arrangement for determining the capacitance of the germanium electrode *Ge*. *Pt* is a platinum electrode; *K* a standard electrode (calomel); *U* a variable voltage source; U_K the electrode potential (reference voltage); U^* an alternating voltage source (100 kc/s, a few mV); and I^* is the alternating current, which is displayed on an oscilloscope.

electrode. The electrode potential U_K ("reference voltage") is measured between this and the germanium electrode. U_K can be adjusted to any value desired by changing U appropriately. Finally an a.c. voltage source U^* (frequency about 100 kc/s, a few millivolts) is included in the circuit. The amplitude of the resulting a.c. current I^* and the phase shift between current and voltage are displayed on an oscilloscope. The total impedance and the phase shift occurring can be interpreted by an equivalent circuit containing a resistance R and a capacitance C connected in series (whether this equivalent circuit is adequate, at least for a certain frequency range, must be checked for each case by testing the circuit at different frequencies [4]). While R is the total series resistance in the circuit we may equate the capacitance C to that of the germanium electrolyte interface alone. The capacitance of the counter electrode, which is in series with that of the germanium electrode, can be neglected since it is much larger because of the large surface area of the platinum gauze. Thus by means of this arrangement one can obtain the capacitance C of the germanium electrolyte interface as a function of the electrode potential U_K.

Before showing how it is possible to get information about surface states from such measurements we have to extend the model derived above for dry surfaces (fig. 3*b* and *c*) to semiconductor surfaces in contact with an electrolyte. The mobile carriers in the electrolyte are ions. In the example under consideration positive ions are attracted by the negatively charged surface states. It is to be expected, therefore, that some of them will be attached to the surface. In order to maintain charge neutrality the space charge in the semiconductor has to be reduced by an equal number of positive charges. The space charge region is thus thinner — although the thickness remains of the same order of magnitude (1 micron). The ions in the electrolyte are surrounded by a so-called solvation shell, and this means that their charge centre remains at a distance from the actual interface corresponding to the radius (several Å) of the solvation shell. The double layer which is formed in this manner in the electrolyte, the "Helmholtz layer", is only a thin layer in comparison with the space charge region in the semiconductor. Consequently, hardly any voltage is developed across the double layer although the charge density and the field strength are rather high. Hence one should expect a potential distribution across the interface as illustrated in a rather simplified way in *fig. 5*. This would mean that the "Galvani potential" U_G is more or less equal to the potential drop U_R across the space charge region. This conclusion is not quite correct because the potential distribution as presented in fig. 5 is simplified in a way which is not really justified. In fact the potential across the Helmholtz layer is not small but it is certainly constant. Consequently, a change ΔU_G in the Galvani potential — which is obtained by varying the external voltage U and which is measured as a change of the electrode potential U_K — appears identically at the space charge layer, i.e.:

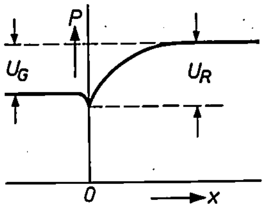


Fig. 5. The distribution of the electrostatic potential *P* of the semiconductor-electrolyte interface. U_G is the Galvani potential.

$$\Delta U_G = \Delta U_K = -\Delta U_R.$$

(The negative sign is simply the result of other conventions as regards the sign of U_K and U_R.) It still remains to be verified whether this equation is satisfied or not. If this is the case then the change of the band bending can be determined by measuring the change of the electrode potential. Furthermore the external a.c. voltage U^* — at least in a certain frequency range — will also be taken up completely by the space charge layer ($U^* = U_R^*$). This leads to a synchronous change Q_R^*

[2] See e.g. J. J. Lander and J. Morrison, J. appl. Phys. **34**, 1403, 1963.
[3] W. H. Brattain and P. J. Boddy, J. Electrochem. Soc. **109**, 574, 1962.
[4] R. Memming, Philips Res. Repts. **19**, 323, 1964.

in the space charge and possibly also in the charge in the surface states (Q_F*). These charges are compensated by corresponding counter-charges in the Helmholtz layer. The charge variations can be related to two differential capacities: the "space charge capacity" as defined by $C_R = Q_R^*/U_R^*$ and the "capacity of surface states" defined by $C_F = Q_F^*/U_R^*$. The sum of these two capacities can be obtained from measurements using an experimental arrangement as shown in fig. 4.

To evaluate such measurements it is necessary to know quantitatively the relationship between the band bending and the two capacities that follows from the model. This relationship can be derived qualitatively as follows.

We shall first examine the capacity of the surface states. A positive value of U_R means, by definition, that the conduction and valence bands are bent *down*wards at the surface. The downward band bending can be made so large that the surface states are shifted below the Fermi level and they are as a result occupied completely by electrons. Q_F then reaches its (negative) maximum value $Q_{F\,\max} = qn_T$ (n_T = number of surface states per unit area). Conversely the level of the surface states can be brought so far above the Fermi level by bending the bands sufficiently *up*wards (U_R negative) that Q_F is equal to zero. The charge change occurs in a relatively small potential range of U_R, namely in that interval where the surface states are quite close to the Fermi level, i.e. Q_F follows the Fermi function, (see *fig. 6*). The slope is equal to the capacity of the surface states. It reaches a maximum when the surface state passes the Fermi level ($U_R = U_{\rm Fermi}$). From the maximum of the capacity curve the maximum charge $Q_{F\,\max}$ in the surface states and hence the density of the surface states can be calculated.

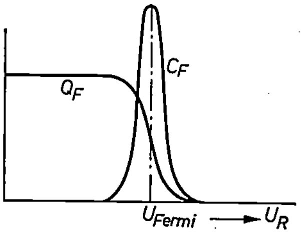


Fig. 6. Charge Q_F and capacity C_F of surface states as a function of the band bending (schematic). At $U_R = U_{\rm Fermi}$ the level of the surface states is equal to the Fermi level.

The derivation of the relationship between space charge capacity C_R and band bending is not so easily performed. We shall treat here only a very simple case, that of intrinsic germanium. The flat band condition ($U_R = 0$) means here that the density of electrons and holes is equal to the intrinsic carrier concentration n_i (*fig. 7*, dotted line) everywhere in the bulk as well as at the surface. By a downward band bending (U_R positive) the density of the electrons is increased at the surface and in the space charge region (upper curve in fig. 7) whereas in the bulk the electron density remains constant ($n = n_i$). The single-hatched area in fig. 7 is then a measure of the number ΔN of excess electrons in the space charge region per unit surface area. ΔN is positive and increases exponentially with U_R (*fig. 8a*). On the other hand the density of holes is decreased (not below zero, of course) by a downward band bending. ΔP, the number of all additional holes in the space charge region, is negative and its absolute value is relatively small (double-hatched area in fig. 7). If the downward band bending ($U_R > 0$) is changed to an upward band bending ($U_R < 0$) then the density of holes is increased and the density of electrons decreased (fig. 8*a*). The total charge per unit surface area in the space charge region is given by:

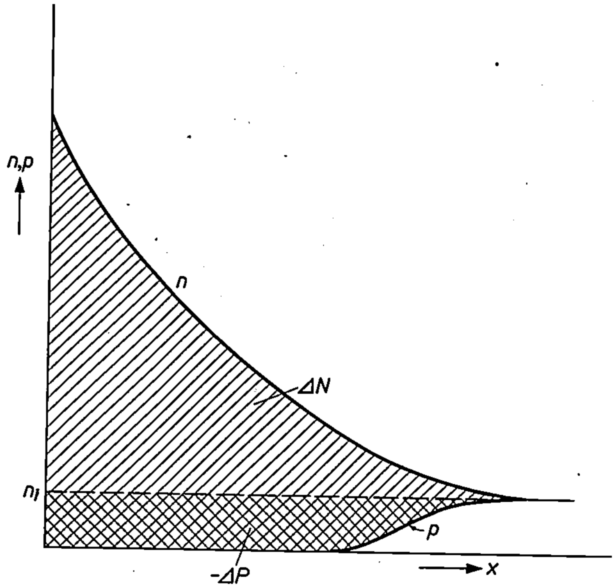


Fig. 7. Schematic representation of the concentration distribution of the electrons (n) and holes (p) in the space charge region of intrinsic germanium. n_i is the intrinsic concentration, x the space coordinate. The single-hatched area is a measure of the density ΔN of the excess electrons present; the double-hatched area is a measure of the density ΔP of the excess holes present.

$$Q_R = q\,(\Delta P - \Delta N),$$

where q represents the elementary charge and where the negative sign is a result of the negative electron charge. Using this equation and the relationship between excess carrier densities (ΔN, ΔP) and band bending (fig. 8*a*) one can plot the charge density Q_R versus band bending as shown in fig. 8*b*. This is an S-shaped curve with a point of inflection at $U_R = 0$. The slope of the curve

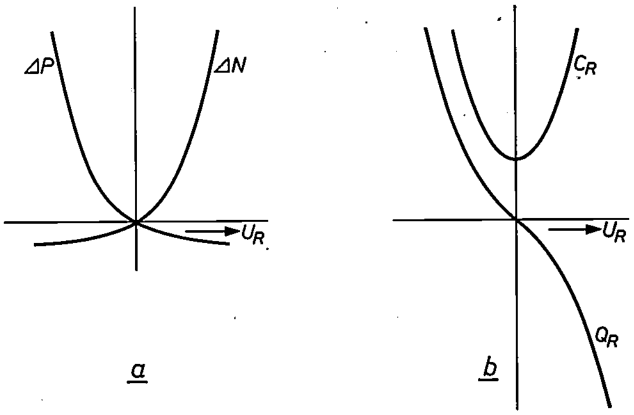


Fig. 8. *a*) The influence of the band bending U_{R} on the excess carrier density ΔN and ΔP in the space charge region.
b) The charge Q_{R} and the capacity C_{R} of the space charge layer as a function of the band bending U_{R} for intrinsic germanium (schematic).

is equal to the space charge capacity C_{R} which has a minimum at the flat band condition ($U_{\mathrm{R}} = 0$).

From figures 6 and 8*b* it may be deduced that the space charge Q_{R}, and the charge in the surface states Q_{F}, both change in the same direction; i.e. the total capacity is given by the sum of the two capacities ($C = C_{\mathrm{R}} + C_{\mathrm{F}}$). If it is found that the measured values agree with the theoretical values of the space charge capacity, then the surface is called "perfect". Strictly speaking this conclusion is not quite correct because it is in principle possible that surface states cannot be detected by capacity measurements.

A typical capacity curve obtained with intrinsic germanium is presented in *fig. 9*. The experimental capacity values are plotted versus the electrode potential U_{K} (upper scale), the theoretical curve versus the band bending U_{R} (lower scale). The initial (and uninteresting) difference between both potentials is determined by shifting the scales to obtain the best fit between the experimental values and the theoretical curve. For this purpose the theoretical curve was also displaced in the *vertical* direction. This means a multiplication of the theoretical values on the logarithmic scale, in this case by a factor of 1.3. This factor is interpreted as a "roughness factor", i.e. it is assumed that the surface area of the electrode is slightly larger than the value given by the edge to edge dimensions.

From the fact that in this case (fig. 9) the experimental and theoretical capacity curves agree so well one may conclude that:

1) changes in the external applied voltage were completely reproduced across the space charge region ($\Delta U_{\mathrm{K}} = -\Delta U_{\mathrm{R}}$),

2) the surface was perfect ($C_{\mathrm{F}} = 0$).

The method of obtaining perfect germanium surfaces turns out to be quite simple. Several layers of the crystal are dissolved by polarizing a germanium electrode anodically in a neutral or alkaline aqueous solution of sufficient purity. After stopping the anodic polarization the germanium surface is perfect for some seconds, minutes or even hours — depending on the purity of the electrolyte.

It has been observed that an addition of copper ions into the electrolyte results in an increase of the capacity (*fig. 10*). Apparently surface states are formed and

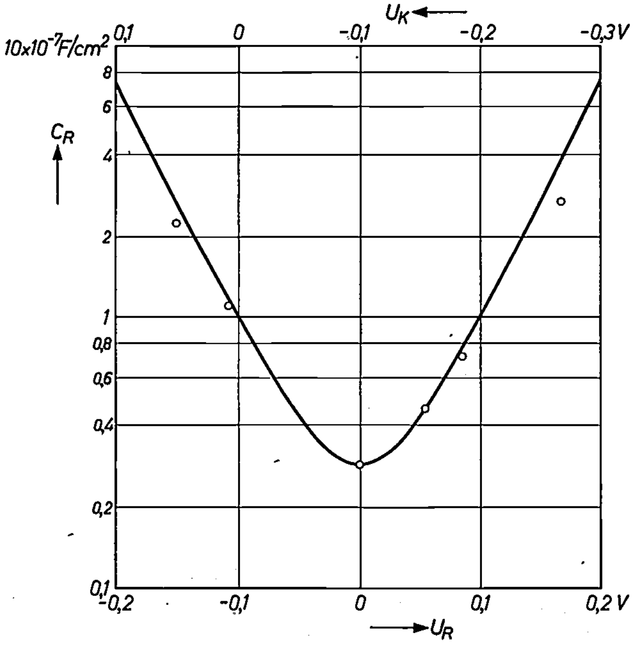


Fig. 9. The capacity of an intrinsic germanium electrode [4]. The experimental values are plotted against the electrode potential U_{K}. The curve gives the calculated capacity of the space charge layer, C_{R}, as a function of the band bending U_{R} (after R. Memming [4]).

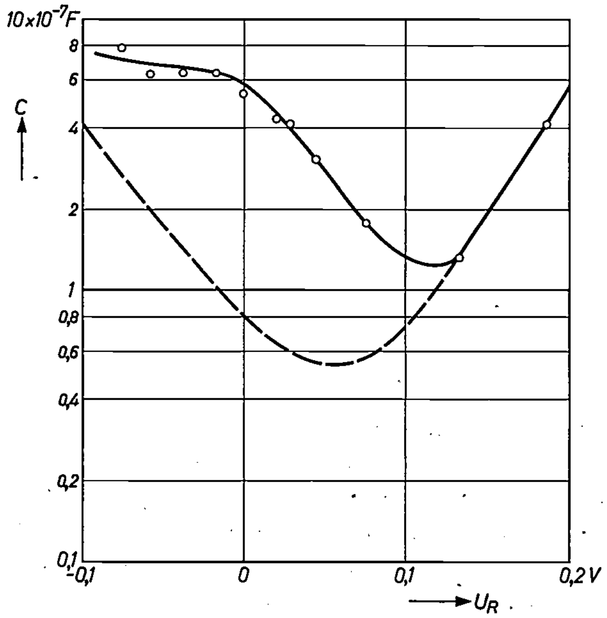


Fig. 10. The capacity C of a germanium electrode in an electrolyte in which 5×10^{-7} mole $\mathrm{Cu(NO_3)_2}$ per litre is dissolved (full line). The dashed line gives the capacity of a "perfect" germanium electrode (fig. 9). (After P. J. Boddy and W. R. Brattain [5].)

these are of two types with different energy levels (*fig. 11*) [5]. Similar surface states are also produced by adding silver or gold ions to the electrolyte. The position of the energy level (surface states) within the band gap is specific for the metal ions whereas the density of surface states is always of the order of 10^{11} cm^{-2}. This corresponds to one surface state per several thousand germanium atoms at the surface. The number of ions needed to reach this density is exceedingly small: a few micrograms per litre are quite sufficient. Further investigations [5] have shown that one of the two types of surface state (the one with the lower capacity maximum in fig. 11) is a "recombination centre". Electrons can drop from the conduction into the valence band via this centre, i.e. electron hole pairs recombine. The other surface state (fig. 11), however, is a trap, i.e. only one type of carrier can be captured by it, trapped for some time and then released again. Whether it is a trap for electrons or holes cannot be deduced from measurements performed with intrinsic germanium.

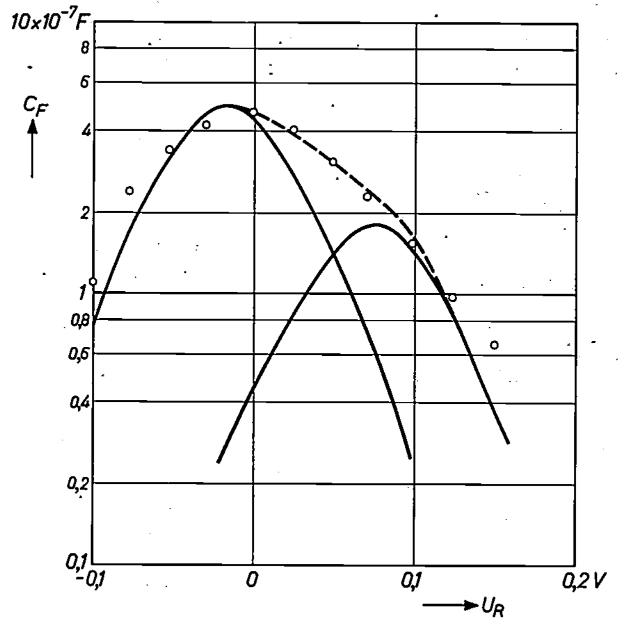


Fig. 11. The additional capacity C_F due to Cu-centres at the surface of a germanium electrode.

Information about the electrical properties of traps can only be obtained from measurements performed with *N*-type germanium. Before discussing these experiments it is necessary to consider more deeply the electrochemical behaviour of germanium, especially anodic dissolution.

It is known that germanium is dissolved in its tetravalent state and that germanate ions are formed in alkaline solutions according to the equation [6]:

$$G^{4+} + 6\,OH^- \rightarrow GeO_3^{2-} + 3\,H_2O.$$

The dissolution rate can easily be determined by measuring the current crossing the interface. This current increases when the electrode is polarized anodically (U_K changed in positive direction). As shown in *fig. 12* (lower curve) the current reaches a saturation value at a certain electrode potential. The fact that at higher anodic potentials the current increases again as a result of a new mechanism does not interest us in this respect. Such a saturation of the current is generally explained by assuming that the supply of ions involved in the corresponding reaction is not sufficiently fast. In the present case this can be the supply of Ge-ions or of OH-ions. Since germanium is naturally present in a sufficient quantity at the electrode surface, it seems probable that OH-ions are the cause. During the anodic dissolution of germanium the OH-ion concentration at the interface is lowered so that a concentration gradient arises in the electrolyte, and this results in a diffusion current which supplies new OH-ions. This diffusion is then the rate determining factor for the dissolution. Since the OH-ion concentration cannot drop below zero the concentration gradient reaches a maximum, i.e. the diffusion is limited.

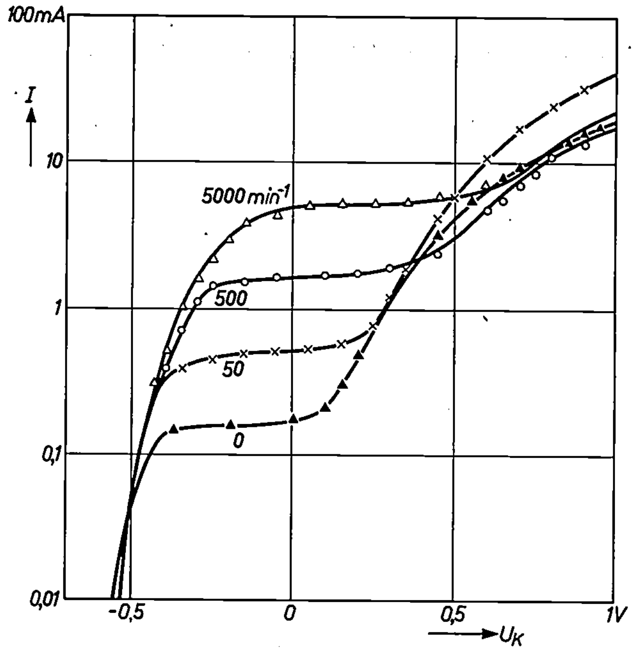


Fig. 12. Current-voltage characteristic during anodic dissolution of *P*-type germanium. The parameter is the revolution speed of the rotating electrode. (After H. G. Schmidt and G. Schwandt [7].)

On the other hand, the supply of ions can be enhanced by convection. Accordingly, one would expect the saturation current to be increased by stirring the solution. One could do this in a simple and effective manner using a "rotating electrode". This is a cylindrical electrode disc rotating around its main axis. *Fig. 13*

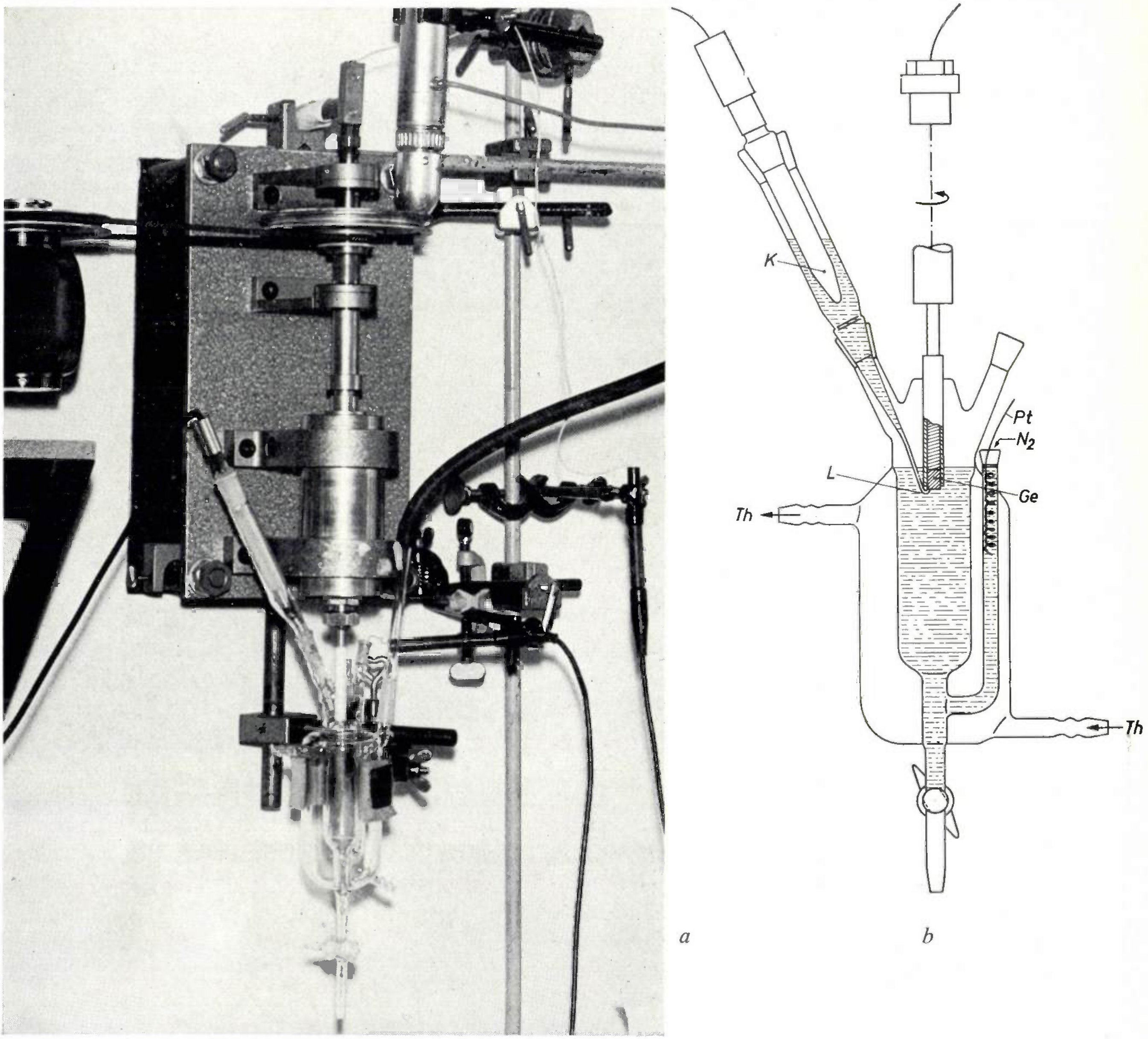


Fig. 13. Experimental arrangement for the rotating electrode: *a*) General view. *b*) Cross-sectional view of the cell. *Ge* is the rotating electrode, *Pt* the platinum electrode, and *K* the calomel electrode with Luggin capillary *L*. *Th* is the connection for the thermostat.

illustrates the arrangement used by us: Theoretically the saturation current may be expected to increase with the square root of the revolution speed of the electrode disc [8]. As shown in fig. 12 and *fig. 14* [7] the dissolution rate of germanium is indeed enhanced and the saturation current depends on the revolution speed as predicted by the theory. The model with which we attempt to explain these phenomena thus appears to be very satisfactory, at least for *P*-type germanium. In the

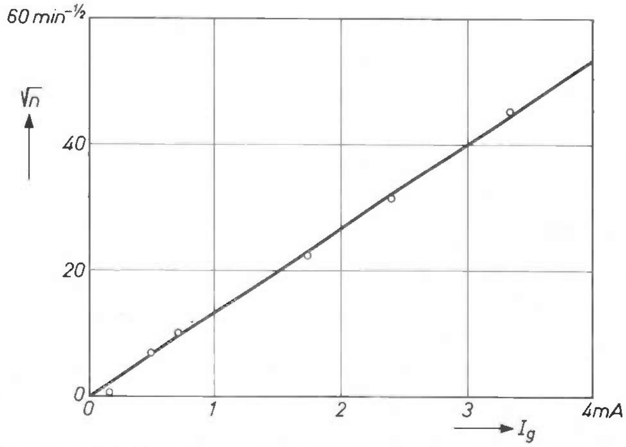


Fig. 14. The saturation current I_g as a function of the square root of the revolution speed *n* of the rotating electrode. Other data as in fig. 12.

[5] P. J. Boddy and W. H. Brattain, J. Electrochem. Soc. **109**, 812, 1962.

[6] F. Jirsa, Z. anorg. allgem. Chemie **268**, 84, 1952.

[7] H. G. Schmidt and G. Schwandt, Phys. Verhandlg. **14**, 44, 1963 (No. 1/2).

[8] B. Lewitsch, Acta physicochimica USSR **17**, 257, 1942.

case of N-type germanium, however, the stirring does not affect the dissolution current; moreover N-type germanium is dissolved much more slowly (*fig. 15*) [9]. This behaviour, of course, cannot be connected with the OH-ion diffusion.

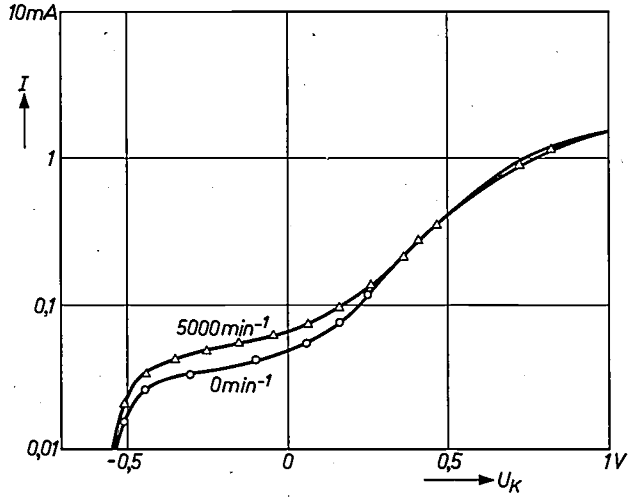


Fig. 15. Anodic dissolution of N-type germanium. The parameter as in fig. 12 is the revolution speed of the rotating electrode. (After G. Schwandt and M. Scheer [9].)

It is evident here we were somewhat premature in our assumption that germanium at the interface is always present in a sufficient quantity. Germanium *atoms* are indeed available; for the reaction, however, *ions* are necessary. In order to dissolve the germanium (ionization of surface atoms) either electrons must be injected into the bulk or holes must be extracted. For germanium both processes occur: for the dissolution of one atom two electrons (e$^-$) are given up and two holes (e$^+$) are taken. The corresponding reaction equation for the dissolution of germanium in alkaline solutions can therefore be written as follows [10]:

$$\mathrm{Ge} + 2^+ + 6\,\mathrm{OH}^- \rightarrow \mathrm{GeO_3}^{2-} + 2\mathrm{e}^- + 3\,\mathrm{H_2O}.$$

Now, in P-type germanium sufficient holes are available for the formation of germanium-ions, and they do not influence the reaction rate. For N-type germanium the situation is quite different, as the number of holes is limited. Since holes are removed for the electrode reaction, their density at the surface is decreased, a concentration gradient is set up, but the corresponding diffusion is limited again. If the maximum diffusion current of holes is lower than that of the OH-ions then the hole diffusion determines the rate of the whole reaction. This explains why the anodic dissolution of N-type germanium is not accelerated by stirring the electrolyte.

On the other hand, the dissolution rate should be enhanced by supplying more holes. This can be achieved by illuminating the electrode: light quanta being absorbed by the germanium create electron hole pairs near the surface from which the holes can be used for the dissolution reaction. *Fig. 16* illustrates that the saturation current during anodic dissolution does indeed increase with illumination (compare with the saturation current in fig. 15) and that it also depends on the stirring. From this it may be deduced that the illumination has created so many holes that it is no longer the holes but the OH-ions again that determine the dissolution rate. This situation may change again at very high revolution speeds. Then the supply of OH-ions for the reaction is increased so much that again the reaction rate is determined by the diffusion of holes. A similar effect has also been observed with N-type germanium of high resistivity without illumination.

We now return to the capacity measurements discussed above. These measurements must in general be performed in a potential range in which the germanium is slowly anodically dissolved. This has no further consequences for intrinsic germanium, but is of importance for N-type germanium. Since in the latter case all holes are removed for the anodic dissolution the density of holes at the surface cannot increase, even if this is required by the upward band bending. Consequently, the holes do not contribute to the charge in the space charge region. Since the space charge Q_{R} is now determined only by electrons, it reaches only a comparatively small value upon an upward band bending (U_{R} negative). That increase of the capacity which is

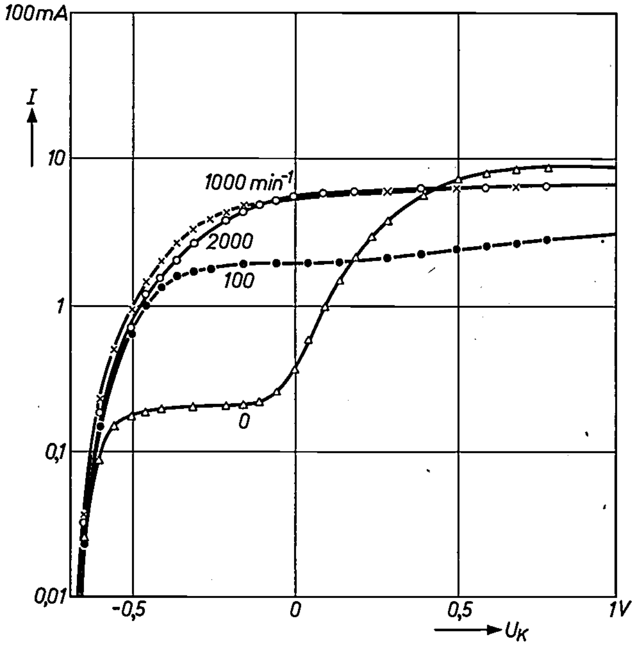


Fig. 16. Anodic dissolution of N-type germanium under illumination. The parameter is the revolution speed of the rotating electrode. (After G. Schwandt and M. Scheer [9].)

normally determined by holes therefore does not occur; the capacity decreases monotonically on shifting the band bending in the upward direction (U_R in the negative direction) (*fig. 17*). Measurements confirm this picture (*fig. 18*) [11].

It has been observed that the capacity does *not* indeed change on adding copper ions to the solution. From this rather surprising result one may conclude that the induced surface states constitute traps for the holes. The reasoning is as follows. Measurements with radioactive isotopes of copper have shown that copper is absorbed by *N*-type germanium in just the same way as by *P*-type or intrinsic germanium [12]. One may consequently expect that the formation of surface states is also independent of the doping of the crystal, i.e. surface states should be present for *N*-type germanium. From the fact that no additional capacity due to surface states appears it may be concluded that, on varying the band bending, the charge in the surface states is not changed. Such a situation can occur if the surface states are traps for carriers which are absent at the surface. In the potential range of anodic dissolution it is the *holes* which are absent, all of them being taken up in this process.

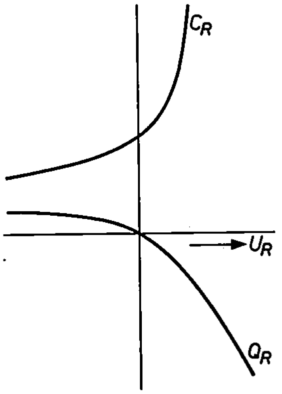


Fig. 17. The charge density Q_R and the capacity C_R of the space charge layer, during a reaction process which removes holes at the surface.

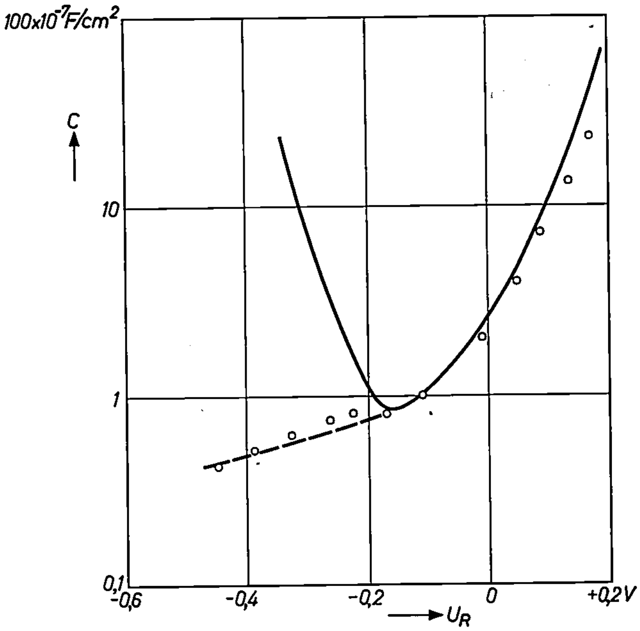


Fig. 18. The capacity C of an *N*-type germanium electrode. The experimental values are plotted against the band bending U_R. The full line gives the theoretical capacity curve of an *N*-type germanium electrode, which is valid if no holes are taken for in the reaction. (After R. Memming [11].)

Summary. In the present paper investigations are reported of surface phenomena of germanium electrodes in contact with an aqueous solution. In certain cases the voltage applied across the germanium-electrolyte interface is completely taken up by the space charge region in the semiconductor. This leads not only to a change in the space charge itself but under certain conditions also of the charge in surface states. These charge variations influence the interface capacity so that capacity measurements are a useful tool to determine the energy level and the density of surface states. Under certain conditions it is possible to obtain "perfect" surfaces, i.e. surfaces without surface states. Extremely small amounts of Cu^{II}-ions in the electrolyte are sufficient to form surface states. Further investigations indicate that OH-ions and holes are removed in the anodic dissolution of germanium. The reaction rate can consequently be increased by stirring the electrolyte and by illuminating the electrode. Knowledge of these phenomena and of measurements based on them, makes it possible to obtain further information on the properties of surface states.

[9] G. Schwandt and M. Scheer, unpublished.
[10] F. Beck and H. Gerischer, Z. Elektrochemie Ber. Bunsenges. Phys. Chemie 63, 500, 1959.
[11] R. Memming, Physics Letters 7, 89, 1963.
[12] R. Memming, Surface Science 2, 436, 1964.

Gallium phosphide light sources and photocells

H. G. Grimmeiss, W. Kischio and H. Scholz

621.383:546.681'183

Preparation and doping of GaP

Amongst semiconductors with a relatively large energy (band) gap, gallium phosphide has aroused interest for various reasons. For one thing, by reason of its 2.25 eV band gap, it is suitable for making *P-N* diodes which in some cases emit light in the visible range of the spectrum.

Our first task was to prepare gallium phosphide of high purity since, as will be made clear below, the efficiency of GaP light sources is very much dependent on the purity of the starting material. GaP is prepared by allowing gallium to react with phosphine; an ample supply of the latter gas, in a very high state of purity, can be obtained by decomposition of aluminium phosphide with water. The aluminium phosphide is prepared by reacting aluminium with phosphorus. A mixture of pure aluminium and red phosphorus in an atomic ratio of 1:1.1 is placed in an iron crucible and ignited. The reaction is fairly violent: some of the phosphorus evaporates and escapes into the atmosphere, where it burns spontaneously (*fig. 1*). The aluminium phosphide yielded by the reaction is a porous sintered substance, yellow in colour.

The GaP is synthesized in the apparatus sketched in *fig. 2*. The round-bottomed flask is charged with aluminium phosphide under a nitrogen atmosphere. The moist gas that evolves when the compound is decomposed by water is first passed through a cooling jacket. This condenses most of its water vapour. Thereafter the gas passes through two traps cooled down to —78 °C by a mixture of acetone and dry ice. Finally it enters the furnace containing a charge of gallium. The metal has first been heated up to 800 °C in a stream of hydrogen; once this temperature is attained the hydrogen flow is cut off and the phosphine fed in. The temperature is then raised to 1200 °C and held there for two hours.

This reaction converts one-fifth of the gallium into GaP, yielding orange-tinted flakes having dimensions of roughly 4×4×0.2 mm. Most of the flakes are single crystals, the large faces being {111} planes. However, crystals twinned on a {111} plane are not uncommon. The overall impurity concentration is less than 10^{-3}%. The methods of analysis used by us have failed to detect the presence of carbon, which interferes with luminescence in GaP, and this implies that the samples must have a C concentration of less than 5×10^{-4}%. In the undoped state these crystals only exhibit very weak luminescence.

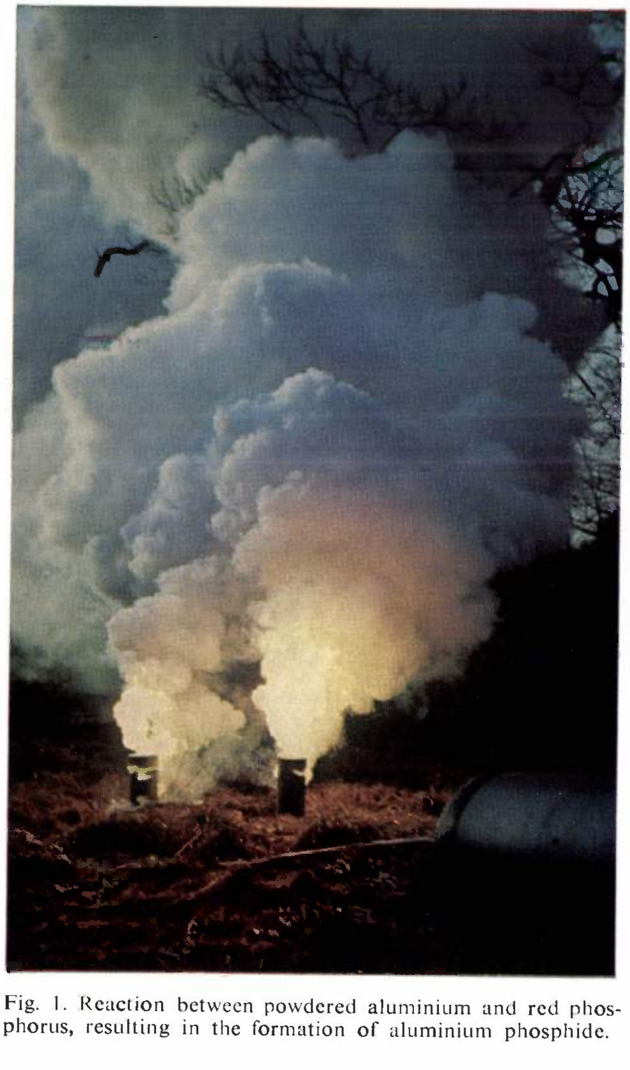


Fig. 1. Reaction between powdered aluminium and red phosphorus, resulting in the formation of aluminium phosphide.

To produce GaP that will luminesce satisfactorily the crystals must be doped, and both zinc and zinc plus oxygen have shown themselves effective doping agents. Zinc gives rise to the emission of green light, zinc plus oxygen to the emission of red light with a peak at 700 nm. As will later appear, it is the red-luminescing crystals that are mainly of interest. The zinc and oxygen can be introduced in the form of ZnO

Dr. H. G. Grimmeiss, Dr. W. Kischio and Dr. H. Scholz are research workers at the Aachen laboratory of Philips Zentrallaboratorium GmbH.

or of $Zn_3(PO_4)_2$, but the use of one particular compound fixes the atomic ratio between the two elements. This can be varied at will if Ga_2O_3 or $GaPO_4$ is taken as a source of oxygen and employed in conjunction with metallic zinc or ZnO.

Doping is carried out in the following way. A quartz container is filled with GaP, metallic gallium and the doping agents, heated up to 400 °C in a vacuum and then sealed off. It is then heated up to 1230 °C in a furnace. The GaP forms a homogeneous solution with the gallium. Cooling must take place slowly if large, well-shaped crystals are to form out of the melt. For that reason the container is lowered slowly through the floor of the vertical furnace chamber. The zinc segregation coefficient between the Ga melt and the GaP is approximately unity. At the same time, recrystallization is accompanied by a sharing of impurities between the solid and liquid phases. The process thus involves further purification of the GaP.

Doping with copper can be carried out by a particularly efficient two-stage process. GaP and copper are brought into contact at 400 °C under an air pressure of 0.5 torr, with the result that copper diffuses into the surface of the GaP crystals, which take on a black coloration. The tinted top layer is about 1 μm deep. It is essential that the GaP should be in intimate contact with the copper during this pretreatment.

The blackened crystals are then annealed for 24 hours at 900 °C in an evacuated container. In the course of the annealing the blackening disappears and the material recovers the appearance it had in the undoped state. It is possible in this manner to convert *N*-type GaP into *P*-type or, alternatively, by modifying the pretreatment conditions, to retain the original *N*-type properties of the material.

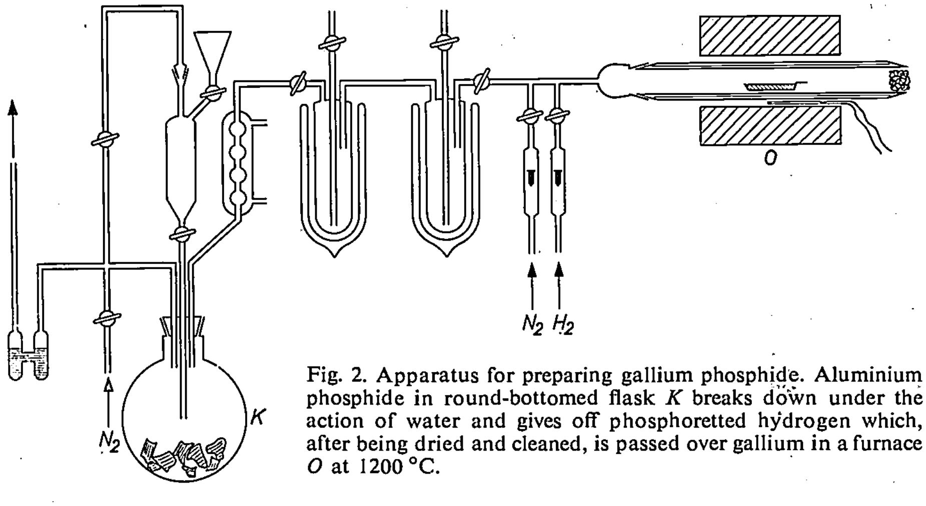


Fig. 2. Apparatus for preparing gallium phosphide. Aluminium phosphide in round-bottomed flask *K* breaks down under the action of water and gives off phosphoretted hydrogen which, after being dried and cleaned, is passed over gallium in a furnace *O* at 1200 °C.

Optical and electrical properties of *P-N* junctions

As already noted, GaP has the relatively large energy gap of 2.25 eV and because of this it is a highly interesting medium to investigate the electrical and optical properties of *P-N* junctions. It will also be recalled that particularly pure starting material is required where luminescence is to be studied. Therefore we used GaP prepared in the manner described above for making experimental diodes.

Of the various methods for producing *P-N* junctions, alloying seemed the best suited to our purpose. But one has to ensure that during the alloying process some gallium phosphide is taken up into the molten alloying metal, with the result that a greater or lesser amount of phosphorus is lost, depending on the reaction time. This phosphorus ceases to be available for recrystallization as GaP during the subsequent cooling stages; the longer the alloying time, the thinner is the recrystallization layer. Since the phosphorus escapes very quickly, we should have to use an alloying method that would ensure rapid heating and cooling of the samples. We therefore decided to heat up the GaP electrically, by passing a heavy current through an iridium strip. In this way we could easily cut down the alloying time to a second or so. In the method adopted by us, the ohmic contact is made of a gold-zinc alloy and the *P-N* junction is formed by alloying the GaP with tin.

The *P-N* junctions thus produced run more or less parallel with the surface of the crystal. *Fig. 3* shows clearly that despite the short alloying time, only a very thin crystallization layer is formed. As measurements of junction capacitance have established, the junctions obtained in this way are step junctions. Where this is so, the inverse square of the junction capacitance shows a linear relationship [1] to applied voltage; *fig. 4* is a plot of this kind for *P-N* junctions made by us.

The electrical properties of the diodes are governed above all by the relatively large energy gap of GaP [2]. Because it has a very low intrinsic conductivity, GaP can be used to make diodes which sometimes do not pass more than 10^{-11} A under a reverse bias of up to about 10 V; some have leakage currents of less than 10^{-13} A. For the same reasons the current in the forward direction shows a logarithmic dependence on applied voltage over several orders of magnitude (*fig. 5*). The result is that at 1.4 V, for example the ratio of forward to reverse current is about 10^{11} : 1.

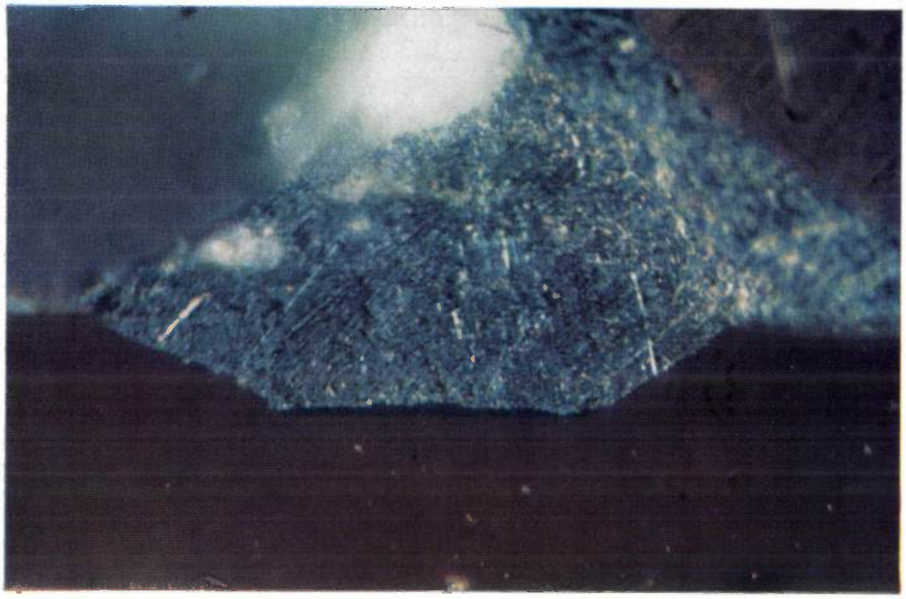


Fig. 3. Photograph of a *P-N* junction in a gallium phosphide diode fitted with a tin contact. The *P*-type GaP is in the lower part of the print; part of the copper wire which carried the diode current can be seen at top right.

It is known that when a *P-N* junction is biased in the forward direction, minority charge-carriers are injected into both the *P*-region and *N*-region, where they are able to recombine with majority carriers already present. In general, recombination may take place by direct transitions from band to band or by way of intermediate levels within the forbidden gap. In GaP however, for various reasons, of which the compound's band structure is not the least important, recombination takes place mainly via impurity levels. The energy released by recombination inside diodes is normally transferred to the crystal lattice in the form of heat. However, by incorporation of suitable impurity levels one can arrange for this energy to be radiated outwards, not as heat, but as light. The effect is known as *P-N* or injection luminescence.

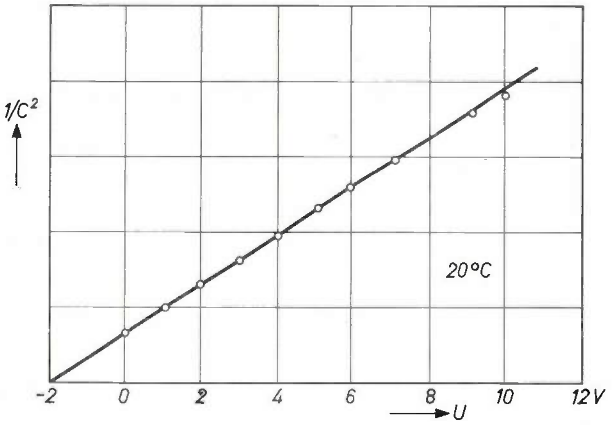


Fig. 4. The inverse square of the capacitance *C* of a *P-N* junction in GaP, as a function of the reverse bias voltage *U* at room temperature. The ordinate scale is in arbitrary units.

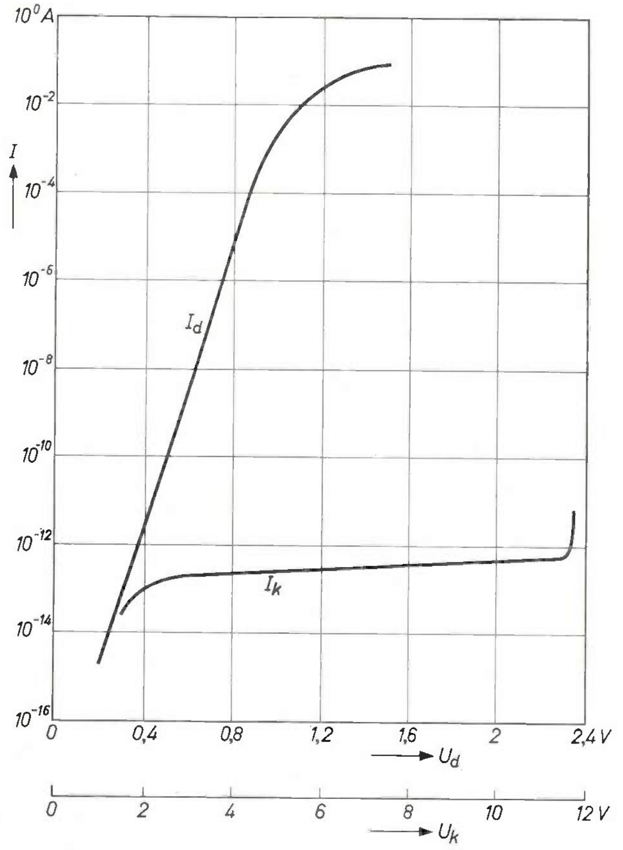


Fig. 5. Current-versus-voltage characteristics of a GaP diods. U_d forward bias voltage, I_d forward current, U_k reverse bias voltage, I_k inverse (leakage) current.

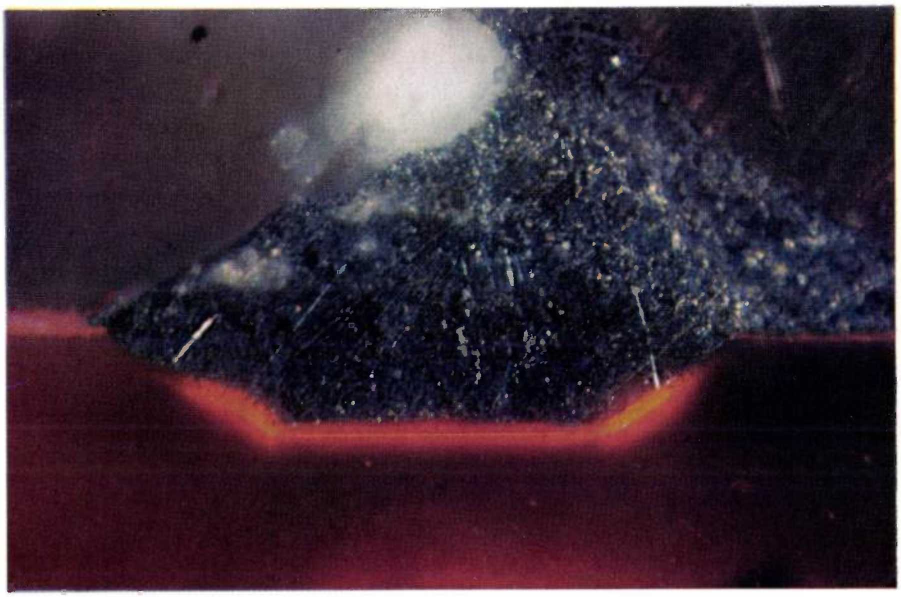


Fig. 6. The same *P-N* junction as shown in fig. 3, now exhibiting luminescence when a forward current (10 mA) is flowing.

The main problem involved in getting GaP to luminesce in this way is that of preventing injected charge-carriers from recombining with the killer centres that are normally present, that is to say, with certain imperfections which cause the recombination energy to be transferred to the crystal lattice in the form of heat, without the emission of radiation. We have managed to avoid these radiationless recombination processes firstly by using extremely pure starting material, and secondly by doping it with zinc and oxygen in the manner already described. Zinc and oxygen create impurity levels in GaP such that, when charge-carriers recombine, the process predominantly involves the emission of light. The nature of these impurity levels has not yet been definitely established. For the sake of simplicity we regard them as acceptor centres lying at a distance of 0.45 eV from the valence band. It is essential, too, that the alloying operation of the Sn contact should occur very quickly. By these means, *P-N* diodes can be obtained that are capable of emitting light with a quantum efficiency of about 1% (the quantum efficiency is the ratio between the number of quanta emitted and the number of electrons injected). Of course, much light is lost by multiple reflection, for GaP has a high refractive index ($n = 2.9$) [3]. But compared with other materials for *P-N* light sources, gallium arsenide for example, GaP has the great advantage of allowing this high quantum yield to be obtained near room temperature as well as at much lower temperatures. *Fig. 6* shows a photograph of a GaP sample with a luminescent *P-N* junction.

The separation of the acceptor level and the conduction band determines the amount of radiant energy released when charge-carriers recombine. In the case of GaP doped with zinc and oxygen the separation is 1.8 eV, which means that the emitted radiation has a peak at 700 nm. This lies in a part of the spectrum to which the eye has low sensitivity, and a device emitting such long-wave radiation is therefore of little use as a light source; interest thus lies in the emission of shorter waves. True, it is also possible to get emission in GaP at about 565 nm, which is quite near the peak in the spectral sensitivity curve of the eye. But the fact that the energy gap of GaP lies around 550 nm (2.25 eV) means that the level responsible for luminescence is very close to the valence band; such a level is largely ionized at room temperature and is therefore no longer available for recombination with electrons from the conduction band. The luminescence effect is in fact "temperature-quenched". For this reason, then, one cannot expect efficient short-wave emission at room temperature based on such transitions except perhaps in the yellow part of the spectrum.

If the possible capture of charge-carriers by traps be neglected then the rise and decay times of the emissions of such a *P-N* light source are dependent only on the lifetime of the injected charge-carriers. It is known that this lifetime can be much shorter than a microsecond. The implication is that the luminescing *P-N* diode has a faster response than any other electrical light source. GaP diodes are not the fastest sources in the *P-N* class, but even so, it is relatively easy by appropriate doping to produce light pulses with durations as short as 10^{-6} to 10^{-7} s. Very short rise and decay times are amongst the most striking properties of *P-N* light sources; in this respect

[1] See for example R. A. Greiner, Semiconductor devices and applications, McGraw-Hill, New York 1961.

[2] C. T. Sah, R. N. Noyce and W. Shockley, Proc. IRE **45**, 1228, 1957.

[3] H. G. Grimmeiss and H. Scholz, Physics Letters **8**, 233, 1964 (No. 4).

the diodes differ completely from other kinds of devices whose electroluminescence is based on the Destriau effect, and which cannot readily be made with switching times shorter than 10^{-3} s.

The *P-N* photovoltaic effect, the opposite process to *P-N* luminescence, is observed when the junctions are irradiated with light. One of the most important prerequisites for obtaining a photocurrent from a *P-N* diode is that, by either thermal or optical excitation of a suitable kind, both mobile holes and mobile electrons should be created in the neighbourhood of the junction; one such process is fundamental lattice absorption, which consists of the transfer of electrons from the valence band to the conduction band. If only one type of charge-carrier is mobile, the electron-hole pairs produced by excitation, which are separated by the electric field in the *P-N* junction, cannot be neutralized by a flow of current through the external circuit, and there will be no photocurrent. GaP differs from other semiconductors like germanium, silicon and gallium arsenide in being photosensitive over a much wider range of energies than that corresponding to fundamental lattice absorption: the spectral sensitivity curve of GaP extends into the infra-red range of the spectrum in proportion to the doping (*fig. 7*).

The current generated by a short-circuited *P-N* photocell depends on the concentration of electron-hole pairs produced by excitation, but also on the lifetime of the minority carriers that are created concurrently. Thus two ways are available of strengthening or weakening the short-circuit current.

In photocells made of GaP doped with zinc and oxygen, it is possible to increase the concentration of electron-hole pairs obtained for instance by fundamental excitation due to "two step processes". These processes make use of radiant energies smaller than the energy gap. Mobile electrons can be generated with the aid of light quanta having an energy of about 1.8 eV, which is smaller than the energy gap but large enough to lift electrons into the conduction band from the acceptor level lying 0.45 eV above the valence band (*fig. 8*). If at the same time the device is irradiated with other light quanta — whose energy need only be greater than 0.45 eV — the empty acceptor centres will refill with electrons from the valence band, and in this way free holes will be created. The electron-hole pairs arising out of these two absorption processes are quite capable of bringing about a photoelectric effect, for the following reason. The time taken by the empty acceptor centres to fill up is shorter than the lifetime of the electrons excited in the first instance; after the second stage of excitation, mobile electrons are available as well as mobile holes. If donors are involved similar processes are possible. Besides exhibiting a "normal" photoelectric effect within the limits of fundamental lattice absorption, then, these GaP photocells are sensitive to light energies smaller than the energy gap. When exposed to sunlight at room temperature they have a measured efficiency of about 3% (the efficiency is the number of quanta irradiated divided by the number of electrons generated). At temperatures of 80 to 90 °C this figure may rise to 4% or 5%, since the centres responsible for dark conductivity are not fully ionized at room temperature. If the unfavourable geometry of our samples be borne in mind it will be appreciated that their efficiency is capable of further improvement — a promising line of development seeing that, under suitable conditions, the cells are sensitive over a range extending from the ultra-violet to the infra-red.

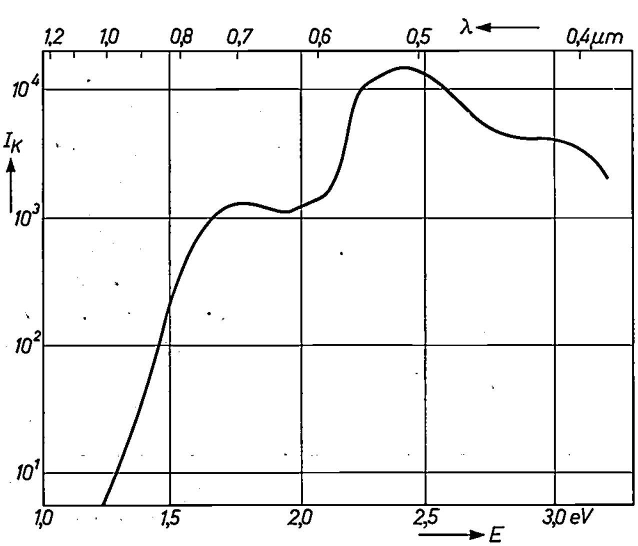


Fig. 7. Spectral sensitivity at room temperature of a *P-N* photodiode made of GaP doped with zinc and oxygen. The short-circuit current I_K is in arbitrary units. E energy of photons.

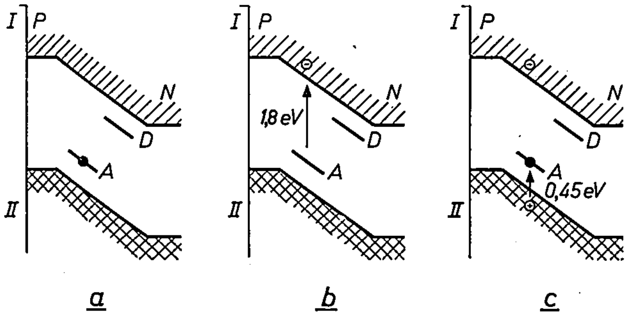


Fig. 8. One way of producing a *P-N* photovoltaic effect with the aid of energies smaller than the energy gap. *a*) ground state, *b*) first stage of excitation, in which a free electron is generated, *c*) second stage of excitation, resulting in the creation of a free hole. *I* conduction band, *II* valence band, *A* acceptor levels, *D* donor levels.

Thus in photocells made of GaP doped with zinc and oxygen, two-stage processes can be used to strengthen the "normal" short-circuit current obtained from fundamental excitation. In copper-doped GaP photocells the "normal" short-circuit current can be modified by altering the lifetime of the minority carriers. A copper-doped cell will deliver a short-circuit current of the usual magnitude, due to fundamental lattice absorption, on exposure to blue light, for instance. By adding infra-red radiation to the blue, one can modify the occupation of the recombination centres within the forbidden zone of GaP. We have been able to demonstrate that the result of so doing is to lengthen the lifetime of the excited minority carriers when the starting material is of *N*-type, and to shorten it when the starting material is of *P*-type. Accordingly, the "normal" short-circuit current is increased in the former and decreased in the latter case. In the latter case particularly it is possible by combination with other kinds of centre to achieve an astonishingly high degree of sensitivity to infra-red light. Devices working on this principle are capable of detecting wavelengths down to about 2 μm with intensities as low as 10^{-7} W.

Opto-electronic applications

Both the light sources and the photocells described above are suitable for embodiment in opto-electronic circuits. The great advantage of these systems, as compared with conventional electronic ones, is that the signal at some point in its path is transmitted in the form of light, a fact which permits complete electrical separation of sub-circuits which would otherwise be subject to unwanted coupling. A further advantage is that if integrated circuits are employed, the overall dimensions of the system can be reduced very considerably. It is true that integrated circuits often sacrifice one of the greatest benefits of *P-N* light sources, namely their speed of response as compared with other types of source, since the usual practice is to use a photoconducting device for detecting the optical signal. Sensitive photoconducting devices have a sluggish response, and consequently the speed of the system is determined, not by that of the *P-N* light source, but by that of the radiation detection. If *P-N* light sources are to be adopted on a wider scale for opto-electric applications, then it will be as well to have an equally fast semiconducting device available for the role of detection.

Signal transmission involving frequencies up to several megacycles is quite feasible using fast optical detection such as photomultipliers. But even with the relatively slow photoconducting devices it is quite possible to devise opto-electronic systems with attractive properties, for a photoconducting device is not only a transducer, it is also an extremely efficient amplifier. Power gain factors as high as 10^6 or 10^7 are now obtainable from commercial types, and even where these are coupled with light sources of poor efficiency, the system still provides a useful amount of amplification.

The set-up in *fig. 9* is a case in point. Here the light from a GaP light source falls on a CdS photodiode whose dark resistance is thereby reduced from about 10^{10} Ω to about 1000 Ω. The advantage of using a *P-N* light source rather than some other type is that it only requires about a two volt DC supply: several hundred volts AC are required for electroluminescence cells based on the Destriau effect.

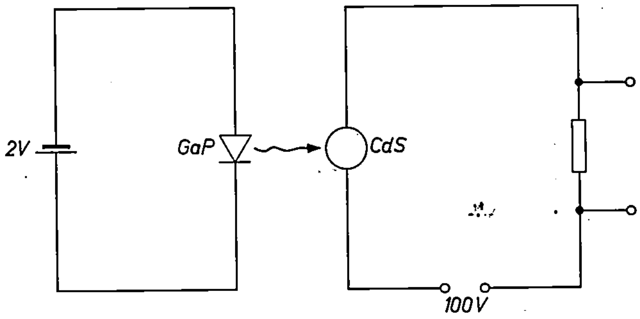


Fig. 9. An example of an opto-electronic relay; here a GaP light source is coupled with a CdS photoconducting device.

An arrangement like that in fig. 9 can therefore be employed as a relay or as a kind of transformer. Further, it can be built in very compact form, the luminance of a GaP source being greater than that of an incandescent lamp.

Again, the system can be given a very steep characteristic curve (*fig. 10*) by exploiting the non-linear relationship between voltage and current that is associated with *P-N* luminescence. The combination then becomes useful for switching purposes. A further inference that may be drawn from fig. 10 is that despite the unfavourable geometrical conditions of the experiment to which the curve relates, the system afforded a certain amount of current gain and, what is more important, a certain amount of power gain. Improvements to the geometry should allow the resistance of the photoconducting pick-up to be reduced by one or two orders of magnitude, and the arrangement would then become a distinctly interesting proposition from the telecommunications standpoint.

Useful opto-electronic systems can also result from the combination of the photocells with other devices. For example, if a copper-doped GaP photocell is exposed to the green light from a zinc-doped GaP

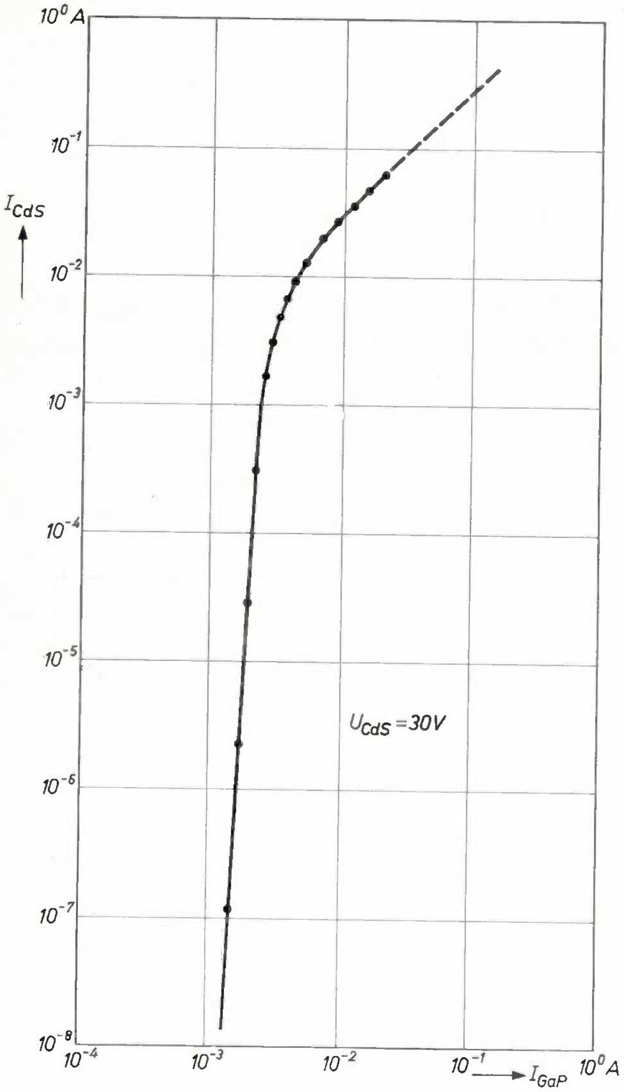


Fig. 10. Characteristic curve of the combination in fig. 9. I_{CdS} the current passed by the photoconducting device, has been plotted as a function of the current I_{GaP} supplying the light source.

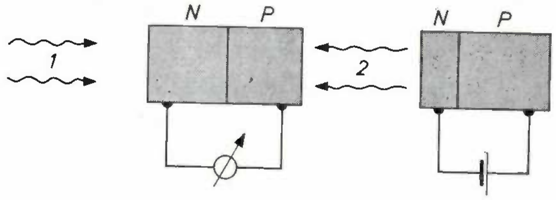


Fig. 11. Infra-red detector embodying a Cu-doped GaP cell. *1* infra-red radiation, *2* green light.

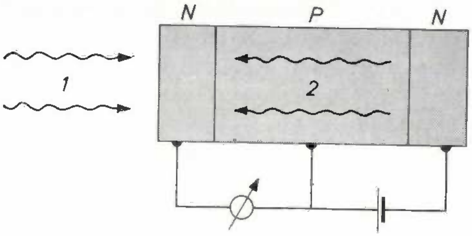


Fig. 12. Infra-red detector made of GaP which has been doped with zinc and oxygen. *1* infra-red radiation, *2* light of wavelength 700 nm.

light source (*fig. 11*) it will deliver a "normal" value of short-circuit current which is sensitive to additional irradiation with infra-red, as has already been noted. Such a combination can therefore serve as a very compact detector of infra-red radiation. The dimensions of the detector can be reduced still further by making use of a Zn+O-doped GaP cell with *two* alloyed *P-N* junctions (*fig. 12*). If one of the junctions is forward biased it will luminesce at 1.8 eV, and cause a photocurrent to flow across the second junction. Since this current will increase under additional irradiation with infra-red, the cell can be employed as an extremely compact infra-red detector.

If, in an arrangement like that of fig. 11 or that of fig. 12, the infra-red radiation undergoing measurement is produced by a *P-N* light source — a gallium arsenide cell, for example, or a Cu-doped GaP one — the result will be a kind of triode whose gain factor mainly depends on the intensity of the infra-red. The great advantage of such a triode, as compared with a phototransistor [4], is that control action is effected by a type of radiation which suffers very little absorption in the semiconducting material.

Summary. The preparation and methods of doping of very pure crystalline gallium phosphide is described. After doping with zinc and oxygen it becomes suitable for making photocells that are sensitive throughout the visible range of the spectrum, and light sources working on charge-carrier injection which have very short switching times (about 10^{-6} s) and a high quantum efficiency (around 1 %) near room temperature. Photocells able to detect very low intensities at wavelengths down to about 2 μm can be made out of Cu-doped GaP. Some opto-electronic applications of GaP devices are cited.

[4] See for example R. F. Rutz, Proc. IEEE **51**, 470, 1963.

Small electric motors

R. Thees

621.313.13-181.4

By "small" electric motors we mean fractional horsepower machines with mechanical outputs up to the equivalent of about 15 watts, of the kind incorporated in domestic appliances and playback equipment. The usual source of electrical power for these machines is the single-phase AC mains or the DC supply from a battery; we shall confine ourselves in the present article to motors running on these two kinds of voltage.

In considering a motor for a certain output power, design requirements will partly be determined by the available power source. For a motor to run on AC mains small size will be the chief requirement. Increasing the efficiency is in general not compatible with this requirement, but the efficiency need only be high enough to keep the temperature rise produced by heat dissipation in the windings, after the motor has been built into a device, below the limit imposed by safety codes. On the other hand, high efficiency is the primary requirement for a battery-operated DC motor.

Irrespective of which of the two types of supply it is to operate on, the motor must have the lowest possible noise level and it must not interfere with radio and television reception. Feeding current to rotor windings by commutation should therefore be avoided. This at the same time helps to satisfy a further requirement, namely that the motor should embody the least possible number of parts subject to wear or requiring maintenance.

The purpose for which the motor is to be used determines the limits within which its instantaneous or mean speed of rotation must remain constant in the face of load and supply voltage fluctuations. In motors powering household appliances it is permissible for the mean speed to show variations of the order of 10%; in playback equipment, the instantaneous angular velocity of e.g. the turntable must not vary by more than 0.06%. We shall now go on to describe a synchronous AC machine that is equally suitable for playback equipment and for domestic appliances. This will illustrate the possibilities opened up by the employment in small motors of permanent magnets made of modern ceramic-oxide materials.

Synchronous motors

Let us first consider a design such as that in *fig. 1a*. The rotor is a solid cylinder of an anisotropic ceramic oxide material (FX-D-300) which has been magnetized in the direction of preferential orientation; this runs transversely, i.e. parallel to one of the diameters of the cylinder. The stator, consisting of stacked iron laminations, carries the field windings and also the bearings for the rotor. A motor of this design is a synchronous motor; it will only deliver mechanical power when running at the rate of one revolution per AC period, i.e. in synchronism with the mains frequency. We shall first enquire into the behaviour of the motor in this synchronous rotation.

The benefits of using a permanent-magnet rotor made of ceramic oxide can be readily demonstrated without going into the theory of these motors. The magnet has a reversible permeability of unity. By virtue of this fact, the air-gap between the magnet and the stator iron has little influence on the magnet working point; it is therefore possible for this gap to be wider than in other types of motor. If the machine is operated as a generator the e.m.f. E induced in the windings will be strictly sinusoidal, as a consequence of the transverse magnetization of the rotor. We shall show below that this is essential for achieving a high efficiency of the machine. Moreover, because of its low permeability, the rotor represents a high reluctance in the magnetic circuit of the field windings; this results in a comparatively small phase difference between the energizing current and the resulting voltage across the windings and this again is important for achieving a reduced bulk and high efficiency. The remanence of the magnet

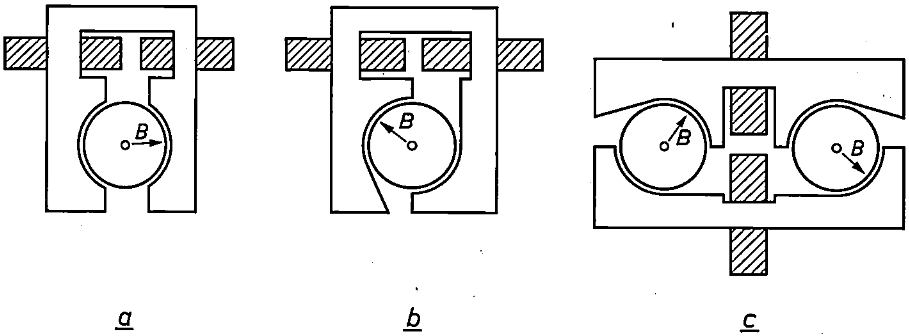


Fig. 1. Synchronous motors with (*a*) symmetrical and (*b*) asymmetrical stator poles, and (*c*) design having twin rotors. *B* direction of rotor magnetization.

Dr. R. Thees is a research worker at the Aachen laboratory of Philips Zentrallaboratorium GmbH.

material should be as high as possible, since the working point of the magnet is always close to the remanence and the higher the remanence, the greater the reduction that can be effected in the size of the motor, other things being equal.

Ferroxdure rotors also enable special designs to be made very efficiently. Fig. 1*c* shows, by way of example, a twin-spindle design; the high reluctance of the two rotors, which lie in series in the magnetic circuit, and the strong mutual coupling between the two magnets they constitute, make this a very efficient motor.

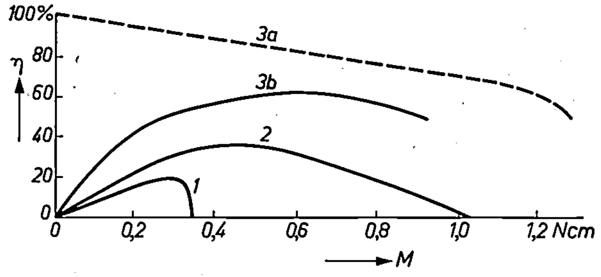


Fig. 2. Efficiency η as a function of torque M (in newton·cm) in various types of motor.
Curve *1*: asynchronous type running on 220 V and having a bulk of 126 cm^3 and a weight of 314 g.
Curve *2*: series-wound machine running on 220 V and having a bulk of 71.5 cm^3 and a weight of 152 g.
Curves *3a* and *3b*: synchronous type;
3a relates to the case where the conditions $U/2E = \cos \alpha$ and $E = U$ are satisfied;
3b relates to an actual model running on 6.3 V and having a bulk of 32 cm^3 and a weight of 104 g.

In *fig. 2*, the synchronous motors that form the subject of this article are compared, in regard to efficiency, output or delivered torque, weight and bulk, with the induction motors commonly incorporated in record-players (curve *1*) and the series-wound motors used in electric shavers. Curve *3a* represents their power output and efficiency attainable under certain conditions. To find out what these conditions are, it must first be ascertained what power output is obtainable from a motor running on a given mains voltage U and having stator windings of a given resistance R. On differentiating the power output equation and setting the first derivative equal to zero we obtain a relation connecting the current phase angle α and the r.m.s. voltages involved:

$$\cos \alpha = \frac{U}{2E},$$

as also the peak output obtainable:

$$P_{\mathrm{am}} = \frac{U^2}{4R}.$$

If we now consider what is necessary for an unloaded or only very lightly loaded motor to have maximum efficiency, while satisfying the above-mentioned phase condition, we arrive at a further condition, namely that

$$E = U.$$

Fulfilment of both conditions results in an efficiency versus output curve such as that marked *3a* in fig. 2. One of the implications is that if E, the induced e.m.f., has the same shape and amplitude as the mains voltage U, then there will be a 180° shift between the two waveforms when no mechanical power is being delivered. In that case, no current will flow and no wattage will be consumed. Thus 100% efficiency represents a limit approached when the output and torque are small. It should of course be noted that the torque just referred to is that exercised via the air-gap, and that no account is being taken of losses such as that due to friction. Within the limits of experimental error, the results of measurements are in agreement with this curve. Curve *3b* relates to a motor of the type produced for record-players in which it is more important to have a small bulk than high efficiency; in fact E in this case is 0.57 U. The region lying between curves *3a* and *3b* shows what latitude is available for a compromise between bulk, efficiency and start-up behaviour. This last point will now be dealt with.

Starting may be difficult if the induced e.m.f. and mains voltage have the same amplitude; we are therefore going to look rather more closely into the manner in which the rotor is set into motion.

When the motor is switched on, synchronism is reached via rotary oscillations with increasing amplitude. Now, we must distinguish between three conditions for successful starting. Firstly, the rotor must start the rotary oscillation as soon as a voltage is applied to the windings. Secondly, these rocking movements must attain an amplitude of at least 180°; that is to say, the rotor must work up an angular velocity corresponding to the synchronous speed. Thirdly, having attained the sychronous speed, the unloaded motor must run stably, i.e. the instantaneous angular velocity may fluctuate, but its sign must not change. It is quite possible for any one of these three conditions to be satisfied while the other two are not.

Consider first a stator design such as that in fig. 1*a*. When no exciting current is flowing through the stator windings the rotor will take up a position with respect to the stator poles such that maximum magnetic flux passes through the windings. In any other position — apart from two unstable positions where the flux is at a minimum — the rotor will experience a restoring couple that tends to turn it to the position of maximum magnetic flux. Thus the magnet has two stable rest positions in the stator. If now the windings are energized with an AC voltage, they will create a field of constant direction whose amplitude changes sinusoidally.

This field has the same direction as the magnetic induction in the rotor. There will be no couple acting on the rotor. Only when the rotor is turned out of its rest position, for instance by a shock, a rocking movement will be initiated and maintained. However, the need to turn the rotor out of its rest position can be avoided in the following way. For the magnetized rotor there is only one air-gap, the true one, but for the stator windings the rotor itself, with its unit permeability, represents an air-gap. This being so, by choosing a stator of *asymmetric* shape, as in fig. 1*b*, one can give the field of the windings a slight slant with respect to the induction in the rotor. This field will now exert a couple on the rotor in the rest position. The lack of balance between the stator poles will thus enable a rocking motion to start. Rotation in either sense is equally probable with a symmetrical stator as shown in fig. 1*a* (provided all conditions for start-up are fulfilled). With an asymmetric stator it is possible to favour starting in one sense or the other. The probability of starting in the chosen sense can be as high as 90%, depending on the shape of the stator, but in practice it never amounts to 100%. This is a matter we shall have to return to later.

The rocking movement is governed by two factors, the restoring couple and that exercised on the rotor by the field of the winding. It is no easy matter to calculate the motion of the swinging rotor in terms of the motor parameters, but experience with these motors has shown that in practice the swings always build up to an amplitude of 180° and the angular velocity always attains the synchronous value.

To get an insight into the requisites for stable running at no-load we must consider the instantaneously developed torque, that is to say, the couple experienced by the rotor in its various positions with respect to the stator. We need only consider half a revolution because the instantaneous torque goes through two full cycles in one revolution. It can be seen from *fig. 3a* that the instantaneous torque, the resultant of the theoretical developed torque and the restoring couple, has a negative value — i.e. a value opposing the sense of rotation — over almost 80° out of the 180°. The rotor generally has a moment of inertia sufficiently large to prevent reversal while it is passing through this 80° sector. But the strong braking couple acting on it at these times is responsible for a great deal of fluctuation around the mean angular velocity. Usually this "hunting" gets progressively worse; the amplitude of the fluctuating component grows rapidly, with the result that the motor reverses after having performed only a few revolutions in the same sense. Here we have a motor which starts successfully but is unstable in operation at no-load.

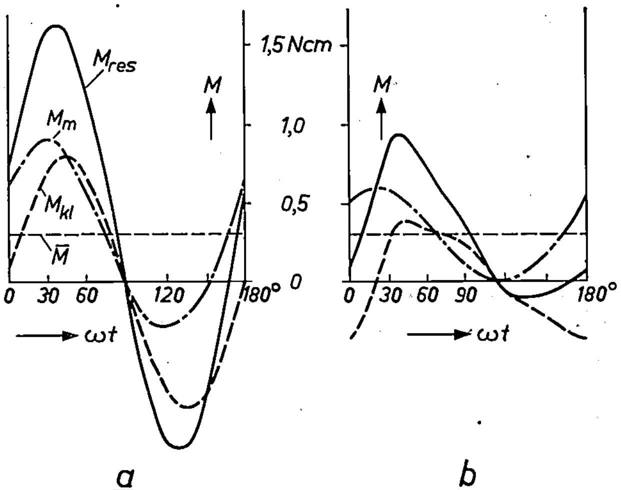


Fig. 3. Instantaneous and mean torque delivered under no-load conditions (*a*) by a synchronous motor whose stator iron is not magnetically saturated and (*b*) by the same machine with saturation.
M_m: theoretically developed torque.
M_{kl}: restoring couple.
M_{res}: resulting (net)torque.
$\overline{M}$: mean torque delivered.

Fig. 3*a* reveals that the restoring couple contributes a good deal towards the amplitude of the instantaneous torque. It is now possible by a simple artifice to make this couple zero. When the angle of the rotor position is φ, the flux through the stator iron is $\Phi = \Phi_0 \sin \varphi$. If the rotor is cut transversely into two equal parts (*fig. 4a*) and if one half is turned through 90° with respect to the other, the flux through the iron will be reduced by a factor of $\sqrt{2}$, becoming:

$$\Phi = \tfrac{1}{2}\,\Phi_0 \{\sin \varphi + \sin (\varphi + 90°)\} = \frac{\Phi_0}{\sqrt{2}} \sin \left(\varphi + \frac{\pi}{4}\right).$$

The restoring couple, however, which was originally given by:

$$M = D \sin 2\varphi,$$

by the mutual displacement of the two rotor halves is changed to:

$$M = \frac{D}{2} \{\sin 2\varphi + \sin\, [2(\varphi + 90°)]\} \equiv 0\,,$$

and thus is eliminated.

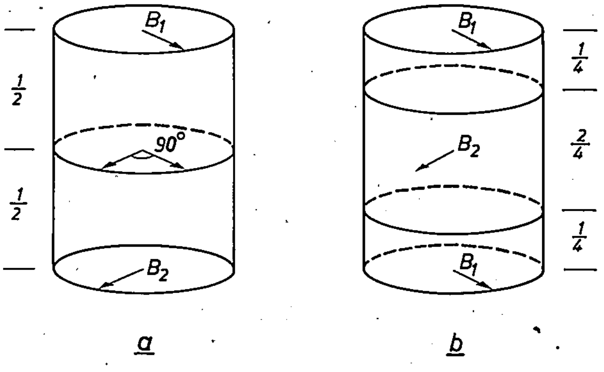


Fig. 4. Split magnet rotors not subject to a restoring couple.

If it is desired to obviate axial forces on the rotor it can be split transversely into *three* parts, as shown in fig. 4*b*, the central third being turned through 90° with respect to the other two.

Even in the absence of a restoring couple, the developed torque is still negative over part of its cycle, as fig. 3*a* shows. We have so far assumed that the magnetization characteristic of the stator iron is linear throughout its operating range. Now, the whole situation can be changed when the cross-sectional area of the stator is reduced to a point such that magnetic saturation of the stator iron is attained already under no-load conditions. Fig. 3*b* shows that when this is done, both the theoretical torque (as calculated from measured current and voltage values) and the restoring couple have highly distorted waveforms. The negative part of the theoretical torque has become practically zero; without employing the artifice illustrated in fig. 4, the resulting net torque is now negative over so small a sector that operation of the motor is stable at no-load. It is true that the instantaneous torque (and hence also the angular velocity) still fluctuates in a roughly sinusoidal fashion around a mean value, but the amplitude of this fluctuation is now very much smaller. The motor runs much more evenly. Theory and experiment show that the lowest mains voltage U at which starting is obtainable, together with stable operation under no-load conditions, is roughly equal to the induced e.m.f. E, when the magnetic saturation of the stator iron is high enough. In general, then, U must be larger than E if supply voltage fluctuations are not to cause starting difficulties. Once a load is placed on the motor the angle of slip will change; that is to say, the main field and the magnetic induction in the rotor will no longer attain their maximum at the same instant. The resulting field in the stator iron will have a smaller amplitude, and saturation will not now occur. This means that the properties of the motor under load are independent of the state of saturation that must occur in the stator iron for stable no-load running.

If the motor is always required to run in the same sense (which is not a necessary condition in electric razors), some mechanical means of preventing reversal is almost always to be preferred to switching of the energizing current, effected by a commutator, or to the employment of shaded poles to create a rotating stator field. The mechanical means will only serve to reinforce the existing preference for rotation in the required sense, and will not need to absorb strong forces. It can therefore be simple in design.

It should be added that these motors will start unaided only if the moments of inertia of elements rigidly coupled to the rotor are not unduly large. The coupling to such elements having large moments of inertia must be by elastic means. Elastic coupling will also afford a convenient means of achieving the necessary degree of constancy in the instantaneous angular velocity, required for gramophones and tape recorders, for example. Because they run evenly at a speed independent of load and supply voltage, and because they cause no interference, the motors described above are eminently suitable for playback equipment.

Direct-current motors

Turning now to battery-operated motors, let us start by taking another look at the set-up in fig. 1*b*. We shall assume that suitable devices for converting DC into AC have been inserted in front of the stator windings. A converter providing a *sinusoidal* AC supply, though theoretically feasible, is ruled out on efficiency grounds. In fact the converter will have to take the form of a switching device, which means that the voltage applied to the stator coil will possess a more or less rectangular waveform. For our present purposes it does not matter whether two voltages of opposite polarity are drawn alternately from two batteries and supplied to one winding or whether one battery is connected alternately to two windings wound in opposite senses.

As in the synchronous motors, the induced e.m.f. must have the same shape and amplitude as the voltage applied to the terminals. A "*radially*" magnetized rotor will induce in the windings a voltage of trapezoidal waveform, which however in practice is a very good approximation of the rectangular waveform. A radial magnetization can be conferred on a rotor made of isotropic ceramic oxides, though the low remanence of the material means that such rotors have to be fairly large. Suspensions of anisotropic ferrites in plastics (plastoferrites) may offer new possibilities in this respect. Considered on its own, without the associated voltage converter, the resulting motors can have an efficiency quite close to that represented by curve *3a* in fig. 2.

If the current energizing the windings is switched by a commutator the start-up difficulties will disappear, and the motor will have an efficiency figure high enough to merit attention, despite the well-known drawbacks of commutation. It is true that brush losses will reduce the overall efficiency, especially in a small machine.

With a transistorized square-wave generator supplying the motor, the overall efficiency of the combination will still be superior to the 55% currently obtainable from DC motors in the fractional horse power class. The constancy of the motor speed will depend on the extent to which the square-wave generator is free from feedback and on the stability of its operating frequency. In all other respects the motor will resemble the synchronous type described above.

The circuitry can be simplified by allotting a dual rôle to certain parts or components of the motor and converter device. For example, the stator windings can at the same time act as inductance for the converter oscillatory circuit. This raises particularly difficult problems of speed stabilization.

One way of getting constant r.p.m. from a motor embodying a commutator is to switch the supply current by means of a thyristor placed in series with the commutator. The thyristor is fired with a predetermined frequency and quenched by commutator action at certain points in the cycle. If the ratio between the firing frequency and the number of rotor revolutions per second is a whole number, any change in load, causing a momentary change in the angular velocity of the rotor, will give the commutator cycle a lead or lag with respect to the instant of firing; this means that the motor output adjusts itself without change in mean speed. This, then, is a way of "synchronizing" a shunt-wound machine for certain running speeds, and at these the motor will possess all the properties of a synchronous motor. The change in load the motor is able to take up without losing synchronism differs very considerably from one of the running speeds to another. The widest and most stable range is obtained by firing the thyristor twice per revolution.

All the available experience indicates that permanent-magnetic ceramic oxide rotors are likely to offer the same advantages in DC motors as in AC machines.

Methods of measurement

Although the results of the measurements cannot be presented in detail here, a brief account of the methods employed will be given.

The mean power delivered by the motors was measured by means of an eddy-current torque meter with a knife-edge fulcrum. In its lowest range the instrument has a sensitivity of 10^{-4} N/m per scale division. The error involved in measuring an output of 2 W from a synchronous motor was ± 1%. Curves of instantaneous power output were produced on an oscilloscope screen with the aid of an inductive pick-up. However, it is difficult to carry out measurements in this fashion, and the results are not very accurate.

The electrical power consumed by the motors was measured with a Hall wattmeter. An instrument was developed for these measurements having four voltage and four current ranges automatically selected by relays, thereby providing seven ranges for wattage measurements. Readings are accurate to 0.015 W in the lowest of these ranges; the measurement of a 2 W power consumption involves an error of ± 2%.

Instantaneous angular velocities were measured with a unipolar generator. The moving parts of this instrument have a moment of inertia so small as to be negligible in comparison with that of the rotor under test. Measurements made with this instrument in conjunction with an oscilloscope have a sensitivity of down to 2 r.p.m. A resolving power of 0.1 r.p.m. is obtainable for slow speed changes traced by means of a chart-recording instrument. Mean values down to about 0.04 r.p.m. can be measured with a millivoltmeter. However, in order to measure down to these limits it is necessary to make fairly elaborate provisions for stabilizing the magnetic field and, in cases where the motor speed is high and fairly constant, its mean value being subject only to small fluctuations, a good deal of attention must be devoted to compensating the DC component. As against this, digital counting of the mean r.p.m. is a technique from which a sensitivity of 0.01 r.p.m. can readily be obtained in any speed range.

Summary. The properties of small motors embodying permanent-magnetic rotors made of ceramic oxides are described. Conditions for unaided starting and stable running at no-load in single-phase AC motors are discussed, and methods are given whereby these conditions can be fulfilled in machines of high efficiency or small bulk. The designs in question can be modified for DC operation, resulting in machines promising the same properties as the AC motors.

Problems and trends in the development of peripheral equipment for computers

G. Haas

681.14

The heart of a data-processing system, the central processor, may be regarded (figuratively) as occupying a really central position. Round about it we then have the various "peripheral devices", some of which serve to form the connection between the central processor, the "computer" proper, and the outside world, while the rest have additional functions such as the storage of large amounts of data.

At present, data-processing systems normally involve close collaboration between man and machine. The data to be processed must therefore be transformed from the form most suitable for man to the language which the machine can understand. This "adaptation" is carried out by an input unit, which at present normally involves the use of an intermediate input medium. The inverse problem involved in the output of the data was solved long ago by using a printer or an optical indication device.

Apart from process control, with which we are not concerned in this connection, the applications of electronic computers may be divided into two main groups: the solution of technical and scientific problems, and commercial data processing. These two groups may be distinguished in many cases by comparing the number of operations required of the central computer between the incoming and outgoing flow of data. In scientific problems, the amount of data fed into the computer is often small, much calculation must be done, and the amount of data flowing out of the machine is again small. In most commercial applications, the situation is just the reverse: large amounts of data are fed into and flow out of the central processor, but the processing of the data is relatively simple. It is clear that in the first case high demands must be made on the central computer, and less high ones on the input and output units. In the second case, the processing rate is often determined by the peripheral equipment. Only when numerous peripheral devices or external data sources are used with the central computer are really high demands made on this central computer in commercial data processing.

The simplest *input device* is a keyboard, e.g. that of a typewriter, book-keeping machine or window machine. Punched-tape readers extract the information from punched tape prepared elsewhere and normally feed the data to the working memory of the central processor; the same is true of the punched-card reader. The often unnecessarily circuitous path from a written or printed text to punched tapes or cards can now be avoided by means of character-reading units, which can feed printed data directly to the central processor. Finally, a teleprinter or telephone line can serve as the input unit, allowing information obtained at another place (e.g. a branch of the main firm) to be placed directly at the disposition of the data-processing system.

The *output unit* delivers the results produced by the data-processing system, printed on paper (by a tapewriter, tabulating machine, high-speed printer), on punched paper tape (by means of a tape puncher) or on punched cards (by means of a card puncher). Finally, the results can be transmitted directly, e.g. by a telephone line, to an outside office or another data-processing system.

Apart from the input and output units and a number of other devices for use with punched cards (e.g. mixer, sorter and duplicator), the most important peripheral equipment is the *mass memory*. Such a memory differs from the working memory in having a much larger capacity (say 10^8 characters), which must however be paid for by a longer access time (say 100 to 1000 ms). This group includes the magnetic-tape memory and the mass memories with random access, such as the drum, disc and magnetic-card memories. (Strictly speaking, the punched-card sorter is also a mass memory, with a capacity as high as one wishes to make it but with a very long access time.)

Characteristics of peripheral equipment

The input-output media

The most important media of information in the peripheral equipment are paper or cardboard and magnetic surfaces. Use is also made of photosensitive or thermoplastic films, especially for the mass memories.

Dr. G. Haas is a research worker at the Hamburg laboratory of Philips Zentrallaboratorium GmbH.

Paper serves as the support for written and printed data, or contains the data in the form of a pattern of holes (punched tape or cards). Magnetic surfaces are used widely on tapes, drums, discs or cards for the storage of large amounts of data. They are also used as an easily readable medium (e.g. on account cards for automatic book-keeping in banks).

Punched tape [1] is the cheapest data carrier (about $ 1 per 100 000 characters) and allows the simplest mechanical input for computers. It is therefore widely used for programmes and sub-programmes for technical and scientific problems. A further advantage, which accounts for its increasing use for commercial data processing, is the fact that it can be prepared simply and cheaply. It is therefore admirably suited for decentralized data-processing systems: it can on the one hand be prepared *in situ* with simple equipment, and on the other hand be fed relatively quickly into large central data-processing systems. In all cases where the data is initially typed out, the punched tape can be automatically punched at the same time, by coupling the typewriter to a tape puncher (e.g. as in the "Flexowriter"). Punched tape thus allows the data to be fed more or less directly into the data-processing system. A further advantage of punched tape is its high storage density of about 1200 bit/cm^3. Punched tape is also gaining importance in data transmission, especially over telephone lines, because of its low cost and because the maximum transmission rate of about 300 characters per second can be handled by fast punched-tape equipment. The disadvantages of punched tape are that it cannot be sorted, that the speed of the tape punchers limits the output speed of large systems, and that the information cannot be erased.

The most important input-output medium at the moment is the *punched card*. The punched card is used so widely because of its low cost (about $ 2 per 100 000 characters), the ease with which it can be sorted, and above all the fact that it has already been used for decades for the mechanization of accounting work of all kinds. When electronic computers began to be used for commercial purposes in the '50s, they were mainly intended as auxiliary equipment for existing punched-card departments (e.g. the IBM 604 electronic card-puncher). With the improvement of electronic computers, the electronic equipment came to occupy a position about equivalent to that of the punched-card equipment, which is about the situation today in commercial data processing. The happy combination of the punched card with electronics is one of the decisive reasons for the success of data-processing systems such as the IBM 1401. The punched card may be expected to maintain its importance for many years. As disadvantages of the punched card we may mention its relatively low mechanical strength, its low processing speed and — as with punched tape — the fact that information cannot be erased.

Magnetic surfaces have a large storage capacity per unit area. They are subjected to little (in magnetic tape and card memories) or no (in drum or disc memories) mechanical wear, and can be erased and rewritten as often as desired. The capacity of a magnetic memory can be very large; when the storage medium can be exchanged, as in magnetic tape and card memories and certain disc memories, there is no limit to the capacity. The processing rate is appreciably higher than for punched tape and punched cards.

The access time

An important characteristic of the peripheral memory is the access time. As is known, this is the time which elapses from the moment when an address is requested to the moment when this address is available for writing or reading. In general, the access time increases with the capacity of the memory. The rate of this increase depends on the storage system used and on the extent of the auxiliary electronic equipment.

The drum memory, originally intended for use as a working memory in the central arithmetic unit, had one or more heads for each track, each head being associated with a certain amount of electronic equipment. Storage capacities of the order of magnitude required for mass memories (10^8 to 10^{10} bits) could only be obtained in this way at the cost of an excessively high expenditure on electronic equipment. In order to obtain a large storage capacity with a reasonable amount of electronics, one must either decrease the number of tracks or use one head for several tracks. Both solutions however increase the access time.

The first approach led first to the magnetic-tape memory and later to the magnetic-card memory: the information recorded on e.g. 8 tracks passes under 8 stationary heads. If one wishes to combine a small number of tracks and a large storage capacity in this way, the tracks must be very long. With magnetic tape, this leads to very long access times (e.g. 100 s); but in the magnetic-card memory the access time is reduced to about 0.2 s by "cutting and stacking" the magnetic tape. The second approach involves moving the read-write heads as well as the storage medium, as is normal in disc and large drum memories.

In general, the average access time t_z consists of half the time taken to cover the total length of the track, $t_s/2$

[1] W. Eicken, H. K. Schuff, W. Henning, H. Pärli and K. Gautzsch, Der Lochstreifen in informationsverarbeitenden Systemen, Elektron. Datenverarb. Beiheft 4, 1964.

plus the average time needed to move the head to the required track, t_k:

$$t_z = \frac{t_s}{2} + t_k.$$

In the magnetic-tape memory, for example, $t_k = 0$, and the access time corresponds to half the time needed to cover the tracks. The same is true of a drum memory with one head per track. In disc memories and drum memories with fewer heads than tracks, the access time is equal to half the period of revolution of the drum or disc, plus the time needed to move the head to the required track. As we have already mentioned, the access time depends on the storage capacity and the number of read-write channels. A large-capacity memory can therefore be better characterized by the capacity divided by the access time multiplied by the number of read-write channels: this ratio gives the amount of data which can be read out or written in per second per read-write channel.

The flow of data

A further important criterion of peripheral equipment is the data flow rate in characters per second. This can vary by orders of magnitude in the various types of equipment. The slowest input is furnished by a keyboard, with about 10-15 characters/s, while disc and drum memories can give rates of up to about 10^6 characters/s. *Table I* gives a survey of the maximum data flow rate possible at present with the various types of peripheral equipment, arranged in order of increasing magnitude.

Table I. Representative values (in characters per second) of the speeds of various types of peripheral equipment for electronic data-processing systems.

Keyboard	15
Tape puncher	300
Card puncher	800
Punched-tape reader	1 800
Character reader	2 000
High-speed printer, mechanical	2 500
Punched-card reader	3 000
High-speed printer, electrographic	60 000
Magnetic-tape unit	150 000
Disc memory	1 000 000
Drum memory	1 000 000
Data transmission	2 000 000

Present-day trends in the development of peripheral equipment

Apart from the further improvement of existing peripheral equipment itself, the trend in this field is influenced by:

a) the study of the application of electronic data-processing systems,
b) the development of control units for the data flow between the central computer and the peripheral equipment,
c) efforts to reach agreement on a universal "interface" between the central computer and the peripheral equipment.

Decentralized data processing

The first decade of electronic data processing has brought enormous advances in circuit and storage techniques. It may be stated today that the technical progress of computers has outstripped the recognition of where and how they can best be used, and the necessary adaption of administrative organization. It has become more and more clearly recognized in recent years that the design and construction of newer and faster machines is not enough. What we need for the further expansion of electronic data processing is much more the study of its applications, in particular the investigation of processes with large flows of important data. It has been found in this connection that while in the initial period of data processing too much stress was laid on centralization, it has in many cases been found to be more efficient to divide the work over a number of places which work together with a central data-processing system. This trend towards long-distance data processing brings new tasks for decentralized input and output units, external checking of errors and data transmission. Just as it is possible today to dial a desired telephone number via the central exchange, so it seems not at all impossible that we shall be able to ask a central data-processing system questions over long distances via a keyboard, or get it to carry out a given calculation, and get the answer or results back by teleprinter [2].

Creation of a "universal interface"

The range of applications of a data-processing system depends on the flexibility with which the central computer can be combined with various types of peripheral equipment. Modern computers are thus characterized by a large number of input and output channels for various peripheral devices, including data transmission (e.g. the Siemens 3003, UNIVAC 490 and the IBM-System/360). It is important that the different input and output channels should be largely independent of one another, and that it should be possible to carry out several programmes at the same time, with optimum time sharing. In such data-processing systems the control of the data flow between the central processor and the peripheral equipment is of great importance. The aim of recent research is therefore to develop suitable control units for this purpose [3] 14] and to agree on a "universal interface" between the central processor and the peripheral equipment [5].

The aim of a universal interface is to make possible a uniform method of connecting different peripheral devices to the central computer: it should be possible to connect any peripheral device to any central computer. If r central computers could be connected to p peripheral devices via a universal interface, a maximum of r matching circuits would be required on the computer side and a maximum of p matching circuits on the peripheral side.

Before we can create a universal interface, we must have a complete knowledge of the signals passing through this interface. Closer examination of the various types of peripheral equipment shows that the control signals used in the different cases are very similar; but differences are found in the *number* of control signals needed and in the speed of the various types of peripheral equipment (cf. *Table I*). Basically, an interface that would be suitable for the most complicated type of peripheral equipment (e.g. a magnetic-tape memory) and the fastest type (e.g. a disc memory) should be suitable for simple peripheral equipment too. Having regard to the great expense which would be involved in the adoption of a really universal interface, it would seem to be advisable to divide the peripheral equipment up into groups according to speed and the number of control signals needed.

Trends in the development of peripheral equipment

Here we may distinguish two basic directions: on the one hand the development of new peripheral equipment to solve problems arising in new applications, and on the other hand the improvement of existing peripheral equipment, in particular as regards its speed, reliability and economic operation. Various output devices, e.g. the high-speed printer, already have such a high output rate that any further increase in the speed seems to be unnecessary. But since in commercial data processing the output data are generally distributed again, there is in general no difficulty with the evaluation of the data here.

New peripheral equipment

At present, only exceptionaly can information be fed into data-processing systems in its original form; generally the data must be hand-punched on tape or cards before the computer can handle it. The development of a character reader [6] is aimed at eliminating this bottleneck in data processing. We still have no idea when the ideal solution, namely the reading of handwritten records, will be a practical possibility. Development in this direction is proceeding step by step, starting with the writing of data in such a form that they can be evaluated by "simpler" reading equipment. The mark-sensing process, in which for example cards are automatically punched at positions indicated by hand-written crosses, may be regarded as a first step in this direction [7] [8] [9]. The next step is the reading (by a magnetic or optical method) of more or less stylized characters. Magnetic character recognition is already very widely used, especially in the American banking world, while the reliable optical reading of normally printed characters is now also possible in practice. Since written records are produced in many places and in many different ways (e.g. by a simple typewriter), but processed in relatively few places, ways of increasing the flexibility of the reading equipment are being looked at so that the demands made on the printing of the characters can be reduced. Another aim is to increase the reading speed. Since in most cases reading equipment is used for large numbers of documents, it is sufficient if the reading time is small compared to the time necessary for the transport of the documents (e.g. for sorting). The next stage in the development may be expected to be the reading of handwritten numbers [10].

A great step towards full automation would be the *automatic recognition of speech*. There is little point in direct dictation into data-processing systems, as this is too slow, but it would then be possible to feed a magnetic tape, recorded wherever one chose, directly into the machine. Attempts to achieve the automatic recognition of at least the ten spoken decimal digits have been going on for a long time [11] [12].

New demands on peripheral equipment are made by the development of *long-distance data processing*. Since this involves the input and output of relatively simple data at a large number of places, what we are interested

[2] J. M. Unk, Isys-Memorandum 62-M7.

[3] J. W. L. Jones, Input/output control systems, Joint Comp. Conf., Edinburgh 1964.

[4] A. W. Nicholson, Peripheral transfer system for a fast computer, Proc. IEE **111**, 15-26, 1964.

[5] IEC Document 53A (Secretariat) 2, Sept. 1963.

[6] I. Sieburg, Verfahren und Möchlichkeiten zur automatischen Zeichenerkennung, Nachrichtentechn. Z. **14**, 349-357, 1961.

[7] Optical reading machine, Data Processing **6**, 24-33, 1964.

[8] C. M. B. Reid, A reader for hand-marked documents, Electronic Engng. **33**, 274-278, 1961.

[9] F. A. Frankl, Transcribing field markings by optical scanning, Electronics **34**, Nr. 31, 49-51, 1961.

[10] E. C. Greanias, P. F. Meagher, R. J. Norman and P. Essinger, The recognition of handwritten numerals by contour analysis, IBM J. Res. Devel. **7**, 14-21, 1963.

[11] K. H. Davis, R. Biddulph and S. Balashek, Automatic recognition of spoken digits, Bell Lab. Rec. **31**, 52, 1953.

[12] E. E. David, Jr., and O. G. Selfridge, Eyes and ears for computers, Proc. IRE **50**, 1093-1101, 1962.

in is small, simple devices, either cheaper versions of existing equipment (e.g. number-checking equipment or numerical indicators), or new devices for cheap data collection, e.g. the combination of a keyboard with a small magnetic-tape device. Long-distance data processing involves important problems in data transmission, the development of coupling equipment (MODEM) and the detection of errors. Particularly for data transmission over ordinary telephone lines, the occurrence of error bursts makes the detection of transmission errors important. Having regard to the costs involved, in particular the costs of renting the telephone line which are still high, it must be carefully decided in each case whether the saving of time is really worth the extra costs compared to the transport of punched tape e.g. by railway or car. Data transmission also increases the demands mᵃde on punched-tape equipment. In order to avoid buffer memories, it must be possible to read asynchronously and to repeat the reading of an information block after the occurrence of an error by using a bidirectional transport.

In the field of *mass memories*, superconducting films show promise for the future; they offer high storage densities and a very short access time (10 μs) [13]. For the moment, however, magnetic surfaces hold the field, with disc and magnetic-card memories coming to play an increasingly important role. One is here concerned with the development of equipment which can read and write large amounts of data in as short a time as possible. Two directions may be distinguished here: first, devices are being developed for moving the magnetic head to the desired address as quickly as possible. Memories working on this principle are characterized by relatively few heads, each of which requires a relatively large actuating mechanism. The other possibility is to use a large number of slower and simpler selection mechanisms [14].

At present, the central processor must be used for certain sub-programmes involved in data processing, e.g. the sorting or recording of magnetic tapes, the writing of data from punched cards or tape on to magnetic tape and the like. For most applications this is still the most efficient solution. However as soon as the central computer, e.g. as a result of increasing long-distance data processing, is fully occupied by its proper tasks, small peripheral devices for carrying out these additional tasks will gain in importance: although each type will only have a limited function, production in large quantities will allow them to be made available at a competitive price.

Improvements in existing equipment

The large difference between the operating speed in the central computer and the rate of data flow in present-day input and output equipment means that the latter do not usually work directly together with the central computer, but via a magnetic-tape unit, whose operating speed is better matched to that of the computer ("off-line operation"). An increase in the speed of the peripheral equipment is thus apparently desirable. In particular the present trend towards real-time data processing (collection, transmission, processing and distribution of data in a closed system, immediately after the start of a process) demands not only large memories but also fast peripheral equipment. However, now that several peripheral devices of the same kind can be connected to the same central processor, the absolute speed of the peripheral equipment is no longer such a pressing problem. What is then more important is the ratio of the speed to the cost. Seen from this point of view, a printer that works twice as fast but costs twice as much will only in exceptional cases represent an advance.

Finally in this connection, a few words about reliability. In general, peripheral equipment is electro-mechanical. The failure rate of most electronic components used nowadays is, after a certain initial period (in which components with hidden defects are eliminated), more or less constant. The failure rate of mechanical components on the other hand increases steadily with the operating time, as these parts are subject to continual wear. Moreover, the failure rate of mechanical components increases sharply with increasing speed. From this point of view it would seem that in fast peripheral equipment, mechanical parts should be replaced by electronic components as far as possible. On the other hand it should not be forgotten that the maintenance which is generally possible with mechanical equipment (the testing and replacing of parts after a certain period of operation) is not in general possible with electronic circuits.

Some possibilities for increasing the speeds of peripheral equipment

Mechanical limitations on the speed

The speed of the peripheral equipment is limited by the mechanical properties of the data carrier on the one hand (tensile strength, elastic limit, etc.) and by the forces and moments of rotation involved in non-uniform motion on the other hand. While in magnetic-tape units the maximum load of the magnetic tape has just about been reached, punched-tape equipment has not yet reached the theoretical limit set by the strength of the tape [15].

Non-uniform motion may be necessary for two reasons:

1) The mechanical recording of data. Here we can distinguish between the motion of the recording device proper (e.g. the punch and the matrix in a puncher or the type and the hammer in a printer) and that of the recording medium. The printing process in the high-speed printer is so fast that the hammer can strike a uniformly rotating type drum or type chain ("on-the-fly printing"). In the punching of tape or cards, on the other hand, not only the punch or matrix moves non-uniformly, but also the recording medium: during punching the medium must be stationary, to prevent the holes from being torn. If the punching could (at least partially) be replaced by a cutting process, a continuous movement of the cards or of the tape would then be possible.

Table II. Mechanical action in modern peripheral equipment for electronic data-processing systems.

	Motion of recording device required for reading or writing	Motion of input-output medium	Start-stop requirement for data carrier
punched-tape reader	none	uniform	yes
punched-card reader	none	card by card	no
tape puncher	non-uniform	discontinuous	yes
card puncher	non-uniform	discontinuous	no
character reader	none	document by document	no
mechanical high-speed printer	non-uniform	discontinuous	yes
electrographic high-speed printer	none	can be uniform	yes
tape memory	none	uniform	yes
magnetic-card memory	none	card by card	no
large drum memory, disc memory	non-uniform	uniform	no

2) The repeated starting and stopping of the recording medium. This start-stop requirement is found with data carriers in the form of tape: for example, punched tape must be able to be stopped and started between two characters, and magnetic tape must be able to be stopped and started within the block interval. In this case, not only the inertia of the recording medium but also the inertia and moments of inertia of the transport device have to be overcome.

Table II gives a survey of the action of recording devices and media normally found in peripheral equipment nowadays. Of the two discontinuity conditions for the recording medium, that mentioned under point (1) is the more difficult to fulfil. While the stopping, e.g. of punched tape, only requires that the maximum stopping distance should be less than 1.5 mm, the tolerance on the location of holes must be within ±0.05 mm for punching.

The desired increase in speed may be achieved in one of four ways:

a) reducing or eliminating the "dead weight",
b) reducing the necessary stroke,
c) replacing non-uniform motion by uniform rotation,
d) use of non-mechanical, e.g. electronic, devices (without in general appreciably increasing the cost of the recording medium).

A typical example of the last point is provided by the non-mechanical, in particular electrographic, printing processes which have been used more and more of recent years [16].

Unfortunately, the processes used so far do not furnish any copies; and in the commercial applications where these high speeds are especially needed one or more extra copies are normally essential. One is therefore compelled to print the extra copies in succession on the same printer (at the expense of the effective speed) or in parallel in an extra wide printer (which costs more, but the costs are much less than proportional to the number of copies needed). Finally it is also possible to copy after printing, for which purpose special equipment has been developed [17].

Low-inertia drives

If a mass M_{n}, e.g. part of a punched tape, has to be moved over the distance s, it is not generally possible to apply the force used for this purpose directly to the mass M_{n}. In general, the transmission of the force from

[13] J. A. Rajchman, Computer memories — possible future developments, RCA Rev. **23**, 137-151, 1962.

[14] Isys Internal Technical Report 63-N10.

[15] A. D. Booth, The transport of paper tape in digital computation, J. Brit. I.R.E. **20**, 657-660, 1960.

[16] E. Webster, The impact of non-impact printing, Datamation **9**, Nr. 9, 24-30, 1963.

[17] Copying printed computer results, Data Processing **6**, 34-37, 1964.

the point where it is produced involves an additional mass M_p, as shown in *fig. 1.* In most cases, in fact, $M_\mathrm{p} \gg M_\mathrm{n}$.

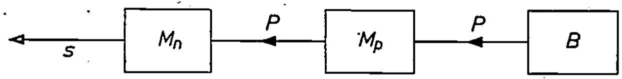


Fig. 1. Transmission of the force P to the mass M_n to be moved, via the mass M_p.

The movement of the mass M_n over the distance s by means of the force P (which we assume to be constant over this distance) requires the time:

$$t = \sqrt{2s\frac{M_\mathrm{p}}{P}\left(1+\frac{M_\mathrm{n}}{M_\mathrm{p}}\right)}\,.$$

In order to move the mass M_n as quickly as possible for a given distance s, we must try to make the ratio P/M_p as large as possible. Mechanical methods (e.g. an eccentric with a connecting rod) can give values of about 0.5 kp/g.

Electromagnets

An *electromagnet* is a useful drive mechanism in this respect. Suitable dimensioning can give a P/M_p ratio of about 5 kp/g. Because of the saturation of the armature necessary for this purpose, this method is only suitable for relatively small values of the distance s (some tenths of a millimetre). A further advantage is that the magnet is practically free from wear. Such magnets are particularly well suited for driving pressure rollers and brakes in tape-transport mechanisms. Their efficiency can be increased still further by electronic means, by appropriate design of the energizing circuits of the magnets (e.g. with thyristors).

Shock waves

Much larger P/M_p ratios can be obtained by letting *shock waves*, e.g. those produced by a spark discharge, do work. With an electrical energy of only a few tenths of a Ws, P/M_p ratios of the order of 10^6 kp/g can be achieved in this way. Such shock waves can be used e.g. instead of punches for the punching of tape. Because the punching speed is increased by about three orders of magnitude, the tape can be moved continuously during punching, so that the intermittent movement of the tape as found in present-day tape punchers which also limits the speed, can be done away with. The additional start-stop requirement can be fulfilled with the above-mentioned electromagnetic drive about an order of magnitude faster than with the discontinuous drive used at present.

A further application of such shock waves is ossible in high-speed printers. In mechanical printers, the speed is mainly limited by the lack of synchronization between the hammer stroke and the position of the type-drum or chain. Since the accuracy with which the shock waves can be produced is roughly an order of magnitude greater, printing speeds of about ten times the present value (rows/s) should be possible. If on the other hand an increase in the speed is not desired, use can be made of the faster printing process by applying it to serial printing, which makes the printer and (particularly in simple machines) the computer itself cheap.

Stroke reduction

In order to achieve a quick movement of the mass M_n for a given value of the ratio P/M_p, the stroke (the distance s) must be made as small as possible. This is not in general directly possible: in the case of punched tape, for example, the stroke is fixed by the standardized displacement of 1/10 inch.

Fig. 2 shows the principle of a drive mechanism for punched tape in which the discontinuously moving parts have a stroke of no more than 0.05 mm. The discontinuous feed of the punched tape is achieved by a suitable combination of two brakes, a capstan wheel and a pressure roller.

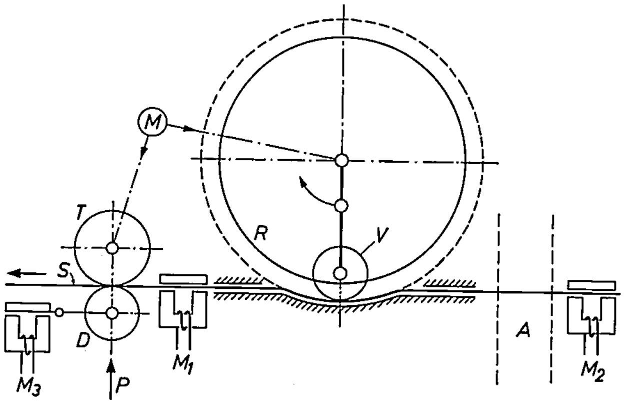


Fig. 2. Principle of the transport mechanism for punched tape. The motor M drives the capstan wheel T and the transport wheel R. When the system is at rest, the brakes M_1 and M_2 are applied and the transport roller V has produced a depression in the tape S. When the tape has to be moved forward, the magnet M_3 is energized, and forces the pinch roller D against the wheel T which now tries to transport the tape in the direction of the arrow. Now brake M_1 is also released and the tape is pulled tight. Then brake M_1 is applied again, and brake M_2 released. The following actuation of the transport roller V produces just enough slack in the tape for the following transport. Punching of the tape takes place at A.

A disc memory with rotating heads

Fig. 3 shows the principle of a disc memory in which the heads do not move discontinuously in a radial direction but in continuous circular paths over the disc. No masses have to be accelerated or braked to select

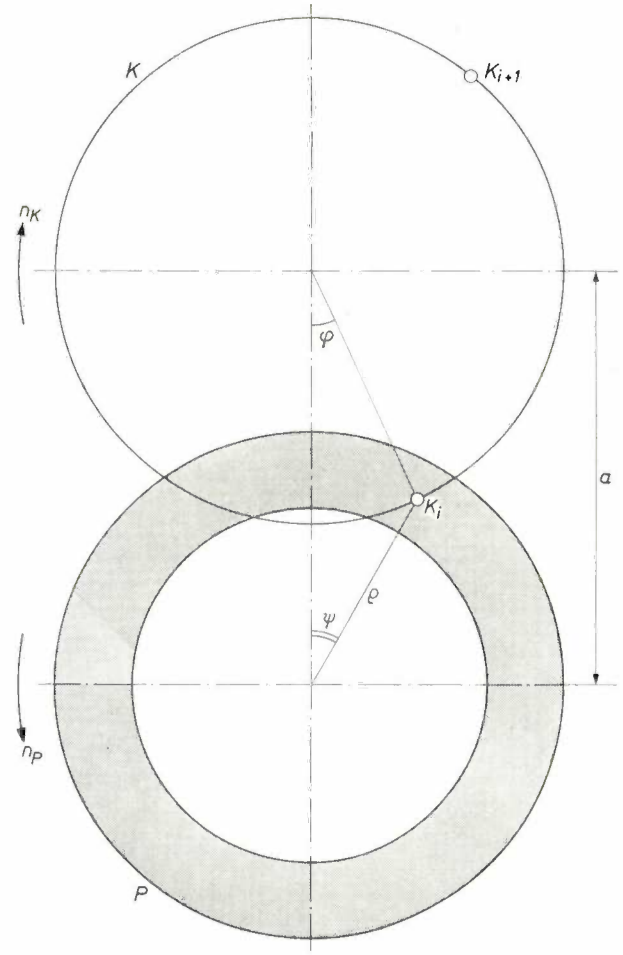


Fig. 3. Principle of the disc memory with rotating heads. The memory disc P rotates uniformly at a speed n_P. One or more magnetic heads K_i are mounted on the disc K, which turns at a uniform speed of n_K. The distance a between the axes of the discs is chosen so that the disc K overlaps the memory disc sufficiently. When K_i represents a block of p magnetic heads (arranged along a radius), then the block K_i allows the reading of p tracks in parallel on passing over the (shaded) region of the disc P. The selection of a given address on the memory disc simply involves choosing the appropriate time, i.e. the appropriate values of the angular positions ϕ and ψ of the memory disc and scanning disc respectively; when the appropriate angles are reached, the read or write channel is activated. The tracks, with radius vector ρ, have the form of parts of hypocycloids, whose parameter is determined by the ratio of the speeds of the two discs.

a given track, as all tracks are covered one after the other during one revolution of the disc. The tracks have the form of parts of hypocycloids, whose parameter is determined by the relative rates of rotation of the memory disc P and the magnetic heads (which are fixed to the disc K).

Such a memory can be designed in various ways, depending on the number m of memory discs arranged round the axis of the disc carrying the magnetic heads, and the number of heads k. The different designs differ in the amount of mechanical and electronic equipment required, in storage capacity and access time. The capacity (C in bits) of one side of a memory disc, divided by the access time τ, depends on the other parameters of the memory as follows:

$$\frac{C}{\tau} \approx \frac{pmk}{7} f,$$

where f is the reading or writing speed in bits/s.
The C/τ ratio reaches favourable values at high values of f (which may be limited by the permissible speed of the disc carrying the heads, or by the auxiliary electronic equipment).

Summary. The task of peripheral equipment lies mainly in the input and output of data and in the storage of large amounts of data. This equipment can be characterized by the nature of the input-output media, the access time, and the flow of data which can be attained. Apart from improvements in the peripheral equipment itself, in particular as regards increasing the speed and the reliability and decreasing the costs, development is at present in the direction of extending the applications of data-processing systems and rationalizing their operation, e.g. with the aid of a universal "interface" between the central processor and the peripheral equipment. The speed of the peripheral equipment is limited by the mechanical strength of the recording medium or by the forces and moments of rotation which occur in connection with the non-uniform motion of the masses of the equipment itself. In certain cases the speed can be increased by reducing the inertia of the moving parts or the stroke, by replacing non-uniform motion by uniform rotation or by using non-mechanical recording effects.

Data checking during input and transmission by means of one or two check characters

G. Renelt and J. Schröder

621.391.088.6

Introduction

With the continual increase in the amount of information to be processed by mechanical and electronic systems, the detection of errors in this information is becoming more and more important. These errors are particularly often produced during the manual input in data-processing installations or the transmission of data via information channels, subject to interference.

To be able to detect errors, a certain amount of redundancy to the information in question e.g. in the form of one or more "check characters", or the redundancy may be present in the first place. It might be thought that the simplest way of checking would be to repeat the original information; however, this method not only involves a lot of work, but also gives relatively poor results, as errors are often repeated exactly the same way the second time. It is better to form the check character from the information according to a fixed rule. In the case of information consisting only of decimal digits, a well known example of such a rule is that the cross-total of all the original digits is taken as the check character and added on after the initial number.

If the information consists of a specific series of characters which represents a definite concept, it is worth while including the check character as an integral part of the series. This applies for example to all reference numbers, such as account numbers, depot numbers, police registration numbers, job numbers, customer numbers, article numbers, order numbers, contract numbers, wagon numbers, etc. The numbers in question can then be checked by means of the standard rule immediately they are fed into a data-processing device (such as a book-keeping machine or accounting machine). One way of doing this is to form the check character again from the original information and to compare it with that contained in the reference number. If the two do not agree, the number in question contains an error, and appropriate steps can be taken to rectify the error, e.g. by feeding the number in a second time.

Now when a check character is used, it can happen, especially when several characters are wrong, that the wrong series of characters gives the same check digit as the right one. Such errors cannot then be detected. If for example we consider a number of three decimal digits with an additional decimal check character, we have a thousand possible numbers and only ten possible check characters. The relationship between the check character and the number proper is thus highly ambiguous, so that very many errors cannot be detected. It should also be realized that errors can also occur in the check character during input or transmission. This means that even an elaborate data-checking system using an adequate number of check characters cannot guarantee to detect all possible errors. However, a good check system should allow as high as possible a percentage of the errors occurring to be detected without too much effort.

In practice, not all possible errors occur with an equal frequency: in fact, nearly all real errors belong in general to a few well defined classes. In particular, most input and transmission errors are normally single-character errors; these are errors in which only one character of the sequence in question is wrong. In data processing by human operators, in particular data input via a keyboard, the most common error apart from the single-character error is the interchange error, these are errors in which two adjacent characters change places.

Now the simple sum of all the digits of a number is capable of detecting all single-character errors. Let us for example consider the sequence of digits 87029, whose cross-total (sum of the digits) is 26. If any one of the digits is increased, the sum of the digits will be greater; and if any digit is decreased, the sum of the digits will be less. Each single-character error will thus lead to another value of the cross-total.

If we consider all five-figure numbers from 00000 to 99999, the sum of the digits can take any value between 0 and 45, while longer numbers can give even bigger digit sums: with numbers of 12 digits or more, the digit sum can have over a hundred different values. If one were to use the sum of the digits directly as the check character, one would thus have to add several digits on to the number to be checked, which is an unnecessarily complicated way of checking for all single-character errors. In fact, the same result can be obtained with one single check character, by a suitable reduction of the number of values which the check character can assume.

This reduction is possible by the use of finite number

Dipl.-Math. G. Renelt and Dipl.-Ing. J. Schröder are research workers at the Hamburg laboratory of Philips Zentrallaboratorium GmbH.

systems, in particular the residual-class system. The residual-class system arises just like the system of natural numbers (0, 1, 2, 3, . . .) by the repeated addition of 1 to the preceeding number, starting from 0, with the single difference that in the residual-class system the initial value 0 is repeated after a finite number (M) of steps, so that further addition gives a cyclic repetition of the same series of numbers. The straight line usually used for representing a series of numbers can here be better replaced by a circle divided into M equal parts. Each number M gives another residual-class system, e.g. $M = 12$ for the hours divisions on a clock, and $M = 60$ for the minutes.

Each residual-class system contains exactly M different numbers, which can be represented by 0, 1, 2, 3, . . ., $(M - 1)$. The definition of this system allows the addition of two numbers, just as with natural numbers. If the addition of the corresponding two natural numbers gives a result greater than or equal to M, M must be subtracted from this to keep the result in the residual-class system. If one starts in general from normal natural numbers, all large numbers can be reduced to values less than M by single or repeated subtraction of M. All natural numbers which differ by M or a multiple of M are said to be equivalent modulo M. For example, the addition of 8 and 7 in the residual-class system with $M = 12$ gives $8 + 7 = 15 \equiv 3 \bmod 12$. (A simple method of reducing large numbers quickly in this way is normal division by M, where all we are interested in is the remainder, in the above example $15/12 = 1$, remainder 3.)

If however one remains in the residual-class system, which can very easily be done in practice e.g. by use of a ring counter with M stages, we get the correct result directly. In the case of the above example, the ring counter would contain a total of 12 stages, and would count 7 stages starting from positions 8 (9, 10, 11, 0, 1, 2, 3), arriving at position 3. If we have to form the cross-total of the number 87029, we must now count 2 stages further, arriving at position 5, and finally 9 stages, bringing us to the final position 2 ($14 \equiv 2 \bmod 12$). The number to be checked plus the check character 2 can then be written 870292. It is easy to see that this reduced sum of the digits of the number also allows all single-character errors to be detected, since any increase or decrease of a single digit (by at most 9) will cause the ring counter to turn through more or less the same number of positions, so that it all cases another final result than 2 and hence another check character will be produced.

Many numbers will give the final result 10 of 11 in this way. If one wants to use a single check character in these cases too, one can use special signs, e.g. + for 10 and — for 11 (so that the final number with check character reads e.g. 87026—). Another simple possibility is to represent all ring-counter positions and residual-class numbers by letters (e.g. A, B, C, . . . instead of 0, 1, 2, . . .). In our example, the number with check character would then be 87029C.

It is also useful to replace the residual-class numbers by letters or more general symbols when series of characters other than decimal digits (or including characters other than decimal digits) have to be checked. For example, if we again use the correspondence $A \longleftrightarrow 0$, $B \longleftrightarrow 1$, $C \longleftrightarrow 2$, . . . and we want to provide a check for the sequence CAB by the "sum of the digits", we find $2 + 0 + 1 = 3$, so that the final sequence with check character is $CABD$. Naturally, this unnecessary reversible correspondence (coding) can also be used for decimal digits, or for binary coded characters.

If letters have to be checked, we need at least 26 different residual-class numbers for the check character, so that M must be at least 26 ($M \geqslant 26$). If mixed "numbers" consisting of letters, decimal digits and possibly special signs such as + — . ,: etc. have to be checked, an even larger residual-class system must be used (e.g. $M \geqslant 42$), while a smaller system will do for the checking of pure decimal numbers. The residual-class system modulo 10 is of special interest in this connection; though as will be shown below it is in general advisable to make M a little larger, even in this case.

General rules for forming the check characters

The simple sum of the digits, even when reduced modulo M, allows all single-character errors to be detected, but does not show up interchange errors. In order to provide a check for the latter errors, a more refined formation of check character must be employed. This can be done by making the correspondence between the characters of the series to be checked and the numbers of the residual-class system different for the different places of the sequence to be checked (e.g. for the first place $A \longleftrightarrow 1$, $B \longleftrightarrow 2$, . . .; for the second place $A \longleftrightarrow 9$, $B \longleftrightarrow 5$ and so on *ad lib.*). If the coding of the various possible characters z in the ith place is described by the function $F_i(z)$ and if in a given sequence (with m places without check character) the character in the ith place is $z = z_i$, then the value P of the check character for this series can be found from the following generalized cross-total:

$$P \equiv \sum_{i=1}^{m} F_i(z_i) \bmod M. \quad . \quad . \quad . \quad . \quad . \quad (1)$$

If for example a check character has to be provided for the sequence $HHDS$, using a residual-class system modulo 29, and if H in the first place corresponds to 7,

H in the second place to 16, D in the third place to 26 and S in the fourth place to 3 ($F_1(H) = 7$, $F_2(H) = 16$, $F_3(D) = 26$, $F_4(S) = 3$), then the check character will be the character (e.g. W) corresponding to the value $P = 23$ ($P = 7 + 16 + 26 + 3 \bmod 29$).

In general, the coding will be chosen so that a check is provided against as many errors as possible. In order to detect all possible single-character errors at the ith place, the image function $F_i(z)$ must be different for all different values of z:

$$F_i(z) \not\equiv F_i(z') \bmod M \text{ for } z \neq z'. \quad . \quad . \quad . \quad (2)$$

The exchange of the characters z and z' between positions i and j will on the other hand be detected only when the *differences* between the image functions are different:

$$F_i(z) - F_j(z) \not\equiv F_i(z') - F_j(z') \bmod M \text{ for } z \neq z'. \quad (3)$$

Various codings which have to a large extent the desired properties can be obtained by the introduction of weighting factors. Here one starts with a fixed coding of all possible characters z in terms of the residual-class numbers Z (e.g. $A \longleftrightarrow 1$, $B \longleftrightarrow 2$, ... or with digits $1 \longleftrightarrow 1$, $2 \longleftrightarrow 2$, ...). The various position-dependent codings are then obtained by multiplication of the fixed coded residue classes Z with the various integral weigths G_i, giving $F_i(z) \equiv G_i \cdot Z \bmod M$.

So if a check character P is calculated according to equation (1), it is equal to the weighted sum of the coded digits Z_i, where each Z_i is counted as many times as its weight G_i:

$$P \equiv \sum_{l=1}^{m} G_i \cdot Z_i \bmod M. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (4)$$

If for example one uses the weights 1 and 2 alternately, the check character is formed from $P \equiv Z_1 + Z_2 + Z_2 + Z_3 + Z_4 + Z_4 + Z_5 + \ldots \bmod M$, which is at least sufficient to detect all interchange errors.

If all single-character errors are to be detected, eq. (2) must always hold. This means in the case of the weighted sum of eq. (4) that the weights G_i must all be prime with respect to M, i.e. G_i and M must have no common factor but 1 (e.g. for $M = 6$, the possible values of G_i are 1 and 5). All interchange errors between the ith and jth places are however only detectable if the *difference* of the weights ($G_i - G_j$) is prime with respect to M. It is therefore highly advisable to use a residual-class system modulo a prime number M. If one then simply uses different weights in forming the check character according to eq. (4) ($G_i = 1, 2, \ldots, M - 1$ in any desired order), one check character is enough to detect all single character errors and all interchange errors in a sequence of up to $M - 1$ characters with 100% certainty.

An equally effective check of other types of errors too can thus only be obtained by taking additional measures. For example the addition of a second check character formed from the simple cross-total modulo M will allow all double errors, i.e. arbitrary errors at two arbitrary positions in the sequence, to be detected with certainty.

A residual-class system modulo M, where M is not a prime number, does not give such a perfect check. The residual-class system modulo 10 is naturally particularly interesting in this connection. The only weights which come into consideration for the formation of a weighted sum in this case are 1, 3, 7 and 9, if we want to be able to detect all single-character errors. Since the differences between these weights are not prime relative to 10, five of the 45 possible interchange errors cannot be detected, e.g. the confusions of 05 with 50 and 16 with 61. In this case it is better to use the original equation (1) for forming the check character. If for example one chooses alternately the two codings ($F_{2s} = F_1$, $F_{2s+1} = F_2$):

$$\left.\begin{array}{c|cccccccccc} z & 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ \hline F_1(z) & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 0 \\ F_2(z) & 2 & 4 & 6 & 8 & 1 & 3 & 5 & 7 & 9 & 0 \end{array}\right\} \quad . \quad . \quad . \quad (5)$$

then all single-character errors and all interchange errors with the exception of the confusion of 89 and 98 can be detected. In this connection, the shift $F_1(z) = z + 1$ makes no difference to the detection of interchange errors; but it has the advantage of showing up errors caused by the addition or omission of zeros at the end of a series. The second coding given above is obtained from the first by a modified doubling procedure which adds the number of times the ring counter goes from M to 0 ($F_2 \equiv 2F_1 + [0.2F_1] \equiv [2.2F_1] \bmod 10$), where [...] means that the decimal parts of the resulting numbers are omitted. Other codings are possible with this residual-class system modulo 10, but at least one interchange error must remain undetectable if one wishes to detect all single-character errors with a single check character. (For if we sum all $F_1(z)$ or $F_2(z)$ for $z = 0$ to 9, we obtain a value 5 modulo 10 since according to (2) all residual classes must occur; the difference $\Sigma\{F_1(z) - F_2(z)\}$ is thus zero, and equation (3) cannot always be satisfied.)

The efficiency of the check

As has already been mentioned, a good check system will detect most errors, but not all possible ones. There always remains a certain residual error probability, given by the ratio of the average number of undetected errors (wrong character sequences) to the total number of character sequences. This probability is what we are finally interested in in practice. If we relate this probability to the probability that errors (detected or undetected) occur at all, the overall reduction factor R thus obtained gives a quantitative measure of the efficiency of a check system [1]:

$$R = \frac{\text{residual error probability}}{\text{total error probability}}$$

$$= \frac{\text{mean number of undetected errors}}{\text{total number of errors}} \quad . \; . \; . \quad (6)$$

If the value of the reduction factor is expressed in percent, it gives the number of undetected errors per 100 errors. For a given error distribution and a given check system the reduction factor is a constant which no longer depends on the frequency of errors. If the various errors which occur can be divided into classes $k = 1, 2, \ldots$ occurring with relative frequencies b_k, and if the percentage of undetected errors of class k is given by the individual reduction factor r_k, then the overall reduction factor is given by:

$$R = \sum_k b_k \cdot r_k, \quad . \; . \; . \; . \; . \; . \quad (7)$$

$$r_k = \frac{\text{mean number of undetected errors of class } k}{\text{total number of errors of class } k}.$$

For performing the calculation, the error distribution for the application in question must be known. For example, if we consider the errors due to noise in data transmission, we may assume that each character has a certain distortion probability β independent of the other characters being in error or not. If the total length of the character sequence is n, then the relative frequency $b_k = b(k)$ of errors involving k characters in a single sequence is given by the Bernouillian distribution

$$b(k) = \frac{\binom{n}{k} \beta^k (1-\beta)^{n-k}}{1-(1-\beta)^n}$$

$$\approx \frac{1}{k}\binom{n-1}{k-1} \beta^{k-1} \text{ for } n\beta \ll 1 . \quad . \quad (8)$$

In particular:

$$b(1) = 1 - \frac{n-1}{2}\,\beta .$$

It follows that with a low noise level (β small) single-character errors will predominate, and multiple-character errors will be extremely rare. If however systematic errors occur, deviations from distribution (8) may be expected. The error distribution b_k must then be determined from the analysis of exhaustive observations. Examples of such errors are cross-talk and short interruptions in telephone lines, errors in the filling in of forms or in input via a keyboard. The error distribution normally associated with the punching of punched cards is [2]:

Type of error K	Relative frequency b_k
1. One digit wrong	76.5 %
2. Dependent error in two digits (e.g. interchanged)	5.9 %
3. Independent error in two digits	7.3 %
4. Digit(s) inserted or deleted	9.5 %
5. Other errors	0.8 %

All errors of the first two types can be detected by a check character according to eq. (4) in which all weights G_i are different and M is a prime number; in that case $r_1 = r_2 = 0$. The other errors are reduced to a relative frequency of $1/M$ by such a check: $r_3 = r_4 = r_5 = 1/M$. The value of the total reduction factor for a residual-class system modulo 11 is thus 1.6 %, which means that if we assume the error frequency to be 0.1 %, the residual error probability is equal to 1.6×10^{-5}, i.e. after such a check with one check character we may on the average expect only two out of 125 000 numbers (character sequences) keyed in to be still wrong. If the number length is constant, the errors of type 4 can be completely eliminated by counting the positions in each number in addition to the above check ($r_4 = 0$). If moreover the check is extended by means of a second check character formed from the simple cross-total of the digits, all errors of type 3 are also eliminated ($r_3 = 0$, $r_5 = 1/13^2$), and, for example, with a residual class system modulo 13 a check is obtained that is still about 350 times better. The residual error probability is thus 4.6×10^{-8}, which means that on the average more than 20 million sequences must be fed in before one will be wrong.

The basic principles of the technical realization

We shall now discuss the technical realization of data-checking devices which are based on the above considerations. Such a device has two tasks to perform: in the first place all data which must later be processed without error must be provided with a check character; and secondly all character sequences already containing a check character must be checked for errors. However, it is not necessary to have two separate devices for this: a device which determines the check character can also be made to carry out the checking with very little modification.

As we have seen, a particularly effective check is obtained when the check character is found from the

[1] H. Aulhorn, H. Lange and H. Marko, Probleme und Anwendungen der Datenübertragung, Elektron. Rechenanlagen 3, 148-159, 1961.

[2] Kontrolle mit Hilfe von Kontrollziffern, Bürotechn. Sammlung 1, Sept. 1956.

weighted cross-total (4), all weights being different, in a residual-class system modulo a prime number M. The formation of all the partial products $G_i \cdot Z_i$ needed for this purpose would require a lot of equipment. However, the desired result can also be obtained more simply according to the following principle: the digits Z_1, Z_2, Z_3, ... of the sequence in question are fed one after the other (if necessary via a parallel-series converter) into the checking device, the main component of which is a modulo-M counter. This counter is initially in the zero position. In the first step the digit Z_1 is added and then the counter position is doubled (modulo M), e.g. by repeated addition of the position; in the second step the digit Z_2 is added and the content of the counter is again doubled, and so on. If for example the number fed in is 11728, a modulo-11 counter will take up successively the following positions 1,2; 3,6; 2,4; 6,1; 9,7. The check character is finally taken as the M-complement of the last counter position in the present example thus 4, which is the 11-complement of 7. In general, this process gives the check character:

$$P \equiv -\sum_{i=1}^{m} 2^{m+1-i}\, Z_i \bmod M. \quad . \quad . \quad (9)$$

Comparison with (4) shows that this is equivalent to forming the weighted cross-total with the weights:

$$G_i \equiv -2^{m+1-i} \bmod M \quad . \quad . \quad . \quad . \quad (10)$$

Unlike the case in (4), however, the weights here also depend on the length m of the character sequence. This has the advantage that the addition or omission of zeros at the end of the sequence has the effect of changing the weights and thus in general the value of the check character, so that these errors are also detected. On the other hand the introduction of zeros at the beginning of a sequence has no effect, as is desired. This allows the numbers to be filled up with zeros to a constant length, if necessary.

For the residual-class systems in which we are chiefly interested, viz modulo $M = 11$ and $M = 13$ (or also $M = 19$ and $M = 29$), equation (10) gives all different weights when the maximum length of the character sequence does not exceed $M - 1$ characters; optimum conditions are thus obtained in this case should the character sequence be even longer, moreover, a cyclic repetition of these weights is obtained without any special measures being taken. For those residual-class systems on the other hand where equation (10) does not give all possible weights, the doubling of the counter contents may be replaced by multiplication by a factor q, so chosen that all weights are obtained from the various powers of q. (One speaks in this case of q as a primitive root of the residual-class system, while the entire process is known as Horner-polynomial formation.) In general this too gives optimum checking.

The actual checking function of the device, by which it detects any errors there may be in character sequences already provided with a check character, is made particularly simple because when forming the check character according to equation (9), the complement has also been calculated. The check sign is handled just like all other characters, and when the whole sequence has been fed in correctly the final counter reading is 0; any other reading indicates an error. In our example, the check character 4 of the complete number 117284 is added to the previous contents of the counter (7), giving a reading of 0, which stays the same on doubling, thus indicating that the number is correct.

The basic principle of an electronic check counter generator working according to this system is shown in *fig. 1*. The data are processed with the aid of pulse groups, the number of pulses in a group corresponding to the value of the character in question. In the ith step, the pulse generator *1* forms a group of Z_i pulses. These are fed to the input *B1* of the modulo-M counter, and move it on a corresponding number of steps. Moreover, pulse generator *2* forms the number of pulses needed to double the contents of the counter, and feeds them to the counter via a second input *B2*. This can simply be realized by feeding another M pulses to the counter via *B1* and switching on pulse generator *2* when the counter reaches the zero position, so that pulse generator *2* passes all the following pulses coming from pulse generator *1* (with a certain delay time) to the *B2* input of the counter as well. Since the pulses still reaching the counter via *B1* return it from 0 to its original position, this arrangement ensures that all these pulses are fed twice into the counter, so that the desired doubling (modulo M) is obtained. The complement of the counter position is displayed and yields the check character. After all steps $i = 1, 2, \ldots, m$ have been carried out, i.e. after the whole sequence has been fed in, the desired check character can be read off from the display device. If on the other hand a complete sequence plus check character was fed in, a departure of the final counter position from 0 indicates an error.

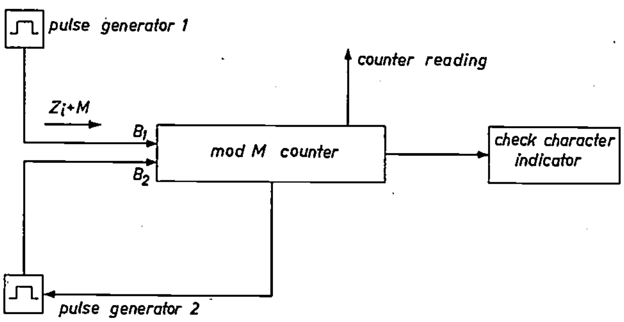


Fig. 1. Basic principle of the technical realization of a check character generator.

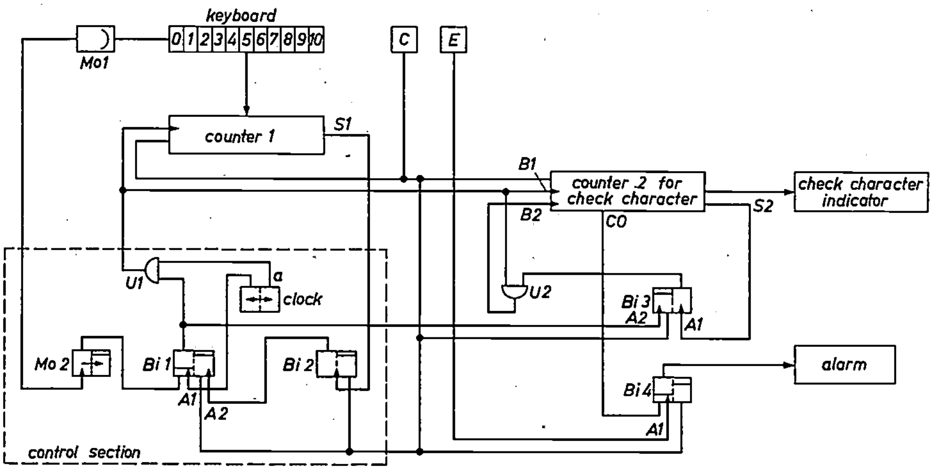


Fig. 2. Verification unit with a single check character modulo 11.

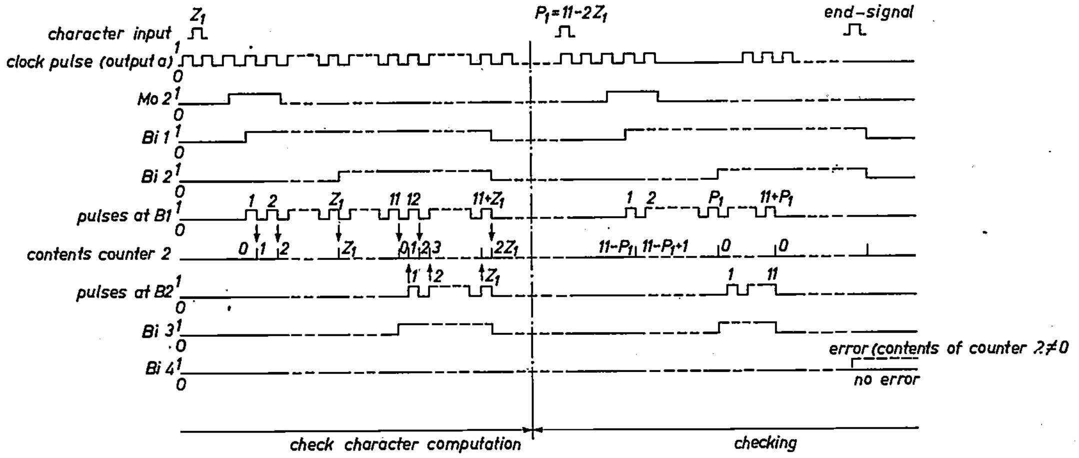


Fig. 3. Timing diagram for circuit of fig. 2.

An electronic verification unit with a check character modulo 11

The electronic construction of a number-checking device working with a check character in the residual-class system modulo 11 is shown in the block diagram of *fig. 2*. We shall explain the internal operation of this device with reference to fig. 2 and the pulse scheme of *fig. 3*, which shows the variation with time of the state of the various stages of the apparatus (timing diagram).

Calculation of the check character

The data and check characters are fed into the device via a keyboard. Naturally, this keyboard can be replaced by the output of a punched-card, punched-tape, book-keeping or other data-processing machine, so that the verification unit can be combined with various data input and output devices. The keyboard controls a counter *1* which can take up 11 different positions and may for example be in the form of a four-stage binary counter with feedback. When the key corresponding to the first character Z_1 is struck, counter *1* is set in position 11—Z_1, and a pulse, delayed by *Mo1* is simultaneously applied to the trigger input of the monostable circuit *Mo2.* The delay via *Mo1* is designed to ensure that the control input of *Mo2* does not receive the pulse until counter *1* has reached its final position. During switching over of *Mo2,* input *A1* of the bistable circuit *Bi1* is prepared so that when the next clock pulse applied to the input goes from "1" to "0" the input *A1* will switch the bistable circuit over. This opens the And gate *U1,* so that the clock pulse can pass to the input of counter *1* and the input *B1* of a counter *2*, which may be constructed exactly like *1.* As the counter *1* is in position 11—Z_1, a carry pulse appears at its output *S1* after exactly Z_1 clock pulses, switching over the bistable circuit *Bi2.* A second carry pulse appears at *S1* after another 11 clock pulses, which switches *Bi2* back to its original position, while *Bi2* delivers a pulse to input *A2,* switching bistable circuit *Bi1* back to its original position. The And gate *U1* is thus closed, and the connection with the clock-pulse generator broken. A total of $Z_1 + 11$ pulses thus appears at the input *B1* of counter *2*. Since this counter was initially in position 0, the carry pulse produced at output *S2* after 11 count pulses causes *Bi3* to switch over via *A1.* This opens And gate *U2* for Z_1 pulses, while the return of *Bi1* to its original position after $Z_1 + 11$ pulses returns *Bi3* to its original position via *A2.* The dynamic counting input *B1* is operated when the input clock pulse goes from "1" to "0", while the counting input *B2* is operated when going from "0" → "1". While the And gate *U2* is open, therefore, each count pulse is counted twice in counter *2*. At the end of the counting process, the desired result $2Z_1$ mod 11 is found in counter *2*. If a further carry pulse should appear at output *S2* during the time that the And gate *U2* is open, this remains without effect at the input *A1*. When the next character is fed in via the keyboard, counter *1* is set to the position 11—Z_2. The process described above is then repeated. $Z_2 + 11$ pulses are produced by the control unit together with counter *1*, and are fed to counter *2*. Moreover, when counter *2* reaches the 0 position the carry pulse at *S2* switches over the bistable circuit *Bi3*, so that the remaining pulses also act on *B2*. The final counter position after this process is completed is $(2Z_2 + 4Z_1)$ mod 11, and so on. After all the characters have been fed in, the calculated check character as the 11-complement of the stored counter reading is available with practically no delay. In the device described, the computation time is one millisecond (not considering the delay time caused by *Mo1* to eliminate chattering of the key contacts).

Representation of the check character

The use of the modulo-11 system allows the check characters to be represented in a particularly convenient way. We have seen that according to (9) the check character is so formed that when the number plus check character is fed in properly, the sum of the digits is always zero. This means that in those cases where the sum of the digits without the check character is already zero, the check character can be omitted. (If the numbers have to be a constant length, this can be arranged by adding a zero to the beginning of the number.) Ten digits are then enough for representing the remaining ten residual classes, e.g. the numbers 1 to 9 for residual classes 1 to 9, and $z = 0$ for residual class $Z = 10$. The latter correspondence need not however hold for all positions of the character sequence, and in particular leading zeros of the sequence can be excluded. This can be achieved by means of an extra bistable circuit. Naturally, it is also possible to use 11 different check characters, for example letters or ten digits and an extra sign (e.g. —).

Verification

A series of characters containing a check character is verified by pressing the key "*E*" (End of series) after all the characters have been fed in. If the series in question contains an error, counter *2* will not be in position 0 after input of the check character (the signal "0" will be present at the counter output *CO*), so that the check bistable circuit *Bi4* is switched over via input *A1*, and an "error" signal is given. This signal can be optical or acoustic, and may also lock other devices. The key "*C*" (clear) serves to make the device ready for use after switching on, by bringing all bistable circuits into their start position, or when an error signal is given to clear counter *2* and reset bistable *Bi4* thus switching the error signal off.

Realizations in modulo-13 and other residual-class systems

It is often desirable to protect various special signs (e.g. decimal point, plus sign) against error as well as the ten decimal digits. Also, *one* check character is not enough when low residual probabilities are required, or when the input data contains a large number of errors. For these cases, a checking device like that described below, working with *two* check characters ($q_1 = 2$; $q_2 = 1$) in the residual-class system modulo 13, is suitable. The check characters are here formed according to the equations:

$$P_1 \equiv -\sum_{i=1}^{m} 2^{m+1-i}\, Z_i \bmod 13$$

and

$$P_2 \equiv -\sum_{i=1}^{m} Z_i \bmod 13.$$

The circuit diagram of the complete device including checking unit is shown in *fig. 4*. The checking unit and gates *U3* and *U4* can be dispensed with if the device is only to be used for forming the check character.

The procedure followed to obtain the check characters is the same as described above for the modulo-11 apparatus. The only difference is that counter *3* is connected in parallel with counter *2*, and is used for calculation of the second check character.

For the checking process, on the other hand, it is necessary that the two check characters P_1 and P_2 fed in via the keyboard at the end of the sequence should be led separately to the appropriate counter, i.e. counter *2* for P_1 and counter *3* for P_2. The direction of the first group of pulses to counter *2* and the second group to counter *3* is achieved with the aid of bistable circuit *Bi4*, which is switched over each time bistable circuit *Bi1* switches over. Before checking begins, i.e. before the check characters are fed in, the check key ("*K*") must be struck. This switches *Bi4* and *Bi5* over, and prepares the apparatus to receive the next character as a check character. When P_1 is fed in, *Bi*$_4$ returns to its original position and the P_1 count pulses produced pass to counter *2* only, as *Bi4* has blocked the And gate *U3* via the Or gate *O1*. Check character P_2 is then fed in, and a group of $13 + P_2$ count pulses are produced, which pass through the And gate *U3* (now open) to counter *3* only, since *U4*

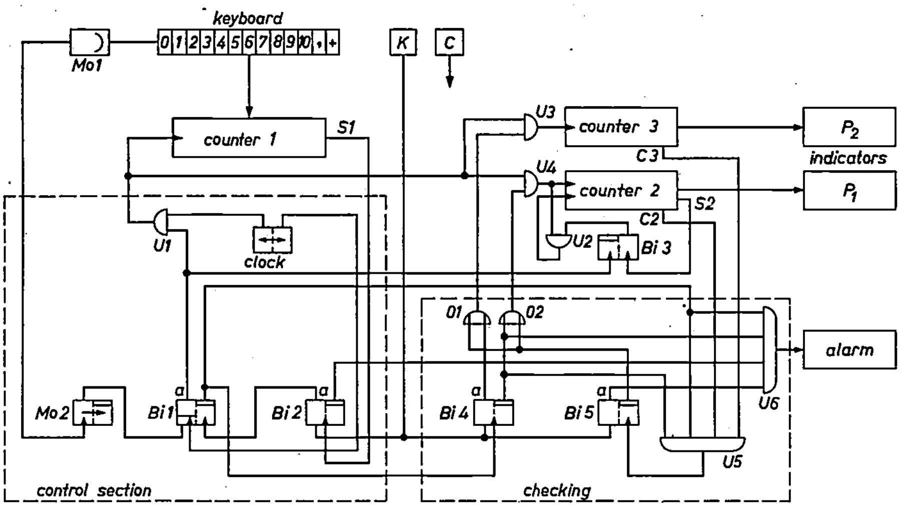


Fig. 4. Verification unit with two check characters modulo 13.

is closed. After these pulses have been processed, and if the check characters fed in agree with those previously calculated, all counters are at 0 and the output lines *C2* and *C3* deliver the signal "L". As a result, when *Bi1* is switched back, "0" → "1" passes through And gate *U5* to the input of *Bi5* and switches this back to its initial position. If counter *2* and *3* are not both in the zero position, the And gate *U5* will be blocked, bistable circuit *Bi5* will not be switched back, and an error signal "1" will appear at the output of the And gate *U6*. If no error is indicated, the device is automatically in its initial state after the check character has been fed in, and the next character sequence can be fed in. For the sake of clarity, the connections of the clear key "*C*", which has the same function here as in the modulo-11 apparatus, to the various bistable stages are not shown.

The verification unit can also have a counter for the number of places in the character sequence. This ensures that all errors produced by the addition or omission of digits will be detected with 100% efficiency. However, the unit described above can give the same efficiency without any modification, if a coding is used in which all the digits or characters correspond to non-zero residual classes. For example, we can use the above-mentioned identical coding of the digits from 1 to 9, corresponding to residual classes 1 to 9, the character $z = 0$ corresponding to residual class $Z = 10$ (e.g. also with the exception of leading zeros at the beginning of the sequence). All errors caused by the omission or addition of a digit can then be detected with absolute certainty with the aid of the second check character.

Checking devices with one or two check characters in any other residual-class system can be realized in just the same way. In particular, quite good checking can be achieved in the residual-class system modulo 10 with one check character, using the modified doubling of (5). In contrast to a prime-number system, this system with one check character (still formed on the basis of (1)) will detect nearly 98% of all adjacent exchange errors and more than 90% of all simple jump errors as well as all single-character errors. Bigger residual-class systems can however be used with advantage in other cases, depending on the number of symbols to be checked and the number of check characters available.

Summary. The timely detection of errors in data for example during input and transmission, is possible when the data are provided with check characters, which form an integral part of the data. These check characters should be formed from the original data in such a way that as many errors as possible are detectable. The conditions under which all commonly occurring types of errors, such as single-character errors and "interchange errors", can be detected with certainty are given. The residual error probability, which is inevitable in any checking method, can in this way be made very low (e.g. about 10^{-5} with one check character and less than 10^{-7} with two). The use of suitable principles for the formation of the check characters allows this process to be carried out very simply in practice by an electronic system. Data verification units working on these principles are described, which serve not only to determine the check character but also to carry out the actual checking of the data, by means of a zero check. This gives optimum checking with one or two check characters using very simple equipment.

A new system of digital circuit blocks for industrial measuring and control equipment [*]

D. Gossel, G. Kaps and W. Schott

621.374.32

Introduction

The characteristic of digital information processing installations is that in their construction a small group of basic circuits is used, each of which occurs in large numbers. The group of basic circuits is called a system.

In order to enable designers to realize a given logical design without going into electronic details, several manufacturers have developed systems consisting of "And", "Or" and "Negation" circuit blocks as well as bistable circuits (flip-flops) for counting and storing purposes [1].

Although it should be possible to realize any logical design using such a system, circuit limitations must be considered, and extensive load tables limit the possibilities of combining the various blocks. The number of prohibited combinations increases with the number of different blocks, making the load tables more and more complicated. Amplifiers for decoupling purposes — e.g. *P-N-P* and *N-P-N* emitter followers — have to be used, although they have no logical function, and this is at variance with the original aim [2]. Consequently there are systems in which the total number of blocks required for a given problem bears no relation to the number that have a logic function. This influences not only costs, but also reliability, since an increase in the number of blocks also causes an increase in the failure rate.

Efforts have been made to avoid these difficulties by decreasing the number of basic circuits. Switching algebra provides a solution through the use of De Morgan's formula [3]:

$$x_1 + x_2 + \ldots + x_n = \overline{\bar{x}_1 \cdot \bar{x}_2 \cdot \ldots \cdot \bar{x}_n},$$

$$x_1 \cdot x_2 \cdot \ldots \cdot x_n = \overline{\bar{x}_1 + \bar{x}_2 + \ldots + \bar{x}_n}.$$

This formula gives the relationship between the AND and OR operation and the Negation. It may be seen from this that the one operation can always be replaced by the other together with the negation.

The NAND and NOR techniques often used at present go another step further. Here it is possible by using suitable combinations of the AND operation together with negation, or the OR operation together with negation, to make only one type of block suffice [4]. *Fig. 1* shows the AND and OR operation realized with NAND- and NOR-blocks. Each NAND and NOR symbol represents one basic circuit containing at least one transistor together with resistors, capacitors and diodes.

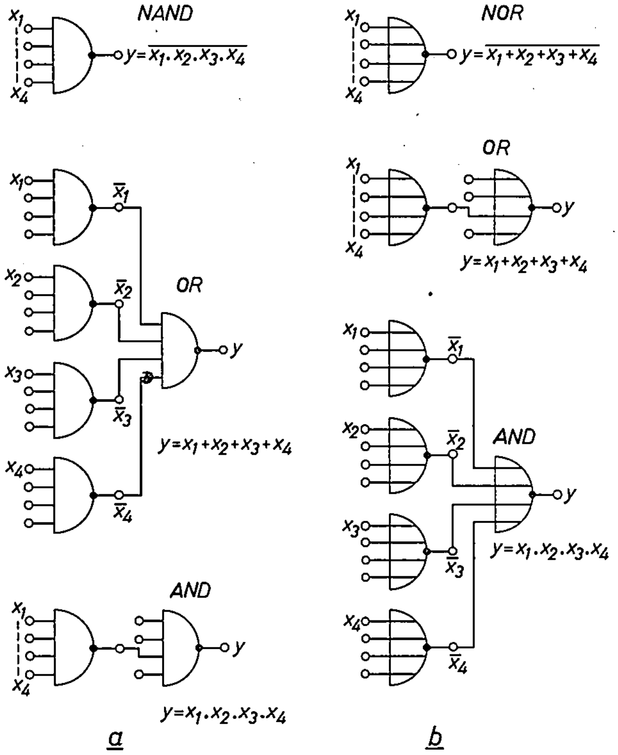


Fig. 1. Logic circuits for the OR operation and the AND operation, built up of *a*) NAND circuits, *b*) NOR circuits.

A second advantage of using only one type of block for all logical circuits is that the load tables in the NAND and NOR systems reduce to one single load rule. It is only necessary to know how many NAND or NOR circuits can be reliably controlled by a previous NAND or NOR. There are no other forms of loading.

A serious disadvantage however is the large number of blocks to be used for certain operations. In order to realize an OR operation, one relatively expensive NAND block is necessary for each input (fig. 1*a*). The AND

Dipl.-Ing. D. Gossel, Dipl.-Ing. G. Kaps and Dipl.-Ing. W. Schott are research workers at the Hamburg laboratory of Philips Zentrallaboratorium GmbH.

[*] *Editorial note:* An article by E. J. van Barneveld, Philips ICOMA Division, Eindhoven, will be appearing shortly in this journal, describing another system of digital circuit blocks, designed for more universal use than the system presented here, and which is now being put into production by the ICOMA Division.

operation, however, can always be realized by two NAND blocks regardless of the number of inputs (within certain limits). This also holds true when only NOR blocks are used; but in this case AND and OR operations are exchanged (fig. 1*b*) [5].

The new building-block system

The new system of digital circuit blocks described in this paper can be regarded as an attempt to reach a good compromise between these two methods. One of the characteristics of the new system is the use of only two basic circuits, an active one, and a passive one. The active block contains a circuit in diode-transistor logic (DTL). With the logic convention chosen ("1" ≙ 12 V, "0" ≙ 0 V) this block works as a NOR circuit (fig. 2). The passive block is an AND circuit (also called an AND gate) with two inputs (fig. 3). The number of inputs can within wide limits be increased by means of additional diodes. Using the passive AND block it is not necessary to realize the AND operation with NOR blocks in the expensive way shown in fig. 1*b*.

This building-block system which is especially designed for industrial measurement and control purposes also has the following characteristics.

1. There are no emitter followers. Only one type of transistor and two types of diode are used.

2. A standardized voltage supply is used (± 12 V ± 5%). This value is considered a good compromise between

a) a high voltage, giving low sensitivity to interference and great freedom in combining various numbers and types of block — i.e. wide signal tolerances are permitted —, and

b) a low voltage which suits the low maximum permissible voltage for available transistors and gives low dissipation.

3. The maximum switching frequency is 80 kc/s, the maximum counting frequency 30 kc/s. (In some cases the maximum counting frequency may be 80 kc/s.) In industrial applications the electronic circuits are mainly used together with moving parts which have a certain mass. Experience here has shown that resolving power (accuracy) and speed call for a counting frequency of 10 kc/s at the most. A frequency of 30 kc/s is thus adequate for quite extreme requirements.

4. Asynchronous or synchronous modes of operation are optional. For simple problems involving low-speed counting, the well-known asynchronous method may be used. Here for example a bistable circuit is triggered by a previous one which also has to supply the switching energy. The maximum counting rate is determined by the sum of the switching times of all stages and thus decreases with an increasing number of stages. The maximum load per stage is much reduced in asynchronous counting techniques.

When somewhat higher demands are made, the synchronous counting mode is to be preferred because of its great advantages, although a few more blocks are needed. Here all the stages of a counting circuit receive switching pulses from a common clock-pulse generator. Each stage contains a separate signal input $\overline{S}$, and the voltage applied to this input determines whether or not a given clock pulse will trigger the circuit. This method has the following consequences.

a) The circuit is very insensitive to interference. As a result of the delaying effect of the pulse gate controlled via the $\overline{S}$ input, parasitic pulses in the signal line will only be able to cause incorrect switching if their duration is > 5 µs, *and* their voltage-time integral is more than 60 µVs, *and* the clock pulse is received more than 5 µs after the start of a parasitic pulse. Under normal conditions it is unlikely that these three conditions will be satisfied simultaneously. In practice therefore there will be almost no interference.

b) The various stages are not subjected to a dynamic load of any significance, since the triggering energy is supplied by the common clock-pulse generator. With the exception of the clock-pulse line, which should be as short and of as low capacity as possible, all signal lines are uncritical.

c) Unlimited use may be made of the maximum switching frequency. All stages which have to be triggered are prepared during the interval between two clock pulses, and are triggered by the next clock pulse received. The switching times of successive stages are not additive. No stringent demands are made for the rise time of the signal voltage at the $\overline{S}$ input.

[1] G. Schinze, Das AEG-Steuerungssystem "Logistat", AEG-Mitt. **50**, 76-83, 1960.
W. Stübchen, Ruhende Steuerungen Logistat — die sinnvolle industrielle Anwendung kontaktloser Steuerungen, AEG-Mitt. **50**, 139-143, 1960.
Valvo-Handbuch Bausteine, Digitale Bausteine, Valvo GmbH, Hamburg 1962.
Catalogue of Akkord-Radio GmbH, Herxheim/Pfalz, Estacord — Das universelle Bausteinsystem für kontaktlose Steuerungen.
Catalogue 3.62 of Ebauches S.A., Neuchâtel, Switzerland, Transistorisierte logische Einheiten.
A. Stopp, Normalkonstruktionen der BBC-Elektronik, BBC-Nachr. **42**, 199-207, 1960.
K. Stahl, M. Syrbe, H. Lisner and G. Hanke, Grundlagen und Aufbau der BBC-Elektronik, BBC-Nachr. **42**, 208-219, 1960.

[2] W. Händler, Digitale Universalrechenautomaten; section 10.1.2.1., Vollständige Systeme, in: K. Steinbuch, Taschenbuch der Nachrichtenverarbeitung, Springer, Berlin 1962.

[3] U. Weyh, Elemente der Schaltungsalgebra, R. Oldenbourg, Munich 1960.

[4] Valvo-Handbuch Bausteine, Norbit Bausteine, Valvo GmbH, Hamburg 1962.

[5] E. Rohloff, Aufbau und Anforderungen bei kontaktlosen Steuerungen für die Industrie, Elektron. Rdsch. **15**, 99-102, 1961.

5. The circuits have been designed to allow for the most unfavourable voltages and resistances within the tolerances quoted, and for the transistor data at the end of the operating life; undisturbed operation under full load is guaranteed in the temperature range from —10 °C to +50 °C.
6. Signal tolerances:

$$\text{“1”} \mathrel{\widehat{=}} +6 \ldots + 12 \text{ V},$$
$$\text{“0”} \mathrel{\widehat{=}} \ 0 \ldots + 1.8 \text{ V}.$$

The two types of block

The *active* block (*fig. 2a*) contains a gate circuit with diodes ($D_1 \ldots D_5$, R_K, R_B), an inverter circuit and a pulse gate (D, C_P, R_S, R_T). The diode gate and the inverter circuit together form a NOR unit, so that the following relationship exists between the output C and the four inputs $B_1 \ldots B_4$:

$$C = \overline{B_1 + B_2 + B_3 + B_4}.$$

If desired, the number of inputs can be increased by connecting extra diodes to one of the inputs $B_1 \ldots B_4$. The operation of the pulse gate will be explained together with that of the counting and memory stages.

The contacts are so arranged (fig. 2*b*) that even with complicated circuits the connections between the blocks can be made without crossovers (see fig. 5).

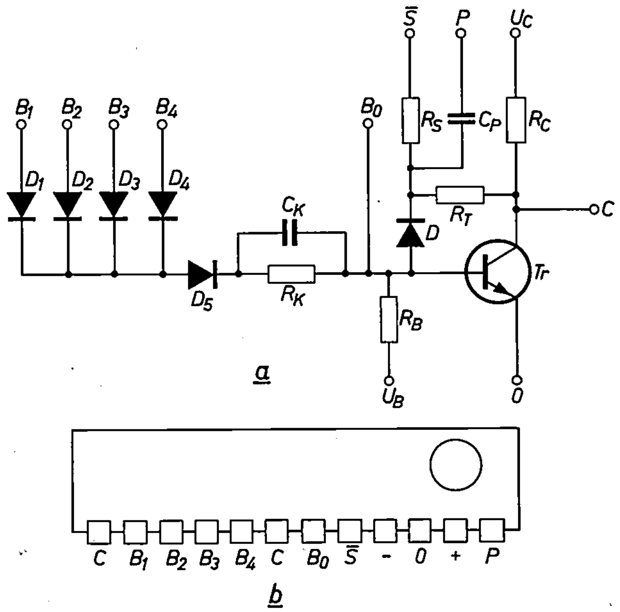


Fig. 2. *a*) Circuit of the NOR stage, *b*) arrangement of the contacts.

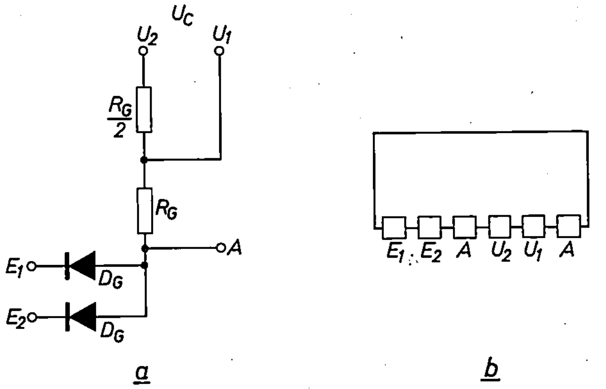


Fig. 3. *a*) Circuit of the AND gate, *b*) arrangement of the contacts.

The *passive* block (*fig. 3a*) contains a gate circuit with diodes and resistors, which gives the following AND relationship between the output A and the two inputs E_1 and E_2:

$$A = E_1 \cdot E_2.$$

The gate resistance used can be either R_G or $R_G/2$, or — by combining these resistors in series or in parallel — $3R_G/2$ or $R_G/3$, respectively. This gives the passive block a good measure of adjustment to circuit requirements in so far as loading and power consumption are concerned. An AND circuit with more than two inputs can be obtained by connecting the outputs of a number of passive blocks, while connecting the resistance of only one block to the power supply. The arrangement of the contacts is shown in fig. 3*b*.

Circuits built up of active blocks

The bistable circuit

With the known systems of blocks the counting and memory functions are realized with the aid of various types of bistable circuit, which are varied to suit the different functions required. Sometimes special blocks are provided for this purpose, while sometimes these circuits are built up from two NOR or NAND blocks [4] [5].

The bistable circuit (flip-flop) consisting of the two types of active block is shown in *fig. 4a* for the asynchronous counting mode and in fig. 4*b* for the synchronous. In both cases negative switching pulses necessary to block the conducting transistor are applied to the input denoted by P. In the asynchronous method, the pulse gate consisting of the diode D, the resistor R_T and the capacitor C_P ensures that the switching pulses can only have an effect at the base of the conducting transistor [6].

In the synchronous method negative pulses are continually applied to the P input. These only trigger the bistable circuit if the signal “0” (collector potential of a conducting transistor) is applied to the $\overline{S}$ input.

[6] G. Rusche, K. Wagner and F. Weitzsch, Flächentransistoren, page 346, Springer, Berlin 1961.
G. Haas, Fundamentals and components of electronic digital computers, page 188, Philips Technical Library, Eindhoven 1963.

[7] For details see the following publications of Philips ICOMA Division, Eindhoven: The “10” series of circuit blocks, The Icomist, No. 71, May 1964, and: Tentative data, circuit blocks, series 10, 32/189/B/E, February 1965.

The logical convention here is thus the opposite of that for the system; this is why we denoted this input by $\overline{S}$ instead of by S. The signal has to be inverted for each signal input. This is done by an active block which, apart from regenerating the potential, can also serve to carry out the OR operation which is frequently required at this place (see fig. 6). These active blocks may only be loaded with AND circuits.

If negative parasitic pulses occur in the output leads of a bistable circuit, they must be prevented from reaching the base of the conducting transistor via the internal feedback, in which case they could trigger the bistable circuit. This can simply be done by placing an AND gate in each feedback loop (decoupled flip flop) [7]. In the circuits described below this is not necessary.

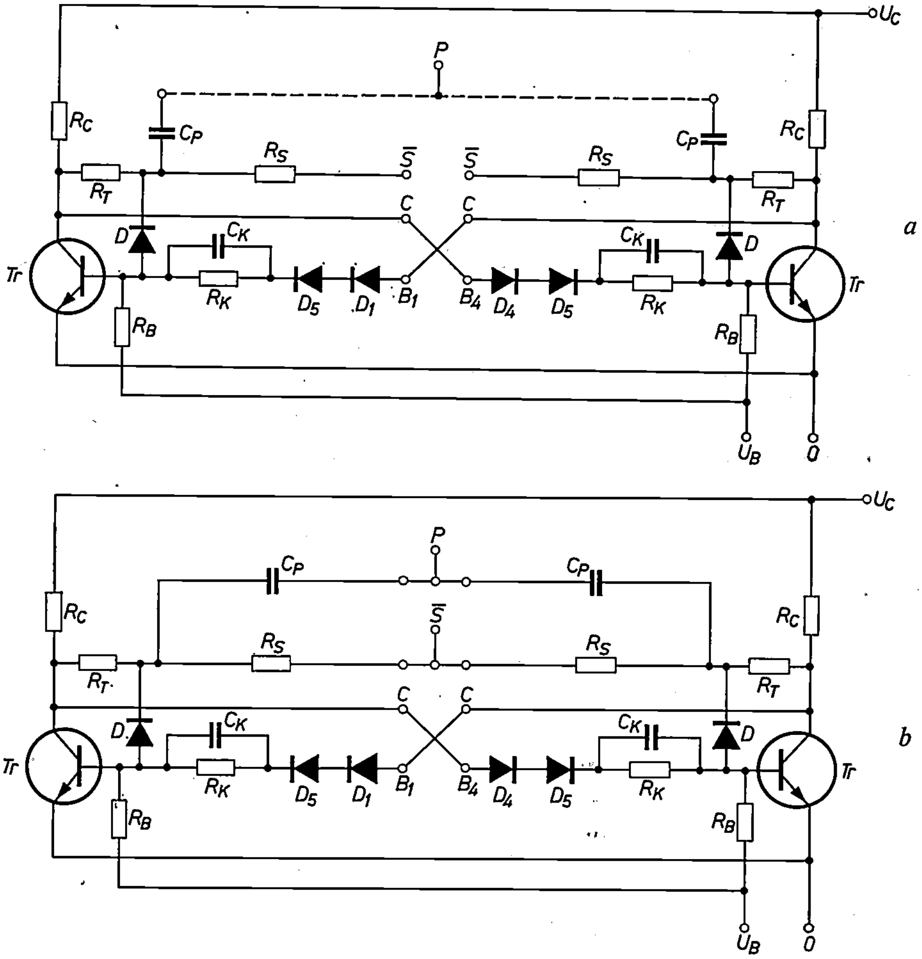
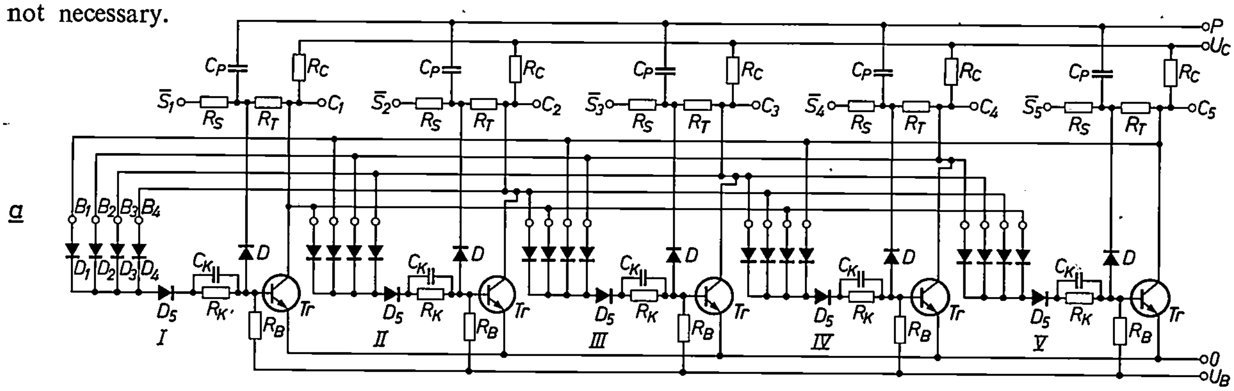


Fig. 4. A bistable circuit (flip-flop), *a*) for the asynchronous type of counting circuit, *b*) for the synchronous type of counting circuit.



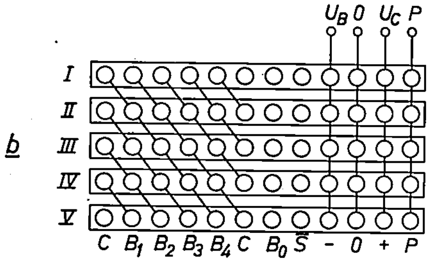


Fig. 5. Circuit with five stable states (quinary circuit), built up of blocks of the system described here. *a*) Circuit diagram, *b*) arrangement of the blocks so that connections can be made without crossover.

Polystable circuits

The new system of circuit blocks can be used not only for bistable but also for polystable circuits with a maximum of five stable states. The number of stable states possible is equal to the number of active blocks used.

In the circuit with five stable states (quinary circuit) (*fig. 5a*), the output of each of the five active blocks is connected with an input of each of the other four blocks. Since the outputs of the blocks are made double, the connections can be made without crossovers (fig. 5*b*). One transistor is always cut off, while the other four are kept conducting via the base inputs connected with the collector of the cut-off transistor. Triggering is from the central clock-pulse generator via the pulse gates, the potentials at the $\bar{S}$ inputs determining which transistor will be marked (cut off).

A quinary circuit can be turned into one with four or three stable states by the omission of one or two blocks respectively, together with those leads which only connect inputs with one another.

Table I. Code of a biquinary decade counter.

Digit	Binary stage		Quinary stage				
	C	$\bar{C}$	Q_1	Q_2	Q_3	Q_4	Q_5
0	0	1	1	0	0	0	0
1	1	0	1	0	0	0	0
2	0	1	0	1	0	0	0
3	1	0	0	1	0	0	0
4	0	1	0	0	1	0	0
5	1	0	0	0	1	0	0
6	0	1	0	0	0	0	1
7	1	0	0	0	0	0	1
8	0	1	0	0	0	0	1
9	1	0	0	0	0	0	1

The reversible biquinary decade counter

A synchronous decade counter can be made very simply by the combination of bistable and quinary circuits [8]. Since the biquinary system of counting is very closely related to the decimal — 2 and 5 are the prime factors of 10 — there is no need to modify the circuit to eliminate superfluous counting capacity. This is important, because such modifications are required for both counting directions. The reading-out of the digits is simpler than with the normal decade counters consisting of four bistable circuits. Ten AND circuits (AND gates) with an average of three inputs each are necessary to read ten digits from the positions of the four bistable circuits. In the biquinary counters, however, ten AND gates with two inputs each are sufficient. *Fig. 6* shows the logic circuit of a reversible biquinary decade counter, using the coding of *Table I.*

Before a counting pulse is fed to the circuit, it is first synchronized with the clock pulse, so that its length becomes equal to the interval between two clock pulses. Each counting pulse, for counting both forwards and backwards, prepares the binary stage B, which is then switched over by the clock pulse. The quinary stage must switch over to another position in two situations, viz 1) when the output C of the binary stage has the potential corresponding to the value "1" and a forward-counting pulse is present, and 2) when the value "1" is found at the output $\bar{C}$ of the binary stage and a backward-counting pulse is present. The counting pulses and the position of the binary stage are combined by the AND gate G_V for counting forwards and G_R for counting backwards.

In forward counting, the pulses for the quinary circuit are fed from G_V to the AND gates $G_{V1} \ldots G_{V5}$, and in backward counting from G_R to the AND gates $G_{R1} \ldots G_{R5}$. A pulse which must be counted forwards by the quinary circuit will e.g. pass the AND gate G_{V2},

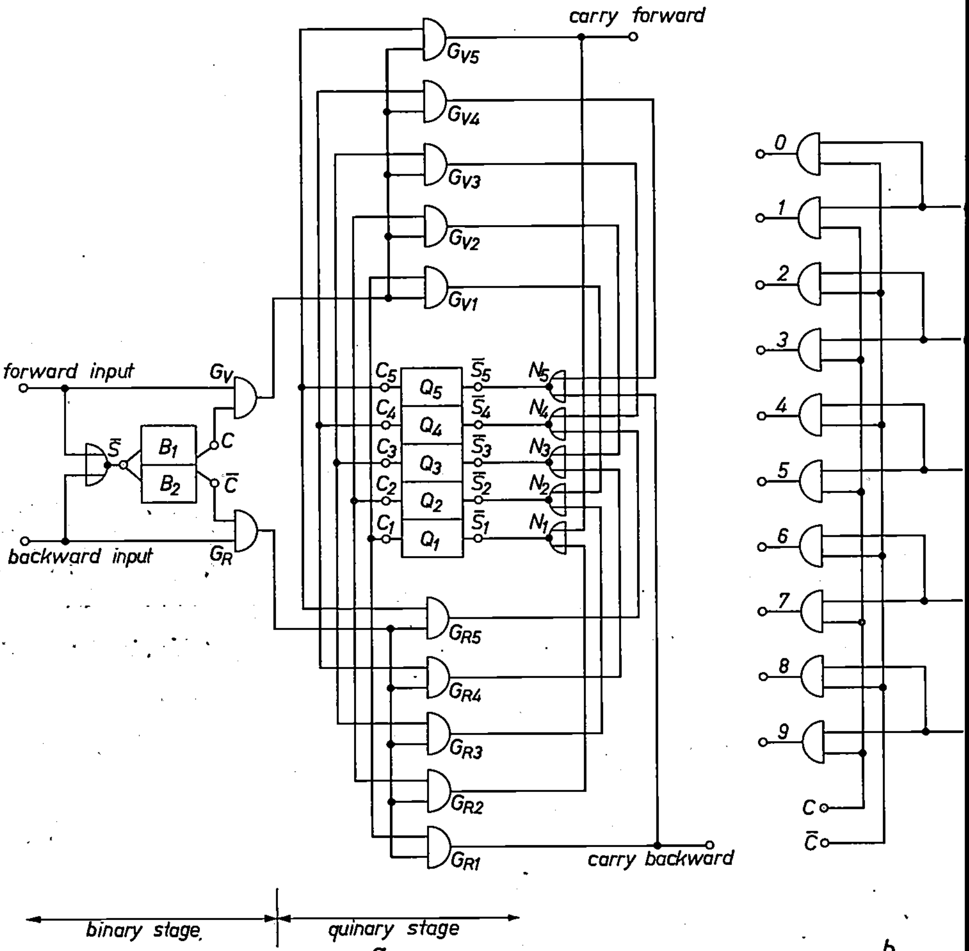


Fig. 6. *a*) Reversible biquinary decade counter. *b*) Circuit for reading-out of digits from this counter. (The P inputs of B_1, B_2 and of $Q_1 \ldots Q_5$ must be connected to the common clock-pulse line.)

which is opened by the "1" signal at the output C_2 of stage Q_2, thus priming the signal input $\bar{S}_3$ of the next stage (in the forward direction) Q_3. Similarly, with a pulse that has to be counted backwards the AND gate G_{R2} is opened and the next stage (in the reverse direction) Q_1 is primed via the signal input. The actual switching-over is always initiated by the next clock pulse to coincide with the counting pulse.

Special circuits

A multivibrator can be built up of two active blocks by connecting their outputs C crosswise via capacitors C_Z to the direct base leads B_0 and by connecting both the leads which are normally intended for the negative and the positive supply voltages to the positive supply voltage.

In the monostable circuit the static coupling — as in the bistable — is formed by the combination R_K, C_K, D_4 and D_5. An external capacitor C_Z, which must be connected between the collector C and the base input B_0, determines the delay time. The base bias resistance of the capacitively coupled transistor is connected to the positive supply voltage, so that this transistor conducts in the stationary state. The monostable circuit is triggered via the P input of the capacitively coupled block.

A Schmitt trigger is formed by connecting two active blocks in series, the emitter leads 0 being earthed via a common resistor. The threshold level can be adjusted by means of another external resistor connecting the direct base lead B_0 of the first block with the positive supply voltage. Excitation is via one of the NOR inputs of the first stage.

Loading rules

The combination of active and passive blocks to give a logic circuit must be done in accordance with the loading rules, since each block is loaded by the subsequent stages and itself forms a load for the previous stage. Because there are only two types of circuit block, the loading rules can be kept simple.

Each input $B_1 \ldots B_n$ of an active block (NOR) represents a "NOR load" for the previous stage. Each input $E_1 \ldots E_m$ of a passive block (AND gate) represents a certain number of "gate loads" for the previous NOR stage; this number depends on the choice of the total gate resistance:

Total gate resistance	$3R_G/2$	R_G	$R_G/2$	$R_G/3$
Number of gate loads k	2/3	1	2	3

These two types of load may not be discounted against one another when calculating the maximum permissible load.

With a few exceptions, which cannot be discussed within the scope of this article, the following simplified loading table is obtained (where allowance must still be made for the restriction mentioned on page 167 for the NOR circuit for signal inversion at the $\bar{S}$ input of a bistable circuit):

Previous stage	Type of load	Load
NOR	AND	$g \leq 8$; $\Sigma k \leq 8$
NOR	NOR	$n \leq 4$
AND	NOR	$n \leq k$

Here g is the maximum number of gate loads which can be carried by a NOR stage (gate loads must be added, no matter whether the AND gates are connected in parallel or in series), k is the number of gate loads corresponding to an AND gate, and n the maximum number of NOR loads that can be carried by the previous stage.

Fig. 7 shows a possible realization of a NOR block

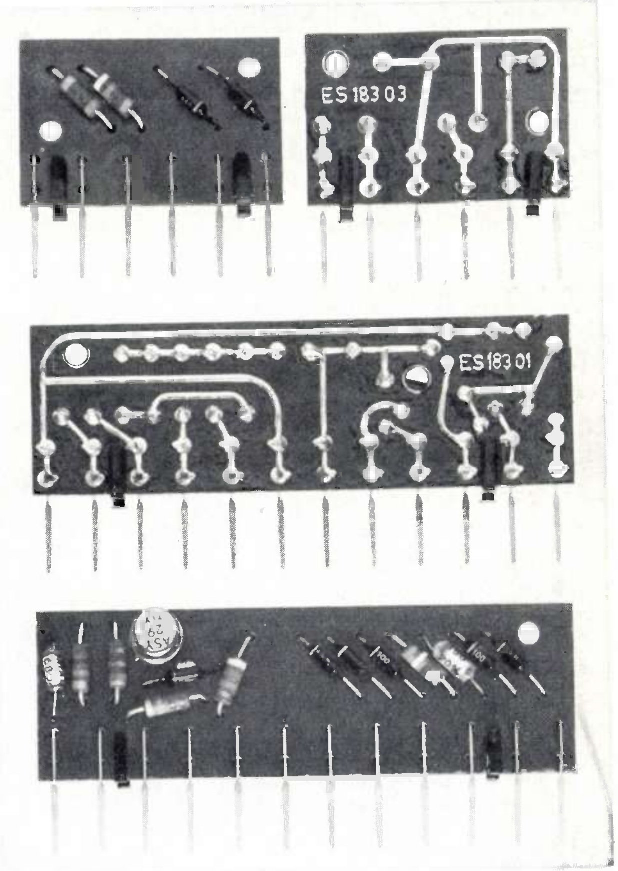


Fig. 7. Examples of the construction of two circuit blocks, as used e.g. in electronic weighing installations. Above AND circuit, centre and below: NOR circuit.

[8] F. Bregman, Counting circuits equipped with transistors, Philips Research Laboratories Eindhoven, unpublished work.
E. Schurig, „UZ 71" — Ein neuer Universalzähler, Elektron. Rdsch. **16**, 111-114, 1962.
R. A. Hempel, A 100 kc add-subtract transistorized decade counter, Semiconductor Prod. **5**, 19-24, 1962.

and of an AND block. In order to increase the economy of this system, these two types of block can be supplemented by another active block consisting of two NOR circuits without pulse gate. These can be used with advantage wherever the NOR does not form part of a bistable or polystable circuit.

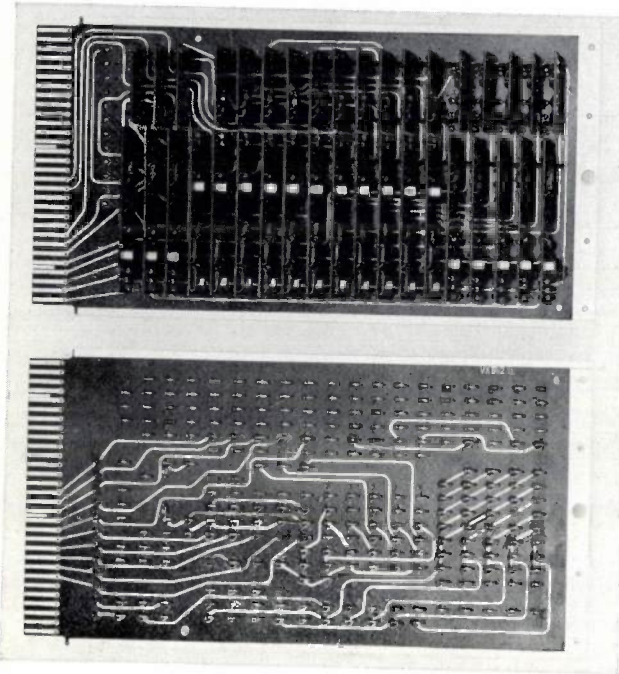


Fig. 8. Construction of a synchronous reversible biquinary decade counter with output amplifiers.

Fig. 8 shows a reversible biquinary decade counter with 10 output amplifiers for the digits 0 . . . 9.

The system of circuit blocks described here has been used in electronic weighing installations with digital data encoding and processing.

The synchronous biquinary decade counter was developed by P. Muuss of the Hamburg Laboratory.

Summary. This paper describes a new system of digital circuit blocks, designed to meet the special needs of industrial measurement and control techniques, characterized by the following.

a) It contains only two different basic circuits: an active logic circuit in diode-transistor logic (DTL), and a passive logic circuit in diode logic.

b) It contains only one type of transistor and two types of diode; there are no emitter-followers.

c) The basic circuits of this system can be combined to give not only bistable but also polystable circuits with for example 3, 4 or 5 stable states.

d) The bistable circuits can be used for either the synchronous or the asynchronous counting mode.

e) The circuit blocks operate reliably under full load in the temperature range from —10 °C to +50 °C with the most unfavourable values of the resistances and voltages within their tolerances, and with the smallest current amplification and the greatest leakage currents which can occur at the end of life of the transistor.

f) The loading table is simple.

g) Special circuits, such as multivibrators, monostable circuits as well as Schmitt triggers, can be realized by simple combination of two active blocks and one or two extra resistors or capacitors.

Generation of musical intervals by a digital method

D. Gossel

534.321.2:621.389

Introduction

The familiar kinds of musical instrument can be divided into two classes:

a) Instruments producing notes whose pitch is not decided upon until the instant of playing: bowed string instruments and certain wind instruments are examples.

b) Instruments possessing a store of notes, from which in the course of playing a selection is made in accordance with a programme. All keyboard instruments belong to this class.

Instruments in class (b) can only be endowed with a limited store of notes for constructional reasons, and because the technique of execution might otherwise be rendered too difficult; also, the access time for whatever notes are available must be compatible with practical requirements for playing the instrument. This implies the existence of some fixed rule or instruction for selecting individual tones from the continuum of pitch.

Several such rules have been laid down at various times in the history of music [1], they find practical expression in the various tonal or tuning systems. The four most important will now be briefly explained and discussed.

Tonal systems

A tonal system has been defined [2] as a scheme for dividing the octave into a progressive sequence of tones, the principle underlying the division being consistently adhered to and designed to produce musically acceptable intervals.

1. Pythagorean tuning

This system dates back to the philosopher who lived during the 6th century BC. It is based upon the fifth,

Dipl.-Ing. D. Gossel is a research worker at the Hamburg laboratory of Philips Zentrallaboratorium GmbH.

and of an AND block. In order to increase the economy of this system, these two types of block can be supplemented by another active block consisting of two NOR circuits without pulse gate. These can be used with advantage wherever the NOR does not form part of a bistable or polystable circuit.

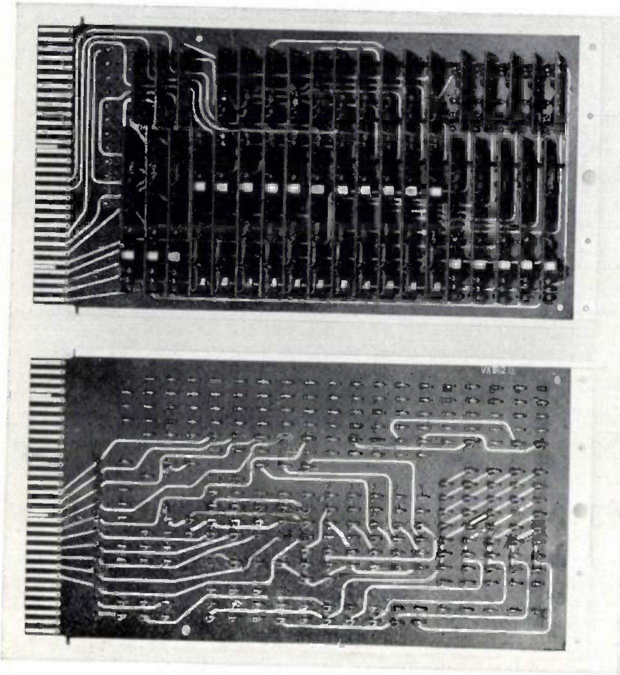


Fig. 8. Construction of a synchronous reversible biquinary decade counter with output amplifiers.

Fig. 8 shows a reversible biquinary decade counter with 10 output amplifiers for the digits 0 . . . 9.

The system of circuit blocks described here has been used in electronic weighing installations with digital data encoding and processing.

The synchronous biquinary decade counter was developed by P. Muuss of the Hamburg Laboratory.

Summary. This paper describes a new system of digital circuit blocks, designed to meet the special needs of industrial measurement and control techniques, characterized by the following.

a) It contains only two different basic circuits: an active logic circuit in diode-transistor logic (DTL), and a passive logic circuit in diode logic.
b) It contains only one type of transistor and two types of diode; there are no emitter-followers.
c) The basic circuits of this system can be combined to give not only bistable but also polystable circuits with for example 3, 4 or 5 stable states.
d) The bistable circuits can be used for either the synchronous or the asynchronous counting mode.
e) The circuit blocks operate reliably under full load in the temperature range from —10 °C to +50 °C with the most unfavourable values of the resistances and voltages within their tolerances, and with the smallest current amplification and the greatest leakage currents which can occur at the end of life of the transistor.
f) The loading table is simple.
g) Special circuits, such as multivibrators, monostable circuits as well as Schmitt triggers, can be realized by simple combination of two active blocks and one or two extra resistors or capacitors.

Generation of musical intervals by a digital method

D. Gossel

534.321.2:621.389

Introduction

The familiar kinds of musical instrument can be divided into two classes:

a) Instruments producing notes whose pitch is not decided upon until the instant of playing: bowed string instruments and certain wind instruments are examples.
b) Instruments possessing a store of notes, from which in the course of playing a selection is made in accordance with a programme. All keyboard instruments belong to this class.

Instruments in class (b) can only be endowed with a limited store of notes for constructional reasons, and because the technique of execution might otherwise be rendered too difficult; also, the access time for whatever notes are available must be compatible with practical requirements for playing the instrument. This implies the existence of some fixed rule or instruction for selecting individual tones from the continuum of pitch.

Several such rules have been laid down at various times in the history of music [1], they find practical expression in the various tonal or tuning systems. The four most important will now be briefly explained and discussed.

Tonal systems

A tonal system has been defined [2] as a scheme for dividing the octave into a progressive sequence of tones, the principle underlying the division being consistently adhered to and designed to produce musically acceptable intervals.

1. Pythagorean tuning

This system dates back to the philosopher who lived during the 6th century BC. It is based upon the fifth,

Dipl.-Ing. D. Gossel is a research worker at the Hamburg laboratory of Philips Zentrallaboratorium GmbH.

which represents a frequency ratio of 3 : 2 and is the simplest musical interval of all with the exception of the octave, with its frequency ratio of 2 : 1. A succession of musically new notes is produced when the gamut is transversed at intervals of a fifth, and in this respect the fifth differs from the octave. The notes of the Pythagorean scale are arrived at by superimposing n fifths in this way and then returning through m octaves to end up in the starting octave; the Pythagorean-type interval $I_{\mathrm{P}\nu}$ is thus the result of multiplying by $(3/2)^n$ and dividing by 2^m. Accordingly, the underlying law is

$$I_{\mathrm{P}\nu} = \frac{(3/2)^n}{2^m} \; ; \quad 1 \leqslant I_{\mathrm{P}\nu} < 2 \, . \quad . \; . \; . \quad (1)$$

In the Pythagorean system all fifths are true 3 : 2 intervals. The thirds, having a frequency ratio of 81 : 64, do not represent a straightforward harmonic interval and are classed as dissonant.

A property shared by Pythagorean tuning with all scales having "just" intonation, and some tempered scales, is that it does not form a closed system. It is impossible in principle to arrive exactly at an octave (say) by superimposing a "pure" interval like a true fifth upon itself, because the relevant frequency ratio represents a simple fraction that cannot yield a whole number when raised to a higher power (which is what the stacking process amounts to) [2]. In fact the octave is not one of the intervals that can be derived from eq. (1), the relation underlying the Pythagorean system.

Nowadays Pythagorean tuning possesses only historical interest; on account of its dissonant thirds, it can only serve as vehicle for a single melodic line.

2. Natural harmonic tuning

This system is the result of introducing a new interval of the simple harmonic type, the major third with its frequency ratio of 5:4, to supplement the fifth (3 : 2) and the fourth (4 : 3); the fourth is that interval which, added to a fifth, is needed to complete the octave. *Fig. 1* shows how all the other harmonic tones are engendered by the octave, fifth, fourth, major third and minor third (6 : 5), this last being the interval which, added to a major third, forms a fifth.

The natural harmonic system is quite a practical propositionforinstruments in class(a) as defined above, but is unsuitable for tuning those in class (b) because too few of the resulting fifths are true to the harmonic series, and this fact limits the possibilities of modulation.

3. Mean-tone tuning

The system, proposed by Arnold Schlick in 1511, was quite commonly employed in the past, especially for organs. The resulting scale contains harmonically true thirds and tempered fifths. The stacking of four perfect fifths engenders the dissonant Pythagorean third (81 : 64), which exceeds the true major third (5 : 4) by an interval known as the syntonic comma (81 : 80). As compared with true perfect fifths, the tempered fifths of the mean-tone system are too flat by a quarter of a syntonic comma, i.e. by $\sqrt[4]{81/80} \approx 1.003$; the difference is too small to be disturbing, and four of these mean-tone fifths engender a true major third. The mean-tone system provides as many as eleven musically acceptable fifths, but the twelfth fifth needed to close the circle is far too wide and scarcely tolerable to the ear (it used to be known as a "wolf").

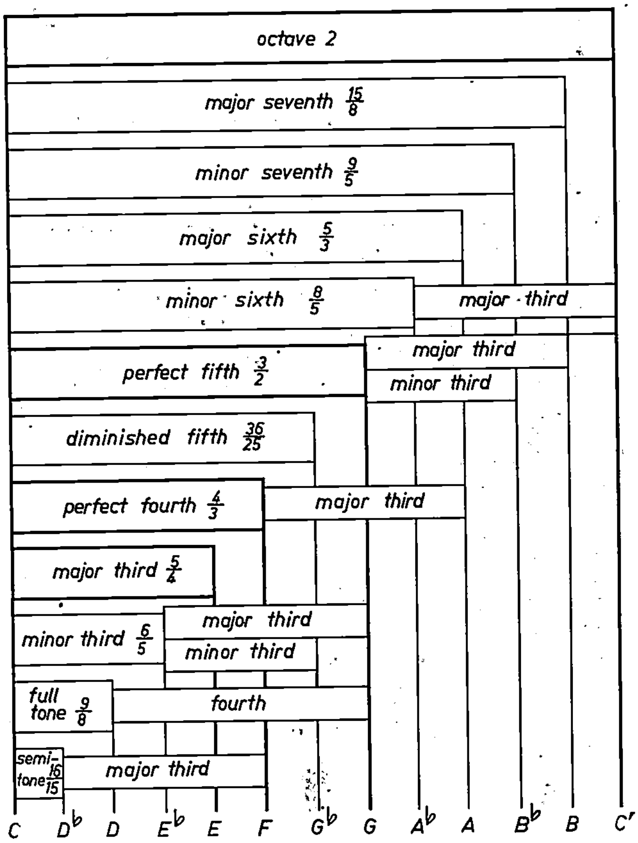


Fig. 1. Natural harmonic tuning: all the intervals of the scale arise out of the octave, fifth, fouth and third.

4. Scale of equal temperament

Practically all kinds of keyboard instrument nowadays are tuned to the scale of equal temperament (or "well tempered" tuning system). Essentially, the system is due to Andreas Werckmeister and Georg Neidhard (c. 1700) although earlier mention of it is to be found in the works of Ramis de Pareja (1440-1500). The system closes back on itself via twelve equally

[1] W. Dupont, Geschichte der musikalischen Temperatur, Bärenreiter-Verlag, Kassel 1935.

[2] Adapted from W. H. Westphal, Physikalisches Wörterbuch, Part II, Springer, Berlin 1952, pp. 546 and 547.

tempered fifths. The tempering consists in the division between these twelve fifths of the ditonic comma (531 441 : 524 288), the interval by which the octave is exceeded when twelve harmonically true fifths are superimposed. This means that the equally tempered fifth is only a twelfth of a ditonic comma (or about 1.001) flatter than a true perfect fifth. To put it another way, the scale of equal temperament divides the octave into twelve exactly equal semitones, or intervals having a magnitude of $^{12}\sqrt{2}$. The advantage of having a closed tuning system is that one can modulate successively from one key to another *ad libitum*, but this is only at the price of major thirds that are a little too sharp and minor thirds that are a little too flat. However, as experience has shown, the ear becomes accustomed to this accentuation of the major and minor character of the relevant modes.

Fig. 2 is a representation of the four tuning systems just discussed. The octave covers an angle of 2π in this diagram. The magnitude of any interval can be expressed in "cents", a logarithmic unit divised by H. Bellerman and H. J. Ellis round about 1880. The multiplication of frequency ratios is thereby reduced to the addition of the corresponding cent values; the formula is

$$\frac{i}{\text{cent}} = {}^{1200}\!\sqrt{2}\,\log I . \qquad (2)$$

Thus the octave, an interval of $I = 2$, has a cent value

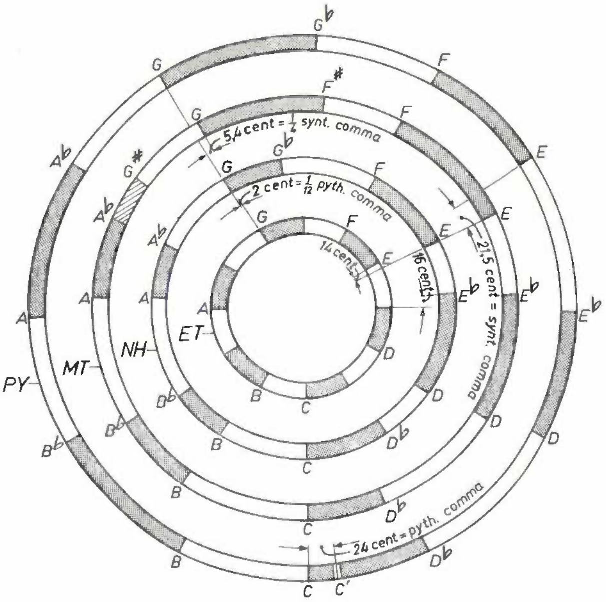


Fig. 2. The four most important tuning systems compared. The octave, having a cent value of 1200, extends over the full circumference of the circle (360°); accordingly, an angle of 1° in the above diagram represents a difference in pitch amounting to 3 1/3 cents.

ET	=	scale of equal temperament
NH	=	natural harmonic tuning
MT	=	mean-tone tuning
PY	=	Pythagorean tuning

of $i = 1200$ and the equally tempered semitone, for which $I = {}^{12}\sqrt{2}$, is equivalent to 100 cents.

On converting Eq. 2 to Naperian logarithms we obtain:

$$\frac{i}{\text{cent}} = 1730 \ln I , \qquad (2a)$$

and the following approximate expression for small intervals such that $I = 1 + \varepsilon$:

$$\frac{i}{\text{cent}} \approx 1730\, \varepsilon . \qquad (2b)$$

It will be seen from fig. 2 that the major third in the scale of equal temperament lies intermediate between the Pythagorean and the natural harmonic third.

Tone production and the tuning of musical instruments

Of particular interest within the framework of the present article are instruments belonging to class (b). With the exception of certain electromechanical organs (the Hammond Organ for example) they all have an independent oscillator for each note, or at least one for each of the twelve notes of the chromatic scale, these oscillators being switched on and off if required [3].

For reasons that are well known the oscillators (which may take the form of strings, wires, air columns, reeds or electronic LF generators) have to be "tuned", i.e. retuned, from time to time, and for instruments employing the system of equal temperament, the operation usually consists in correcting the pitch of a chain of fifths [4]. The tuner's professional skill lies in an ability to find the right tempering for the fifths; only then will the circle close in twelve steps. In principle, tuning could alternatively be done in a sequence of fourths, major sevenths or semitones. In fact the use of the semitone as a tuning interval would do away with the need to work back continually to the starting octave; this must invariably be done if any other interval is employed. But it is almost impossible to judge the tempering of a dissonant interval by ear, and so the consonant fifth and fourth are preferred.

The successful tuning of a musical instrument, in fact, calls for concentration, time and a trained musical ear. Specialists able to do the work are nowadays becoming fewer and fewer.

In the next section a small electronic device is described which permits of exact tuning to the scale of equal temperament in the shortest possible time. Adjustment of a note to the right pitch is done with a visual aid, the needle of a measuring instrument for example, and does not in any way necessitate a trained musical ear. Operation of the device is so simple that it might be worthwhile considering the desirability of simplifying it together with certain keyboard instruments, occasional tuning of the instrument thus being left to the user. The device is also likely to be widely adopted for tuning church organs, a job that has to be done every so often on account of seasonal temperature changes. This routine is particularly laborious and time-consuming in the case of the larger instruments with their multiplicity of pipes and registers.

Generation of intervals by digital means

As is well known it is possible in digital technique to divide a given frequency f_0 by any desired whole number z. Where the value of z is on the large side the usual practice is to feed f_0 into a scaler which has been adjusted to recognize a preselected z value; having counted this number of input pulses, the scaler emits a zeroing pulse and starts to count anew. The zeroing pulses form a train with a recurrence frequency of $f = f_0/z$. It is an easy matter to select divisors z_i such that the corresponding recurrence frequencies f_i represent musically acceptable intervals.

If for example we choose $z_1 = 2$ and $z_2 = 3$, we shall obtain a true perfect fifth:

$$\frac{f_1}{f_2} = \frac{f_0/z_1}{f_0/z_2} = \frac{z_2}{z_1} = \frac{3}{2}. \quad . \; . \; . \; . \quad (3)$$

All the other intervals are obtainable in a similar way. The great advantage of the digital method of generating musical intervals is that the intervals produced are independent of f_0. The absolute position of an interval in the gamut can thus be changed by varying f_0, and this fact can be exploited for the purpose of transposition. Digital equipment and methods could conceivably be used for carrying out investigations into musical aesthetics, enabling concords with various degrees of tempering to be quickly and conveniently produced and compared, and appraised as a function of absolute pitch.

A digital tuner

The intervals between notes in the scale of equal temperament are given by the law underlying the system:

$$I_{G\nu} = {}^{12}\sqrt{2^{\nu}}. \quad . \; . \; . \; . \; . \; . \; . \quad (4)$$

Now, ν is always a positive integer, so the above relation yields irrational numbers which do not correspond exactly to any interval that can be produced by digital means (the only exception is the octave, for which $\nu = 12$). However, W. Schott of this laboratory found that the quotient of 196 : 185 approximates very closely to the equal-tempered semitone, whose value is ${}^{12}\sqrt{2}$, the error being only 5×10^{-6}. It is on this, and on the fact that the absolute pitch of intervals generated by frequency division can be varied at will, that the tuner represented schematically in *fig. 3* is based.

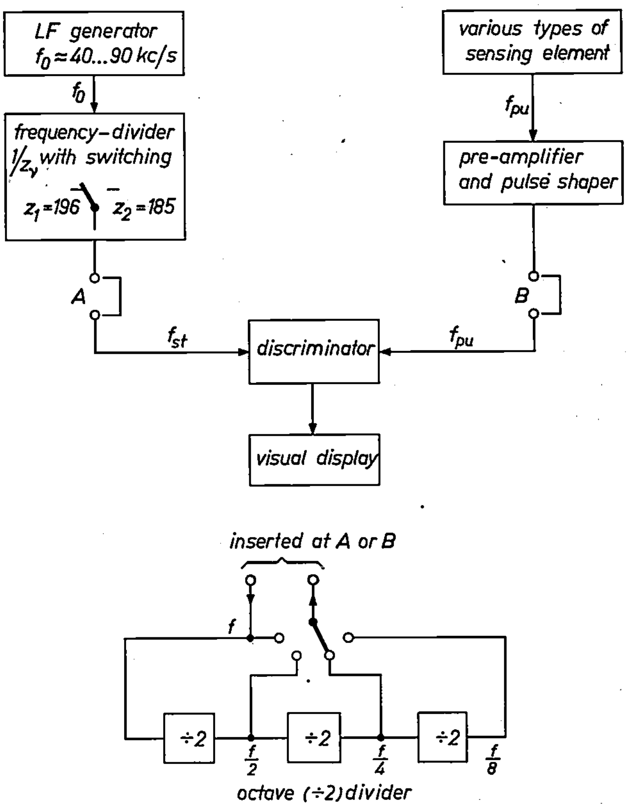


Fig. 3. Arrangement for tuning a musical instrument to the scale of equal temperament. The sensing element can take the form of a microphone, for example, or a magnetic pick up; none is needed for tuning electronic musical instruments. A moving coil measuring instrument, an electric lamp, a magic eye or the like can serve as a visual display device.

The output waveform from an LF generator whose frequency f_0 is continuously variable over the range between 40 kc/s and 90 kc/s, which covers little more than an octave, is fed to a frequency-dividing stage which has facilities for division by $z_1 = 196$ or by $z_2 = 185$, as desired. Thus two frequencies, $f_1 = f_0/196$, and $f_2 = f_0/185$, are alternatively available from the output of frequency-divider A; and provided f_0 is constant, the separation between them is almost exactly equivalent to an equally-tempered semitone. The frequency f_{pu} of the note to be tuned is picked up by a sensing element whose nature depends on the musical instrument being dealt with, and after amplification and conversion into a pulse train it is applied to a discriminating circuit in which it is compared with either f_1 or f_2, one of these serving as a standard or test frequency f_{st}.

The difference-indicating arrangements are phase-sensitive; they should preferably take the form of a visual display. As the picked-up frequency is gradually

[3] D. Wolkov, Electronic organ tone generators, Audio 46, No. 2, 34-44, and No. 3, 30, 32, 65, 1962.

[4] O. Funke, Theorie und Praxis des Klavierstimmens, published by Das Musikinstrument, Frankfurt a.M. 1958.

adjusted to exact equality with the standard frequency a fluctuation in the visual display becomes slower and slower and finally stops altogether; the fluctuation may appear in the movement of a needle against a scale, ceasing when the needle finally comes to rest, or in the dimming and brightening of a small electric lamp, which finally acquires a constant brightness level.

The tuning procedure is as follows.

a) Adjust the LF generator to give its standard frequency of 81 400 c/s, which can if desired be controlled by a built-in quartz crystal. In switch position z_2 = 185 a standard test frequency of $f_{\rm st}$ = 440 c/s will be obtained, and can be used to correct the A above middle C on the musical instrument being tuned. It is scarcely necessary to point out that this standard A can be adjusted to any other desired pitch — 435 c/s for example.

b) Switch now to z_1 = 196, with the result that the standard frequency is lowered by a semitone. This new standard can be used to tune A flat above middle C on the instrument.

c) Switch back to z_2 = 185 and decrease f_0 until $f_{\rm st}$ is in unison with A flat on the instrument, as just corrected.

d) Switch to z_1 = 196 and tune G above middle C.

e) Switch back to z_2 = 185 and decrease f_0 until $f_{\rm st}$ is in unison with G on the instrument, as just corrected.

f) Switch to z_1 = 196 and tune F sharp above middle C, and so on.

In the course of twelve downward semitonal shifts, carried out in the manner described above, all the required intervals can be found and one finishes up an octave below the note first tuned. Taking account of the systematic error involved in each semitonal shift, which is 5×10^{-6}, the octave thus arrived at is true to within about 6×10^{-5}. This may be compared with the smallest deviation from unison that the ear is capable of perceiving in the most sensitive range of its response curve, around 1 kc/s; this smallest detectable difference is 4×10^{-3}, or two orders of magnitude greater than the error in the octave.

Where facilities are required for tuning the parallel octaves in bass or treble along with the twelve notes of the middle register, it is an easy matter to incorporate a chain of octave-dividers at *A* or *B*, i.e. on the standard-frequency or pick-up side of the tuner, the lower registers being covered in the former case and the upper ones in the latter. These octave-dividers are straightforward bistable circuits ("flip-flops") that divide the incoming frequency by two.

One type of frequency comparator circuit is shown in *fig. 4a*; its mode of functioning is explained in figs. 4*b* and *c*.

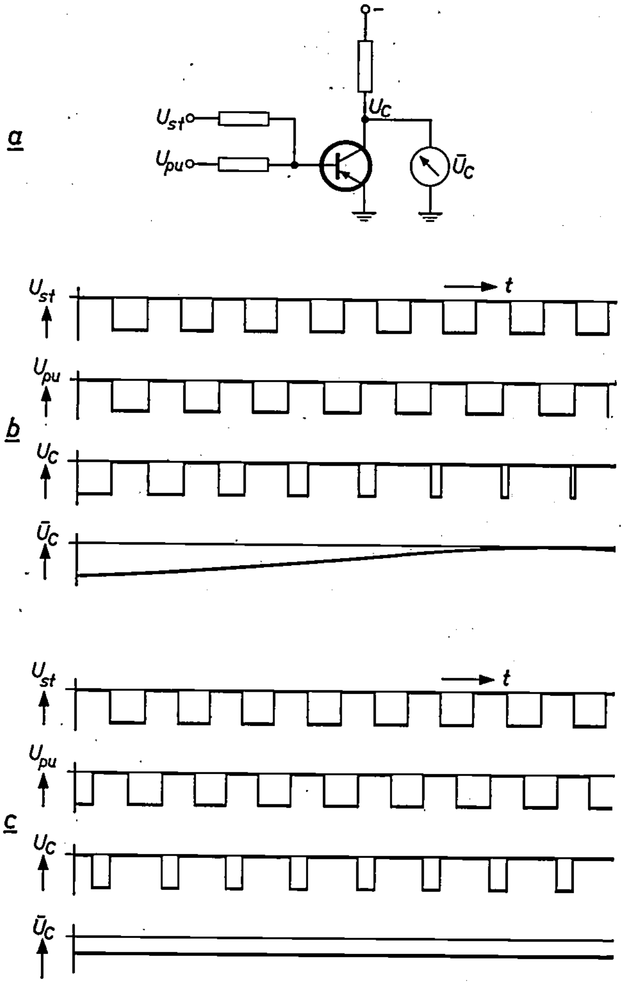


Fig. 4. *a*) Discriminator circuit.
b) Pulse trains operative in the case $f_{\rm st} \neq f_{\rm pu}$.
c) Shapes assumed by the same voltages when $f_{\rm st}$ = $\rm f_{pu}$.

It is assumed that both $U_{\rm st}$ and $U_{\rm pu}$, the standard and picked-up voltages, can each assume either of two values only, namely zero or the negative maximum. Further, the ratio between the pulse duration and the recurrance period is assumed to be constant. The circuit functions as a NOR gate: the presence of either of the two negative voltages on the transistor input suffices to switch it on. The transistor only switches to the non-conducting state when both circuit inputs carry zero potential.

If the standard and picked-up frequencies differ (fig. 4*b*) there will arise at the collector of the transistor a train of pulses $U_{\rm c}$ whose breadth fluctuates cyclically. The corresponding mean voltage $\bar{U}_{\rm c}$ also fluctuates cyclically, but is a continuous function of time; it is this quantity that is visualized. If on the other hand $f_{\rm st}$ is equal to $f_{\rm pu}$, voltage $U_{\rm c}$ will be a train of pulses whose breadth is uniform, though dependent on the purely fortuitous phase relationship between the two incoming signals. The result will be a steady reading on the measuring instrument. The visual display method makes it possible to detect very slow fluctuations (down to a frequency of 0.01 c/s) that cannot be perceived by the ear.

The type of sensing element depends on the kind of musical instrument to be tuned. A microphone is to be preferred for picking up organ tones; for piano tuning a special magnetic pick-up is proposed, fitted with permanent magnet pole pieces which allow it to be attached to the wires on either side of the one being tuned, in such a way that it has a damping effect on vibrations in these neighbouring wires. The pick-up also embodies a small transistor amplifier coupled to a feedback circuit, which serves to keep the wire being tuned in a state of sustained vibration. For correct frequency discrimination it is necessary that the resonator under test should supply a continuous train of oscillations. Experiments have shown that damped vibrations, such as are produced by percussion of a piano wire, are useless for this purpose.

An electronic musical instrument like the "Philicorda" is the least demanding as regards the ancillary equipment required for tuning. No electro-acoustic transducer is required, because alternating voltages at the frequencies under test are available from the instrument anyway. In experiments on a "Philicorda" which had first been completely detuned in a random manner, it was possible with the aid of the digital tuner to bring the instrument back to exact conformity with the equal-temperament scale in a matter of barely ten minutes.

Musical instruments based on digital interval generation

Musical instruments of the purely melodic type (*fig. 5*) or of the type equipped for harmony (*fig. 6*) can be devised on the basis of this digital method of generating musical intervals; the "polyphonic" models are naturally dearer, the price depending on the number of melodic lines required.

Both types have much the same fundamental design as the digital tuner. One essential difference is that the purely melodic type incorporates one frequency divider with facilities for selecting 12 different divisors $z_1 \ldots z_{12}$, whereas the harmonizing type is equipped with n frequency dividers up to a maximum of 12, for each of which at least one divisor z_ν, and possibly a set of m such divisors, is available.

A feature common to all such instruments is that it is impossible in principle for them to go out of tune, since all the notes are produced by numerical division of a single master frequency f_0; and all can be transposed at will, simply by altering f_0.

The all-important factor governing the choice of divisors is the type of tonal system desired and, in some cases, the exactness or truth the individual intervals are required to have. To finish up, we shall briefly describe a possible choice for the natural harmonic tuning.

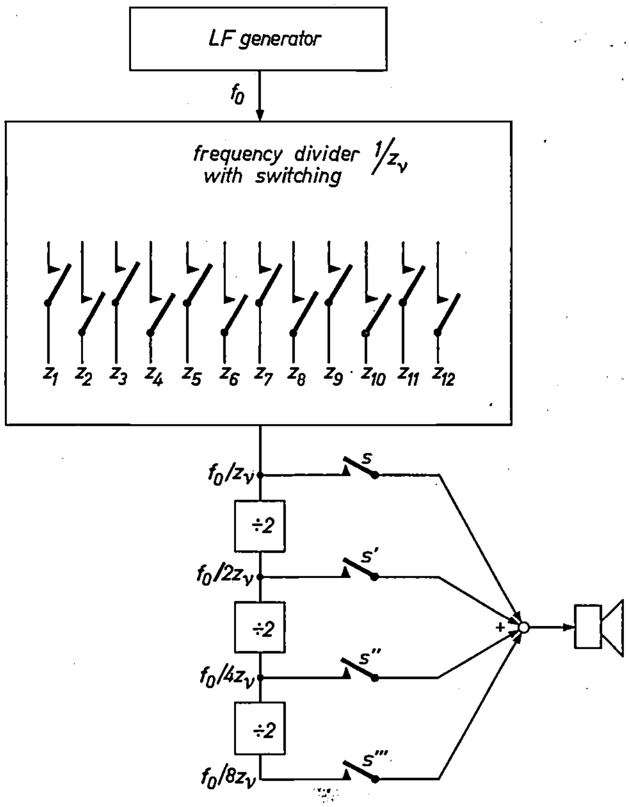

LF generator
f_0
frequency divider $1/z_v$ with switching
z_1 z_2 z_3 z_4 z_5 z_6 z_7 z_8 z_9 z_{10} z_{11} z_{12}
f_0/z_v
s
÷2
$f_0/2z_v$
s'
÷2
$f_0/4z_v$
s''
÷2
$f_0/8z_v$
s'''
+

Fig. 5. Purely melodic musical instrument using digital interval generation.

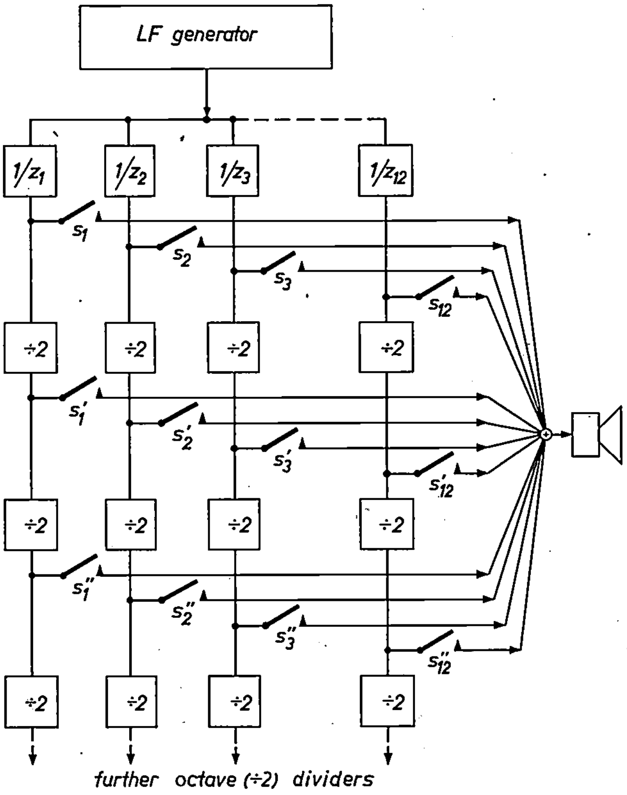


Fig. 6. "Polyphonic" musical instrument using digital interval generation.

. The lowest note to be produced is associated with the highest divisor z_1:

$$f_1 = \frac{f_0}{z_1} . \qquad \qquad (5)$$

All the other tones are given by

$$f_\nu = \frac{f_0}{z_\nu} . \qquad \qquad (6)$$

where $z_\nu < z_1$. This means that the divisors must be

$$z_\nu = z_1 \frac{f_1}{f_\nu} = \frac{z_1 \, b_\nu}{a_\nu} , \qquad \qquad (7)$$

where a_ν represents the numerator and b_ν the denominator of any of the fractions appearing in fig. 1; these fractions stand for cancelled-out frequency ratios. Since all z values must be integers and no b_ν value is a factor of its a_ν, a sensible move will be to make z_1 the lowest common multiple of all the values assumed by a_ν:

$$z_1 = 2 \times 2 \times 2 \times 2 \times 3 \times 3 \times 5 = 720$$

All the other smallest whole-number values of z_ν can be found on inserting $z_1 = 2 \times 2 \times 2 \times 2 \times 3 \times 3 \times 5 = 720$ in eq. (7) (they are set out in *Table I*). The choice of 720 as highest divisor entails a certain outlay of circuit elements — 10 bistables, say, and the associated decoding circuits. This outlay can be greatly reduced if the designer confines himself to producing only the consonant intervals at their true values, accepting slight deviations in the dissonant ones.

Table I. Divisors required for producing the intervals of a natural harmonic scale by the digital method, shown against the notes of the scale.

Note	f_ν/f_1	z_ν
C	1	720
D flat	16/15	675
D	9/8	640
E flat	6/5	600
E	5/4	576
F	4/3	540
G flat	36/25	500
G	3/2	480
A flat	8/5	450
A	5/3	432
B flat	9/5	400
B	15/8	384
C′	2	360

For example, let us suppose that only the octave, fifth fourth and the major and minor thirds need be taken into account in fixing z_1. The highest divisor then becomes

$$z_1 = 2 \times 2 \times 3 \times 5 = 60.$$

The new set of divisors and the errors they involve are displayed in *Table II*; a positive sign means that the tone in question is too sharp. The quantity of components required is now greatly reduced. Six bistables will be ample, and even so, seven of the intervals produced are true as against five which are not. Some of the deviations are rather large; for example, that in the minor sixth amounts to almost the fourth part of a semitone.

Table II. Divisors required for, and errors involved by, an approximation to natural harmonic tuning.

Note	z_ν	Error	
		percentage	in cents
C	60	—	—
D flat	56	+ 4.4	+ 7.7
D	53	+ 6.3	+10.8
E flat	50	—	—
E	48	—	—
F	45	—	—
G flat	42	— 8.0	—13.8
G	40	—	—
A flat	38	—13.3	—23.1
A	36	—	—
B flat	33	+10.0	+17.3
B	32	—	—
C′	30	—	—

However, a better approximation can be obtained by choosing z_1 three times as large. The result of so doing is that the minor seventh and diminished fifth all become true intervals. With a highest divisor of

$$z_1 = 2 \times 2 \times 3 \times 3 \times 5 = 180$$

the number of bistables required is eight, and the tuning that results is as shown in *Table III*. The scale arrived at in this way would be quite acceptable musically.

Table III. Divisors and errors entailed by a closer approximation than that in Table II.

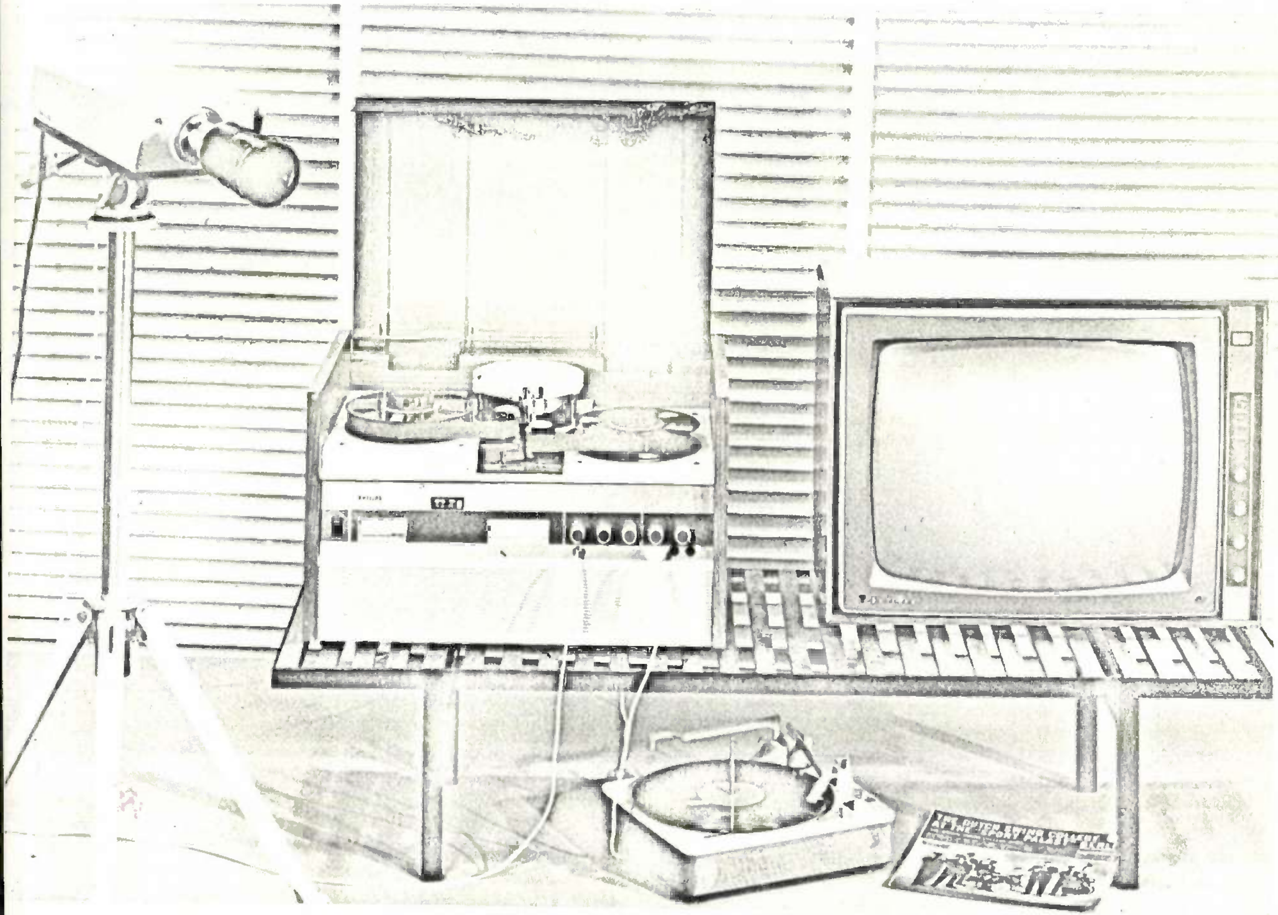
Note	z_ν	Error	
		Percentage	in cents
C	180	—	—
D flat	169	—1.5	—2.6
D	160	—	—
E flat	150	—	—
E	144	—	—
F	135	—	—
G flat	125	—	—
G	120	—	—
A flat	112	+4.4	+7.7
A	108	—	—
B flat	100	—	—
B	96	—	—
C′	90	—	—

Summary. A method familiar from digital techniques, enabling a given frequency to be divided by any desired whole number, can be exploited for producing musically acceptable intervals. The more important tuning systems are briefly reviewed. A description follows of a device suitable for quick and accurate tuning of keyboard instruments; it produces an equally-tempered semitone, which can be transposed through the gamut at will. The true semitonal interval, whose value is $\sqrt[12]{2}$, can be closely approximated by performing the division 196 : 185, the error being only 5×10^{-6}. Also described is a procedure for arriving at a complete scale of notes for a melodic instrument, or one affording facilities for harmony. By way of example, numerical values are given for the intervals of natural harmonic scales obtainable by the digital method.

Volume 26, 1965, No. 4/5/6 pages 89-176 Published 21st December 1965

PHILIPS TECHNICAL REVIEW

VOLUME 26, 1965, No. 7



A video tape recorder for non-professional use

H. K. A. de Lange

681.84.081:621.397

While the recording of sound signals on magnetic tape has become common practice for many non-professional purposes, the magnetic recording of television signals was until recently confined to places where the expensive and complicated equipment involved can be operated by technically trained personnel. Recently, however, video recorders have been constructed which can be operated by persons with no technical training.

The possibility of recording television signals on a agnetic tape has existed for a number of years. Until ecently, however, its application was limited to television studios. For non-professional use the equipment hould be more compact and less expensive, and must be capable of being operated by persons who have had no technical training. This article describes a video tape recorder of this kind, developed in our laboratories.

It is likely that these recorders will come into use in the near future both in the entertainment sector and elsewhere. The video recorder will be useful, for example, in schools, where it can allow a particular

. H. K. A. de Lange is on the staff of Wiener Radiowerke GmbH, ienna.

television programme to be played back as often as required, so that parallel classes can see the programme at different times. Because of the fact that the pictures can be shown immediately after recording, a video recorder can render other valuable services in education, especially in vocational schools. Related fields of application are motion studies in industry, coaching in games and athletics, and theatre and ballet.

Video recordings can also be used as evidence of traffic offences as well as for studying traffic situations. More generally, the use of this system of recording pictures is indicated in those cases where there is more interest in the recording of the pictures than in storing all of them for archive purposes. Finally, mention should be made of the application of video recording in X-ray diagnostics. This application, for stationary pictures, has already been described in this journal in an article on the recording of television signals on a magnetic wheel store [1].

Requirements to be met by the recorder

Before describing the apparatus itself, we shall mention the principal requirements to be met by a non-professional video recorder.

1) Since it has to be operated by non-technical people, the apparatus should require no critical adjustments, and there should be no risk of damage by errors in operation.
2) It should be possible to connect the recorder to a normal television set without the latter having to be radically modified. This applies both to recording and playing back a programme. The recorder must also be able to work in a closed-circuit television system, i.e. in combination with a television camera, a microphone and a video monitor.
3) The apparatus must be easily transportable.
4) Tape consumption should be such that the tapes are not too expensive or too bulky.
5) A reasonable playing time without interruption should be possible.
6) No particularly high demands should be made on the constancy of the voltage or frequency of the supply mains.
7) The picture quality need not be equivalent to that obtained with professional recorders, but should still be acceptable.
8) The price must be substantially lower than that of professional video recorders.

This summary discloses the paradoxical fact that in some respects the demands on a video recorder of this kind are greater than those imposed on the more expensive studio equipment. We refer in particular to the safeguards needed against errors in operation, and to the possibility of connecting the recorder directly to a normal television set. This creates problems in connection with supplying and extracting the signal; we shall return to this later. The less stringent demands on picture quality allow a simplification compared with studio video recorders: the maximum frequency of the video signals to be recorded can be lower. Studio video recorders are required to record frequencies up to say 5 Mc/s, whereas the highest frequency for the recorder described here is 2.5 Mc/s [2]. This made it possible to reduce the tape speed and also the overall dimensions of the recorder.

Principle of the recorder

In the simplest magnetic recording system the recording and playback head is stationary and the tape is moved past the head. For the recording of video signals, however, this system presents considerable difficulties. If we take the highest video frequency as 2.5 Mc/s, and the shortest wavelength that can be recorded on the tape as 2 μm, which is about the best that can at present be achieved, then the tape speed needed is 5 metres per second. If the playing time is not to be too short, this would mean the use of very large tape reels. Furthermore it is difficult at such a high speed to maintain reliable contact between tape and recording head. An added difficulty is that such large and heavy reels call for relatively large motors and a complicated braking system.

For these reasons a system was adopted in which the scanning speed is made much greater than the tape speed by mounting the head on a rotating disc. The system is basically the same as that of an experimental apparatus for studio use, a brief description of which was published some time ago in this journal [3]. In the non-professional recorder to be described here, however, the limited frequency band has allowed reductions in the size of the apparatus and the tape speed.

The principle of the system is illustrated in *fig. 1.* The magnetic tape B, 25 mm wide, is drawn around very nearly the complete circumference (355°) of the stationary drum T in a helical track. The tape speed is only 19 cm per second. With this approach, using reels with a diameter of no more than 20 cm, which can take a tape with a length of 540 metres, a playing time of 45 minutes can be achieved. The pitch of the helix is slightly less than the tape width. The drum, the diameter of which is 15 cm, contains a disc S which rotates at a speed of 50 revolutions per second. Mounted on this disc is the recording head (*video head*) K_V, which rotates with the disc, so that it runs in a slot G round the complete circumference of the drum T. In one revolution the video head thus records a track about 47 cm long obliquely across the tape, and owing to the slight displacement of the tape during each revolution, each successive track lies beside the preceding one. The entire tape is filled in this way with recorded tracks except for a narrow space at each edge. *Fig. 2* shows the relative positions of a few tracks on a part of the tape. The distance between centres of two successive tracks is 180 µm.

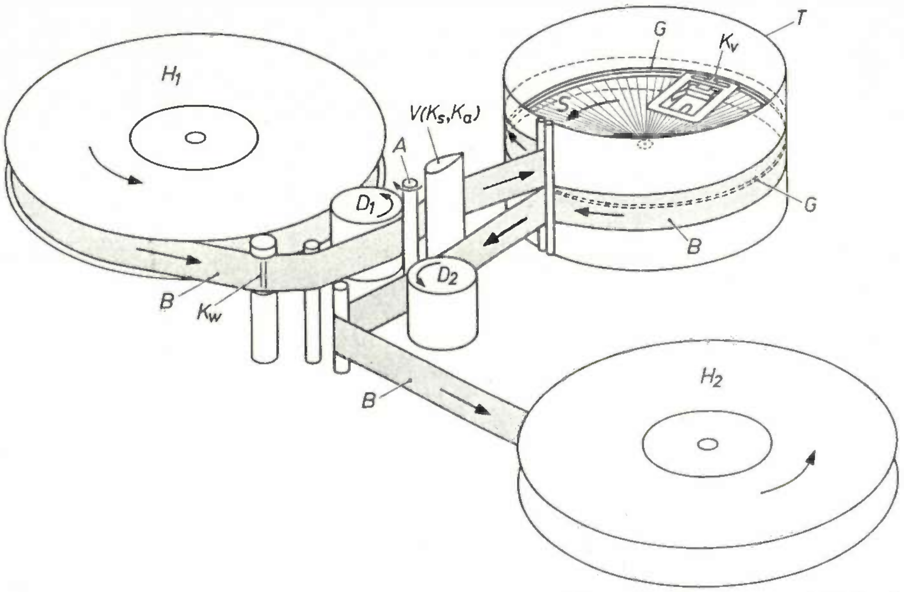


Fig. 1. Principle of the video recording system. *T* drum with slot *G*. The video head K_v fixed to the rotating disc *S*, runs in the slot. *V* guide-vane housing the synchronizing head K_s and the audio head K_a. K_w erasing head. *B* magnetic tape. H_1 and H_2 tape reels. *A* tape drive capstan. D_1 and D_2 pressure rollers.

Since the video head rotates at a speed of 50 r.p.s. it follows that each track is recorded in the time needed for one picture frame (1/50 s), so that a complete picture is recorded on each track (although with half the number of lines). During recording, the rotation of the disc is synchronized with the picture frames, the phase being set so that the recording head crosses from the incoming to the outgoing tape just before the flyback. Each frame synchronizing pulse is thus completely recorded and there is no interruption between the recording of this pulse and the next frame. As the interruption necessary for passing from one track to the next occurs just before the flyback, a few lines in the picture are lost, but this is hardly if at all noticeable on playback as the last lines in every frame are usually covered by the mask round the picture tube.

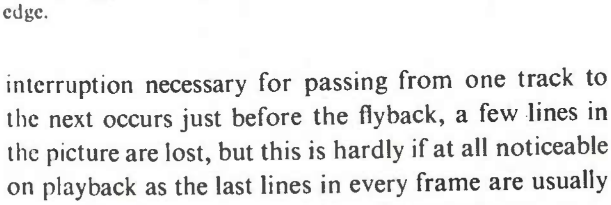


Fig. 2. The video signals are recorded in oblique tracks on the 25 mm wide tape. Each track is 180 µm in width and roughly 47 cm in length. Frame synchronizing pulses are impressed along the top edge of the tape and the audio signals along the bottom edge.

For the *playback* of the recorded pictures the same head is used. Obviously the head must run exactly over the tracks and not, for example, on the boundary between two tracks. It is therefore necessary to ensure that the rotating disc remains in the correct phase in relation to the tracks. To ensure this, the normal frame synchronizing pulses are impressed during recording along one of the edges of the tape, at intervals of 190/50 = 3.8 mm. This is done by means of a stationary head (*synchronizing head*) K_s which is mounted with the audio head K_a, to be discussed presently, in the guide-vane *V*. During playback the head K_s supplies pulses which control the tape speed and the phase position of the rotating head K_v in relation to the tracks. The recorded frame pulses thus have the same function as the perforation on films, and could be referred to here as a magnetic perforation.

The sound signals are recorded on the other edge of the tape. This is done by the *audio head* K_a which is accommodated with K_s in the guide-vane *V*. During recording, the magnetic tape unwinding from the reel H_1 is first passed over the *erasing head* K_w, which is wide enough to erase the picture, sound and synchronizing tracks of any recording already present.

[1] J. H. Wessels, A magnetic wheel store for recording television signals, Philips tech. Rev. 22, 1-10, 1960/61.

[2] In this connection it should be noted that in many normal television transmissions there are no video signals with frequencies higher than 2.5 Mc/s; the quality of the picture reproduced by the video recorder is no poorer than that of a directly received television picture from such a transmission.

[3] F. Th. Backers and J. H. Wessels, An experimental apparatus for recording television signals on magnetic tape, Philips tech. Rev. 24, 81-83, 1962/63.

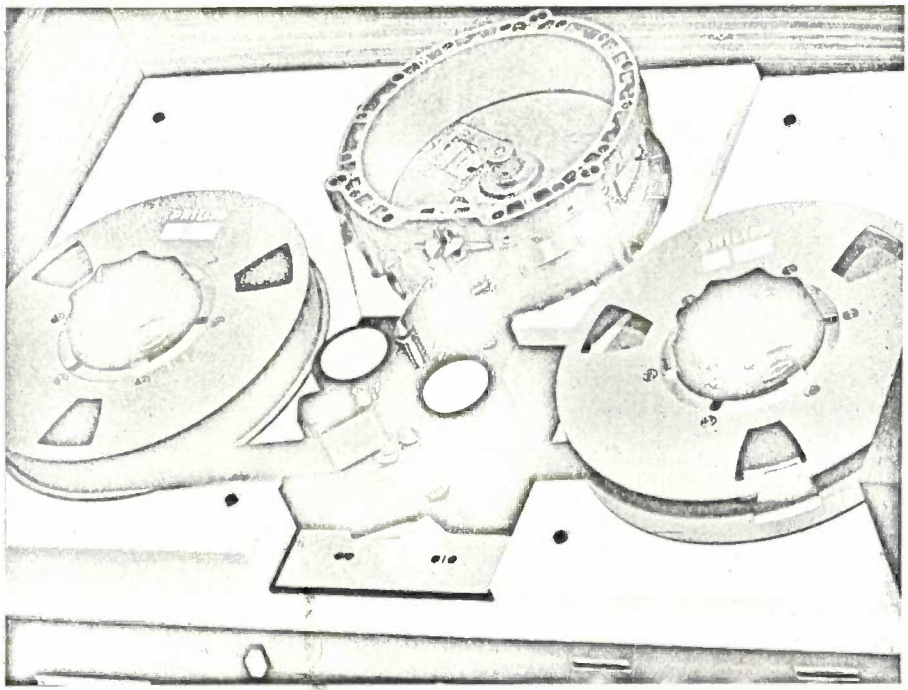


Fig. 3. The tape deck. The cover on the drum T (see fig. 1) has been removed to show the video head.

Fig. 3 shows a photograph of the tape deck, where the various parts discussed can be seen. The top cover of drum T has been removed in order to show how the video head K_v is mounted on the disc S. *Fig. 4* shows a photograph of the video head. The magnetic material used is ferrite, and the ferrite is partly enclosed in glass to avoid disintegration, which is otherwise particularly liable to occur at the head gap.

Some constructional details

The disc S is driven by a pair of asynchronous motors coupled to the spindle of the disc by a pulley drive. The motors exert a greater torque on the disc than is needed to obtain the right speed of rotation. The required speed is maintained, however, by an eddy-current brake, the energizing current for which is supplied by a control circuit. We shall return to this circuit when dealing with the electronic circuits.

The reason for control by braking and not by means of the motors is that the power to be supplied by the control circuit can then be lower. If the motors were controlled, the control circuit would have to supply the energy needed for starting and sustaining the movement of the disc and its video head. Moreover, the circuit technique is much simpler for eddy-current braking than for motor control.

The rotating video head is coupled to the signal-handling part of the circuit by means of a transformer consisting of two halves, one having a stationary winding and the other rotating with the disc. This coupling system dispenses with the need for slip rings and brushes, and rules out the risk of unreliable contact.

The *tape drive system* consists of the spindle A and two rollers, D_1 and D_2, pressing both the tape arriving at the drum, and the tape leaving the drum, against the capstan. With this drive system there can be no irregularities in the tape speed at the drum, caused for example by the reels. Moreover the tension of the tape is much less than it would be if the take-up reel had to move the tape around the drum.

Winding tape under high tension on to a reel can, after a certain time, cause plastic deformation of the tape and this could give rise to picture distortion. Since the tape speed has to be highly constant, the spindle A must be driven by a motor which can easily be regulated. It is moreover necessary, with a view to the fast forward winding or rewinding of the tape, to drive the spindle very fast. A d.c. motor is used for the best solution of these two requirements. To avoid wear and interference that might be caused by brushes an "optical collector" is used. With this device the direction of current flow in the stator coils is reversed at the appropriate moment by a bistable circuit (flip-flop), driven by

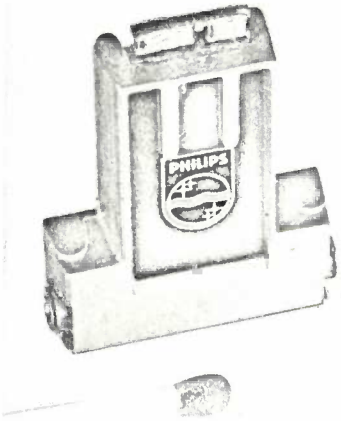


Fig. 4. The video head K_v (see fig. 1) with holder.

a phototransistor. The arrangement is shown schematically in *fig. 5*. Fixed to the spindle *A* of the motor *M* is a disc S_b, which is situated between the lamp L_m and the phototransistor *FT*. This disc has four slits through which the phototransistor is periodically illuminated as the disc rotates. As a result the bistable circuit is intermittently switched from one state to the other, reversing the direction of current in the stator windings.

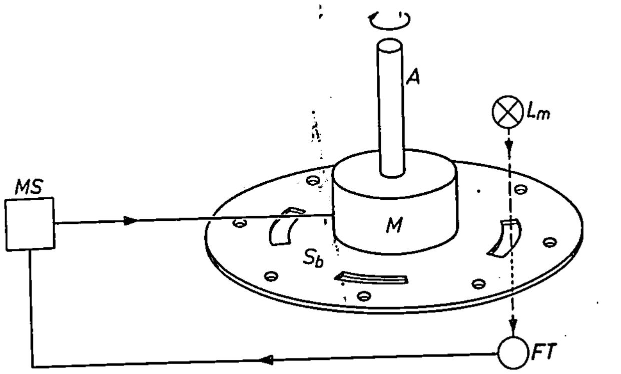


Fig. 5. Schematic illustration of a motor with "optical collector". *M* motor. *A* spindle (see fig. 1). S_b rotating disc with slots. L_m lamp. *FT* phototransistor. *MS* motor control circuit.

Recording the video signal

As already mentioned, the frequency range of the video signals that can be recorded is from 0 to 2.5 Mc/s. This is such a wide range that it is difficult to record the signal directly; the lower frequencies especially cause difficulties. This was discussed in detail in the article under reference [1], which also describes a method of avoiding these difficulties, i.e. by recording not the video signal itself but a carrier modulated by the video signal. In this way one can ensure that in the spectrum of the modulated signal the ratio of the highest to lowest frequencies is smaller than in the video signal. *Frequency modulation* is employed to minimize the effects of possibly disturbing variations of the signal amplitude (due for example to bad contact between tape and head). The modulator is a multivibrator whose frequency is controlled by the video signal. The multivibrator frequency can be varied from 3 to 4.3 Mc/s, and the circuit is designed so that the lower of these two frequencies corresponds to the peaks of the synchronizing pulses. The higher of the two frequencies corresponds to the white level.

In *fig. 6*, where the method of recording the video signal is represented in a block diagram, the modulator is denoted by Mod_1. The video signal (*Vid*), which may be supplied by a camera or, as we shall describe later, by a television receiver, goes via the amplifier A_1 to the modulator while the FM signal reaches the video head K_v via the amplifier A_2 and the rotary transformer *Tf*.

Since the frequency of a multivibrator is in general not very constant and may vary for a number of reasons (e.g. temperature and mains voltage fluctuations, and ageing of the valves) a stabilizing circuit is employed which keeps the lower frequency limit at exactly 3 Mc/s. For this, the output signal from A_2 is fed not only to the transformer *Tf* but also to the discriminator *Di*. This is essentially a tuned circuit, whose impedance-frequency characteristic has a steep slope at 3 Mc/s. The signal from the discriminator is supplied to the modulator after rectification and amplification in the d.c. amplifier A_3.

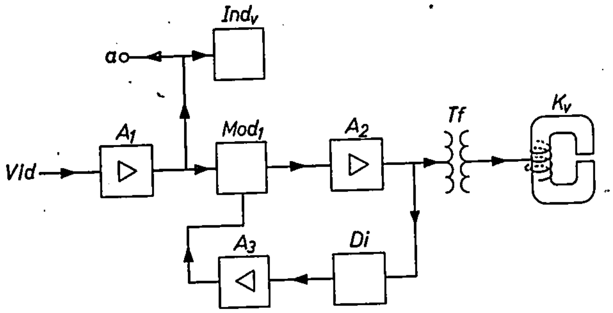


Fig. 6. Block diagram of the circuit for recording the video signals *Vid*. A_1 and A_2 a.c. amplifiers. Mod_1 modulator (multivibrator). *Di* discriminator. A_3 d.c. amplifier. K_v video head. *Tf* transformer, one half of which rotates with the video head. Ind_v visual indicator. *a* terminal for the synchronizing circuits (see fig. 12).

The amplifier A_1 also supplies a signal for a visual indicator Ind_v, necessary when adjusting the video signal by means of the gain control in A_1. A further signal is taken from A_1, via the terminal *a*, for the synchronizing circuits discussed below.

For playback the recorded FM signal has to be converted back into a video signal. A block diagram of the method used is shown in *fig. 7*. The signal delivered by the video head K_v reaches the amplifier A_4 again via the rotary transformer *Tf*. The amplifier is followed by a limiter *B*. The output signal from this limiter is passed to the frequency detector FD_v.

The limiter has to be really effective, because at the

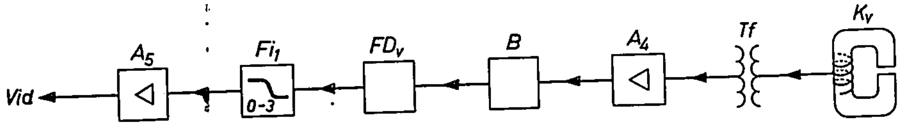


Fig. 7. Block diagram of the circuit for playback of the video signals. K_v video head. *Tf* rotary transformer. A_4 and A_5 amplifiers. *B* limiter. FD_v frequency detector. Fi_1 low-pass filter. The video signal *Vid* obtained from A_5 can be supplied to a monitor or, in the manner illustrated in fig. 11, can be used for playback with a television receiver.

high speed of the video head the contact between head and magnetic tape is relatively poor and consequently fairly considerable variations can occur in the strength of the FM signal. The limiter used here is a *multivibrator*, having a constant output voltage and over a wide range a frequency identical to that of the FM signal. The FM signal is applied to the grids of both valves.

Another reason for using an effective limiter is that, because of the small value of the modulation index of the FM signal, the components of this signal lie in a wide frequency range. A frequency demodulator is therefore needed which has a very wide and hence relatively flat characteristic. This means that undesired amplitude variations of the FM signal would have a relatively large effect on the output signal.

The essentials of the frequency detector FD_v are a diode and a tuned circuit whose resonant frequency is lower than the frequency of the FM signal to be detected. This signal is thus converted in the conventional way into an amplitude-modulated signal, high instantaneous frequency of the signal corresponding to a small amplitude.

Here it is useful to note an advantage of the "active" limiter, such as a multivibrator, compared with most other ("passive") limiters. This advantage appears when the video signal is briefly interrupted, as a result for example of a small irregularity in the magnetic tape. When this happens a passive limiter gives no output, and this shows itself at the output of the discriminator described above as a signal with a very high instantaneous frequency. This causes white streaks in the picture. With the triggered multivibrator employed here this effect is much less troublesome. If the trigger signals are interrupted the multivibrator continues to oscillate. If this "self-oscillation frequency" is arranged to lie in the middle of the frequency band covered by the FM signal, this signal then corresponds to grey, causing less annoyance.

In order to remove any remaining FM signal, the output signal from the frequency detector is passed through the low-pass filter Fi_1, whose cut-off frequency is 3 Mc/s. Via the amplifier A_5 the video signal can now be fed to a monitor or to a normal television receiver in the way described below.

Recording the audio signal

The audio signal is recorded at one of the edges of the tape in the same way as in normal tape recorders. A block diagram of the circuit used is shown in *fig. 8*. The audio signal *Aud*, which may come from a microphone, a gramophone pick-up or from a television receiver, goes via amplifier A_6 to the audio head K_a and to the visual indicator Ind_a as well. This is used when adjusting the signal level by means of a gain control in A_6.

Following established technique [4], a biasing signal of much higher frequency than that of audible sound is superimposed, in order to reduce distortion in the magnetic recording of the audio signals. This biasing signal is supplied by an oscillator Osc_a, the signal from which is also used for the erasing head K_w.

The *playback* of the audio signal is also conventional. The signal supplied by the audio head K_a is amplified and can be fed to a final amplifier or, after further processing, to a television set.

Connecting the video recorder to a television receiver

The requirement of being able to connect the video recorder to a normal television receiver without radical modifications to the receiver gave rise to a number of special problems. One of these relates to the circuit employed for synchronizing the rotating video head with the frame synchronizing pulses. We shall return to this circuit later. Another problem arises from the fact that in Europe hardly any television sets are fitted with a mains transformer, which means that the TV receiver chassis may carry mains voltage. Taking video and audio signals from the video and audio amplifiers of a television receiver in a conventional way would often result in excessive mains hum. Moreover this would generally amount to an infringement of safety regulations.

For these reasons the signals in our apparatus are taken from the *intermediate frequency amplifier*. Due to use of the intercarrier sound television system [5] both the video and the sound information are available here. To avoid modification of existing television receivers, the signal is taken off by an electrode capacitively coupled to the anode of the last I.F. valve. This electrode is in the form of a ring encircling the valve: it is mounted, properly insulated, in a screening can.

Fig. 9 shows a sketch of this can, with the transformer coupling between the ring (*El*) and the coaxial cable connected to the video recorder. The video recorder contains a circuit corresponding to the last I.F. stage

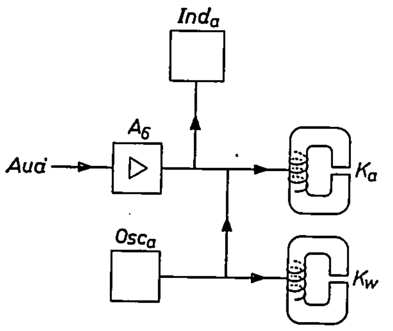


Fig. 8. Block diagram of the circuit for the recording of audio signals *Aud*. A_6 amplifier. K_a audio head. K_w erasing head. Osc_a oscillator. Ind_a visual indicator.

[4] See for example W. K. Westmijze, The principle of the magnetic recording and reproduction of sound, Philips tech. Rev. **15**, 84-96, 1953/54.

[5] See for example W. Werner, The different television standards considered from the point of view of receiver design, Philips tech. Rev. **16**, 195-200, 1954/55, in particular pages 198-199.

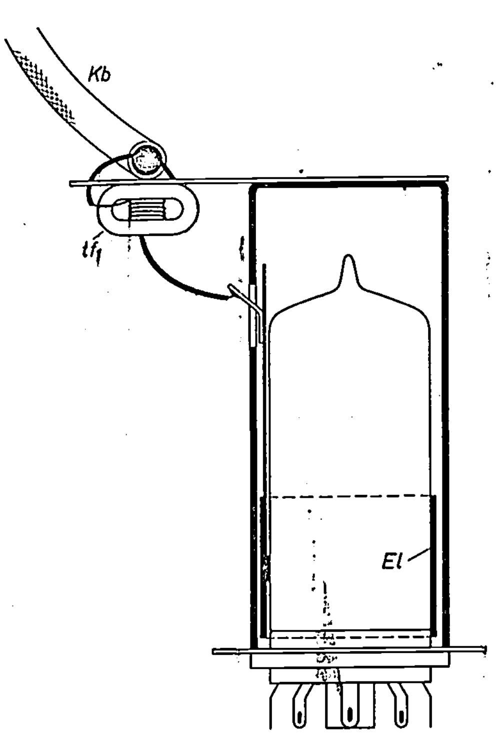


Fig. 9. The video signal to be recorded can be obtained from a normal television receiver without radical modifications to the receiver. This is done by means of a ring electrode (*El*) which is capacitively coupled to the anode of the last I.F. valve. The ring electrode, properly insulated, is mounted in a can fitted around the valve. *Kb* coaxial cable. *tf*₁ transformer.

and the video detector of a television receiver working on the intercarrier sound system. The relevant block diagram can be seen in *fig. 10*. The coaxial cable *Kb* is connected via a transformer tf_2 to the I.F. amplifier A_7, which is followed by the video detector *VidD*. The video signal and the intercarrier sound FM signal are obtained from the output of this detector. After amplification (A_8), the FM sound signal is demodulated in the frequency detector FD_a where the audio signal (*Aud*) is produced. Video and audio signals can then be connected to the points indicated in fig. 6 and fig. 8.

Fig. 11 shows a block diagram of the circuit enabling *playback* of the recorded signals through a television receiver. Here video and audio signals are modulated on to carrier waves with frequencies corresponding to the video and sound carriers of one of the television channels 2, 3 and 4, the frequency range of which extends from 48.25 to 67.75 Mc/s.

The *vision carrier* is generated in the oscillator Osc_v, and is modulated by the video signal in the modulator Mod_2. It would be possible to generate the sound carrier by means of another oscillator, this oscillator then being frequency-modulated by the audio signal. This method, however, involves a complication due to the fact that there must be an exact frequency difference of 5.5 Mc/s between the two carriers.

In order to maintain this difference with sufficient accuracy it would be necessary to generate the two carriers with the aid of crystal oscillators, in spite of the fact that there is no call for great absolute accuracy of the frequencies. To avoid the fairly expensive solution which this would represent, an auxiliary carrier with a centre frequency of 5.5 Mc/s is frequency-modulated by the audio signal. Due to the much lower frequency sufficient absolute accuracy can be obtained by simple means (see below). Next, the vision carrier is amplitude-modulated by both video signal and FM auxiliary carrier. This produces, as one of the sidebands, a signal component whose frequency is 5.5 Mc/s higher than that of the vision carrier and which contains the audio signal as frequency modulation. This is the modulated sound carrier required.

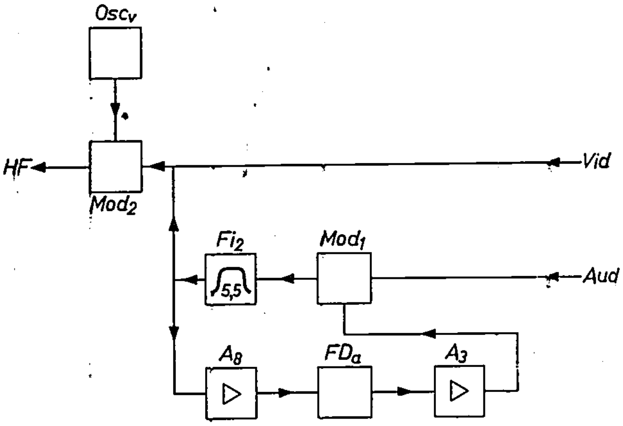


Fig. 11. Block diagram of the circuit for playback of the recorded signals through a normal television receiver. *Vid* video signal. *Aud* audio signal. Mod_1 frequency modulator. Mod_2 amplitude modulator. Osc_v oscillator. Fi_2 band-pass filter. A_8 intercarrier amplifier. FD_a frequency detector. A_3 d.c. amplifier. *HF* connection to the aerial terminal of the receiver.

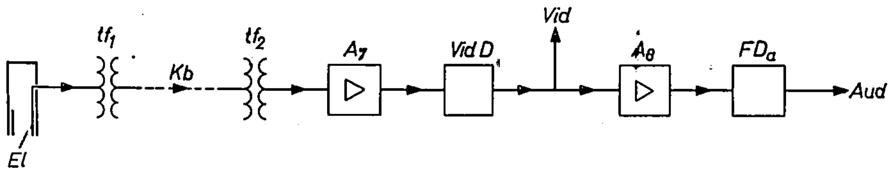


Fig. 10. Block diagram of the circuit for deriving the signals to be recorded from a television receiver. *El* ring electrode (see fig. 9). tf_1 and tf_2 transformers. *Kb* coaxial cable. A_7 I.F. amplifier. *VidD* video detector. A_8 intercarrier amplifier. FD_a frequency detector. *Vid* video signal. *Aud* audio signal.

To economize on circuit components, the same multivibrator that is used for converting the video signal into an FM signal during *recording* serves for generating the modulated auxiliary carrier. As in fig. 6, this multivibrator is denoted in fig. 11 by Mod_1. (The frequency, which is between 3 and 4.3 Mc/s for recording, is set, upon switching over, to a centre value of 5.5 Mc/s.) The signal from this modulator, together with the video signal, goes via the band-pass filter Fi_2 to the modulator Mod_2 mentioned above. The output voltage from Mod_2 is the aerial signal for the television receiver. A free channel can be selected with a switch on the oscillator Osc_v; the receiver must also, of course, be set to the same channel.

The centre frequency of the multivibrator Mod_1 is stabilized at a value of 5.5 Mc/s by means of a simple feedback system. The output signal from filter Fi_2 is passed via amplifier A_8 to the frequency detector FD_a. (These two units are the same as used in the *recording* of the audio signal, and are denoted by corresponding letters in fig. 10.) The d.c. voltage supplied by FD_a is amplified in the d.c. amplifier A_3, which was also used in the recording of the video signal (see fig. 6) and applied to the multivibrator Mod_1. Following this scheme, the centre frequency of the auxiliary carrier wave can be kept with sufficient accuracy at 5.5 Mc/s without the use of extra components.

The synchronizing circuits

To ensure that, during recording, the rotating disc with the video head completes exactly one revolution per frame and that, during playback, this head travels over the recorded tracks, two synchronizing circuits are provided, one of which governs the rotation speed of the disc S and the other the tape speed. The demands made on these circuits are rather severe, due to the fact that the video recorder must function satisfactorily in combination with a normal television receiver. In most television receivers the line time-base has flywheel synchronization. This strongly reduces the influence of interference on the line synchronization, since the playback line frequency adjusts itself to the *mean frequency* of the line synchronizing pulses received over a certain time interval. Incidental interference pulses cannot then cause the line flyback to occur at undesired moments, and so in spite of the presence of interference an undisturbed picture is obtained. For recording and playing back the pictures through a tape recorder, however, this imposes the requirement that the rotation speed of the disc S and the tape speed, which govern the line frequency on playback, must be extremely constant — just as constant as the line synchronization frequency of the signal sent out by the television transmitters. In the video recorder described here we have met this requirement rather simply by taking as our basis the mains frequency, which is sufficiently constant for the purpose. Further details are given below.

A diagram of the synchronization circuits in the "record" position is shown in *fig. 12*. The video signal, obtained from the point marked a in fig. 6, is connected via a in fig. 12 to the sync separator SS, which is identical with the corresponding circuit in a television receiver. The frame synchronizing pulses obtained from the signal in this circuit are fed to the synchronizing head K_s, which records these pulses at the upper edge of the magnetic tape.

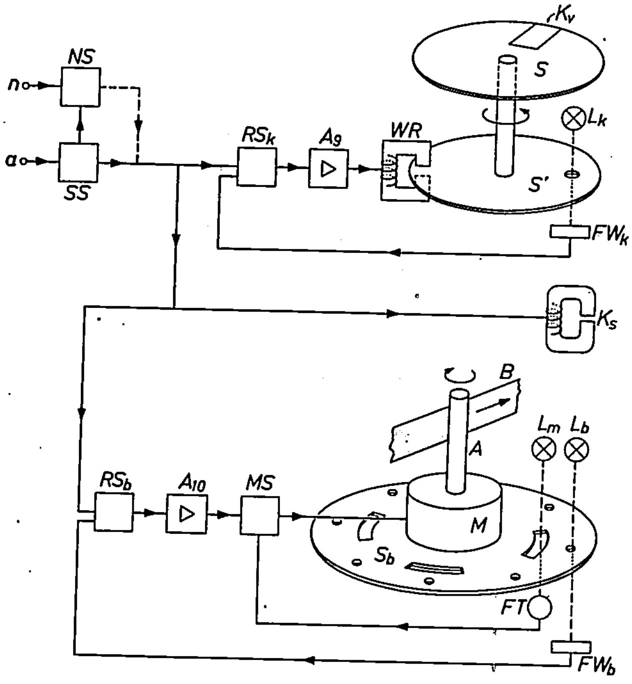


Fig. 12. Diagram of synchronizing circuits used in recording. a connection for the amplified video signal (see fig. 6). SS sync separator. n mains connection. NS circuit for replacing frame synchronizing pulses during a temporary interruption of the video signal. RS_k and RS_b control circuits. A_9 and A_{10} amplifiers. S' disc mounted on the same spindle as the video head disc S. K_v video head. WR eddy-current brake. L_k, L_m and L_b lamps. FW_k and FW_b photoresistors. K_s synchronizing head. M motor with spindle A for driving the magnetic tape B. S_b disc mounted on the spindle of the motor M. FT phototransistor. MS motor control circuit.

To ensure that the disc S with the video head K_v completes exactly one revolution per frame, a second disc S' is fitted to the spindle; this rotates between the electric lamp L_k and the photoresistor FW_k. As there is a small hole in the disc a current pulse is generated at every revolution. These pulses, together with the frame synchronizing pulses obtained from the sync separator SS, are supplied to the control circuit RS_k [6]; this delivers a d.c. voltage related to any frequency or phase difference between the two pulse trains. This d.c. voltage, amplified in A_9, actuates the eddy-current brake WR. In this way the number of revolutions per second described by the disc S', and thus also by the video head, remains equal to the frequency of the frame pulses; at the same time, suitable design ensures that, at every frame pulse, the video head is at the proper place (just beyond the point where the recording head moves from the incoming to the outgoing tape).

During recording it can happen that the synchronizing pulses are temporarily interrupted. This happens, for example, during a transmission failure and also, briefly, when the receiver is being tuned or switched over from one television channel to another. In such cases strong braking action would be exerted on the discs S and S'. When the synchronizing pulses reappear, it would normally take some time before the correct speed of rotation was reached again and the video signals recorded in the proper manner. This undesired effect is reduced in that, if the synchronizing pulses are interrupted, other pulses derived from the mains voltage are automatically supplied in their place. The relevant circuit, marked NS in fig. 12, thus "opens" when the synchronizing pulses cut out. The speed of the disc is in this way maintained at 50 revolutions per second, and, when the pulses return, only the phasing of the disc has to be readjusted.

The control system governing the *tape speed* works broadly in the same way as that for the video head disc. The disc S_b employed for the optical collector (see fig. 5) also serves for this purpose. This disc and the motor control circuit MS are shown again in fig. 12. In addition to the four slots for the optical collector the disc contains eight holes which, when illuminated by an electric lamp L_b, throw light pulses on the photoresistor FW_b. Since, at the correct tape speed, the disc S_b rotates at a speed of 6¼ revolutions per second, 50 pulses per second are generated. These, together with the frame synchronizing pulses, are fed to a second control circuit, RS_b. The d.c. voltage obtained from this goes, via amplifier A_{10}, to the control circuit which supplies the power for the tape transport motor.

During fast winding and rewinding of the tape the motor is not governed by the control circuit but is connected to a higher voltage, so that the speed increases to about 75 revolutions per second.

When the video recorder is used for *playback* the two synchronizing circuits function in essentially the same way (*fig. 13*). Instead of the frame synchronizing pulses however, mains-derived pulses are now used. These are taken from the point marked n and fed to RS_k and also, via the phase regulator FR, to RS_b. Both the rotation speed of the video head disc and the frequency of the frame pulses supplied by the synchronizing head K_s are thus made equal to the *mains frequency*. As the mains frequency never fluctuates very rapidly, this gives a sufficiently constant frame frequency to enable playback of television pictures in conjunction with a receiver having flywheel synchronization.

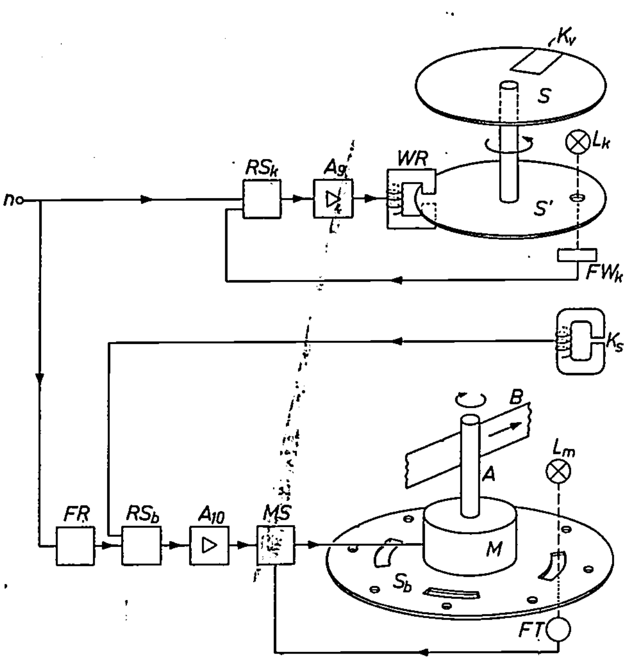


Fig. 13. Diagram of the synchronizing circuits used for playback. FR phase regulator. The other letters have the same significance as in fig. 12.

With the phase regulator FR the user can change the relative phase of the mains voltage pulses supplied to RS_k and RS_b. This can be set so that the video head runs exactly over the tracks. If this is not the case, a moiré pattern can be seen in the reproduced picture; an example is shown in fig. 5 of the article quoted[1].

Some technical details

The electronic circuits of the recorder are mounted on a chassis which can be swung out at the rear of the cabinet for servicing; see *fig. 14*. For the signal-handling circuits valves are mainly used, whereas the synchronizing circuits are mostly transistorized. The recorder has in all 21 valves, 45 transistors and 25 semiconductor diodes.

For change-over from recording to playback and vice versa, there are a large number of switches for the various electronic circuits. The user need only turn a single switch, however; a servo motor does the rest. The two tape reels are each driven by a fan-cooled motor. Including the motors for the video head disc and for the capstan, there is a total of eight motors.

Tape reels of 20 cm diameter can be used which take as already mentioned, a magnetic tape 540 metres long, giving a playing time of 45 minutes. Fast forward winding or rewinding of the whole tape takes 4½ minutes.

The video recorder has five control knobs, which can be seen in the title photograph. They serve respectively for regulation of the video signal level, adjustment of

[6] The design of a similar circuit is described in reference [1], pp. 8-10.

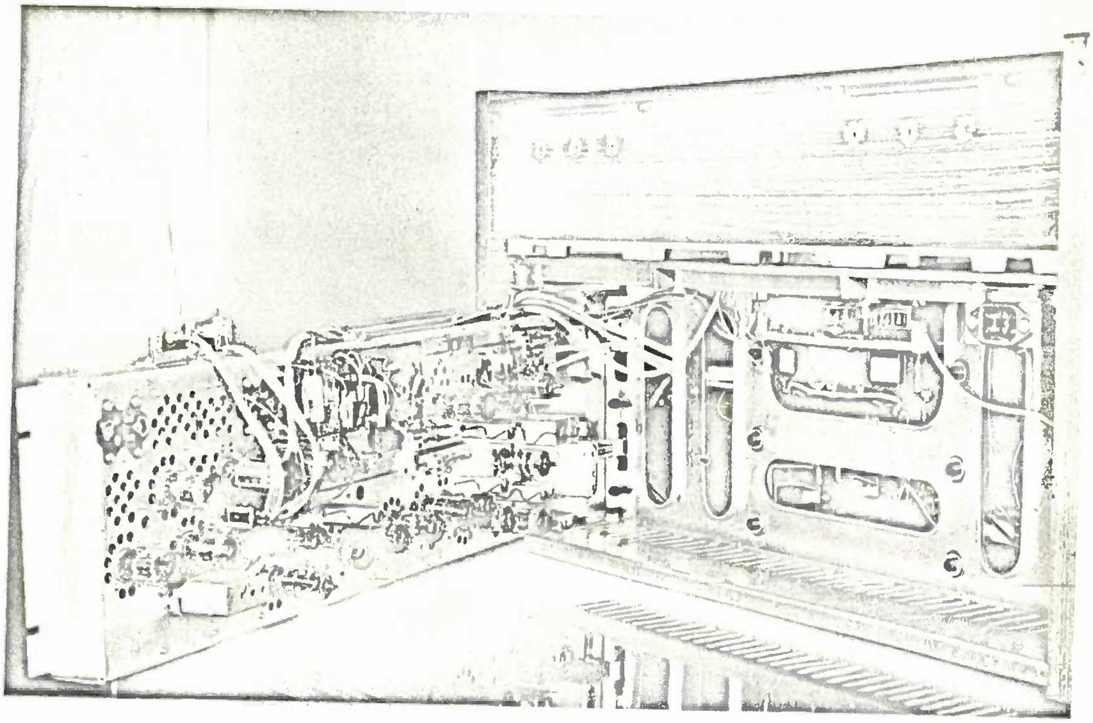


Fig. 14. Rear view of the video recorder. The chassis carrying the electronic assembly has been swung outwards.

the phase of the mains pulses which govern the tape speed during playback (*FR* in fig. 13), regulation of the tape tension, regulation of the audio signal level when recording with a microphone, and regulation of the audio signal when recording from a television receiver. Above these knobs are located the visual indicators for controlling video and audio levels.

The recorder measures 63×42×39 cm. It weighs 45 kg. The power consumed from the mains is 400 W.

An important part in the early development of this video recorder was played by Ir. W. van den Bussche of Philips Radio, Gramophone and Television Division, Eindhoven.

Summary. To record television pictures on magnetic tape it is necessary to record video signals with frequencies up to a few Mc/s. The high relative speed of tape and recording head which this calls for is obtained at a low tape speed (19 cm/s) in the apparatus described by arranging for the head to rotate in a drum around which the tape travels. The tracks (one per frame) are recorded slanting across the tape. At the same time the frame synchronizing pulses are recorded along one edge of the tape, and the sound signals along the other edge.

Pictures and sound can be recorded using a camera and a microphone (or record player) and the recorder can also be connected to a normal television receiver. The receiver does not have to be radically modified for this purpose, the signal being taken from the last intermediate frequency valve by means of an electrode enclosed in a can around this valve. During recording, the rotational speed of the video head is kept in synchronism with the frequency of the frame synchronizing pulses. During playback the tape speed and the rotational speed of the head are controlled in such a way that the frequency of the frame synchronizing pulses equals the mains frequency. If required, playback is also possible with the aid of a monitor and an amplifier with loudspeaker.

Crystal growth of silicon carbide (II)

H. B. Haanstra and W. F. Knippenberg

548.52:546.281'261

Some time ago [1] we discussed here crystal forms which appear in the initial stage of the growth of SiC crystals during the reaction between an intimate mixture of carbon and quartz powder at a temperature of 1500 °C in an argon atmosphere. The observations were made with an electron microscope. The crystals grow as "whiskers" and are in general very long. They are formed mainly on the carbon particles.

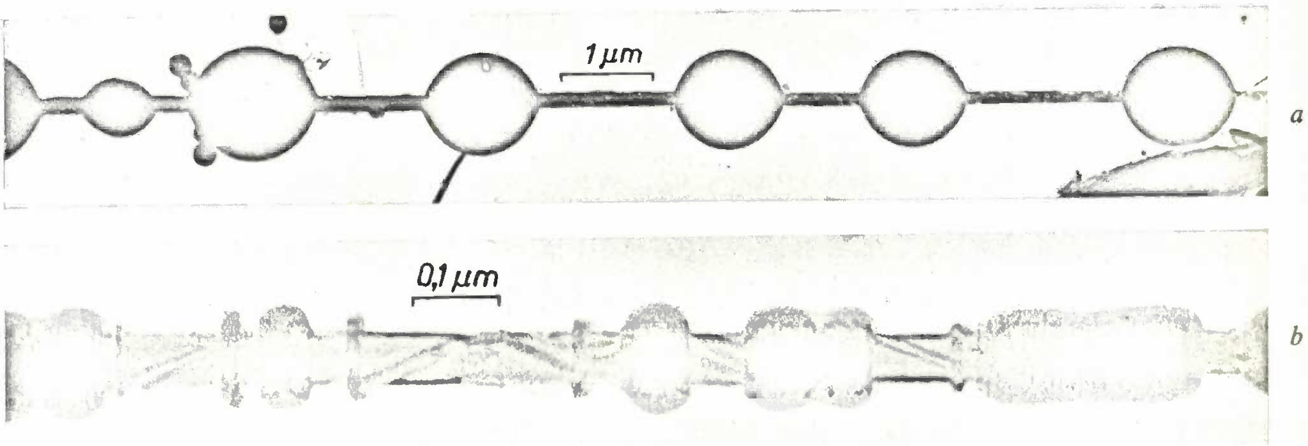


Fig. 1. Electron-photomicrographs of two types of whiskers grown in the reaction between carbon and quartz powder.

Some of these whiskers show local thickening, which may be globular, giving the crystal the appearance of a string of beads (*fig. 1a*), or may take the form of "sleeves", in which can be seen a fine structure of dark lines perpendicular to the long axis of the crystal; these lines do not continue right through to the surface of the sleeve (fig. 1*b*). The rounded outer part of the sleeve gives the impression of being amorphous. The micrograph does not show any internal structure of the beads, the latter being so thick as to be opaque to electrons of the energy used.

Recently we also found crystal growth in the form of whiskers in the reaction between silicon *vapour* and carbon at a temperature of about 1500 °C. Here again, very long and whiskery crystals grow in the first stage (*fig. 2*); the striking thing here, however, is that no local thickening is found on these whiskers. The crystals have a constant diameter over their whole length and there are no marked differences between individual whiskers.

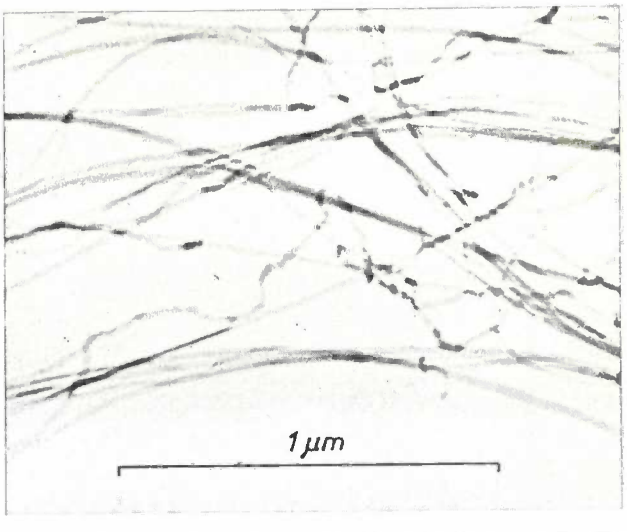


Fig. 2. Electron-photomicrograph of whiskers grown in the reaction between carbon powder and silicon vapour.

When comparing the crystal forms produced by the two different reactions it occurred to us that the local thickening on the whiskers grown from the mixture of carbon and quartz powder might have something to do with the presence of globules of SiO_2. A simple experiment showed that this supposition is very probably right.

H. B. Haanstra and Dr. W. F. Knippenberg are research workers at Philips Research Laboratories, Eindhoven.

[1] W. F. Knippenberg, H. B. Haanstra and J. R. M. Dekkers, Crystal growth of silicon carbide, Philips tech. Rev. **24**, 181-183, 1962/63.

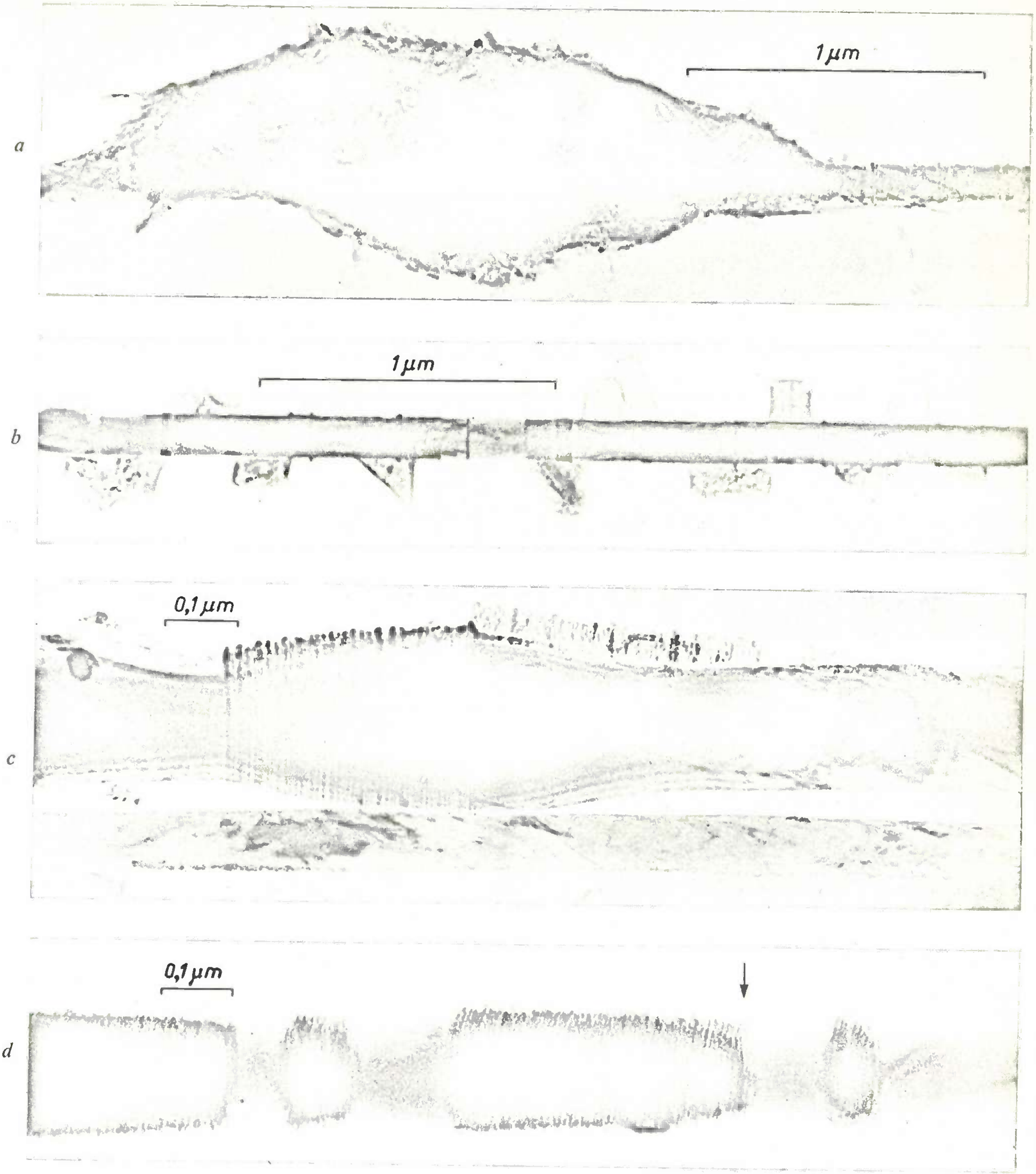


Fig. 3. Some whiskers obtained from carbon and quartz powder after exposure to the action of hydrofluoric acid.

In this experiment the whiskers were exposed for some time to the action of hydrofluoric acid vapour; the reaction of this acid with SiO_2 results in the formation of gaseous SiF_4, while it does not attack SiC. This treatment affects only the thickened parts of the whiskers. The beads have disappeared, and in their place is seen as a rule a tangled mass of whiskery crystallites around the continuing body of the whisker (*fig. 3a*); in some cases outgrowths in the form of platelets are found (fig. 3*b*). On the sleeves the amorphous-looking part disappears, while the cross-lines become much more distinct (fig. 3*c*). The experiment makes it seem likely that the parts which disappeared consisted of SiO_2. On a whisker whose long axis is oriented obliquely to the direction of the incident electron beam, it can be seen that the bands are due to the presence of rings around the continuing body of the crystal (fig. 3*d*). At these places the crystal itself is usually slightly thickened. Sometimes one has the impression that the rings are hexagonal on the outside (the hexagon can be best seen at the part marked by the arrow in fig. 3*d*).

In view of these observations the following hypothesis to explain the growth seems reasonable. Small globules of SiO_2 on a growing SiC whisker react locally with carbon diffusing over the crystal surface [2]. The SiC thereby produced first forms rings around the whisker and then grows epitaxially on the crystal already present. This assumption is supported by the fact that, on whiskers not yet exposed to hydrofluoric acid, an amorphous layer is visible around each ring (see fig. 1*b*). It is not yet clear why the growth in the beads takes place in such a disordered manner; it may be bound up in some way with the size of the beads, the distance between the surface of the bead where the reaction takes place and the whisker being relatively large.

In many cases a globule was also found at the *end* of a whisker (*fig. 4*). This might be a confirmation of the hypothesis of Wagner and Ellis [3], who assume that the whiskers originate in a globule. Whether, however, *every* whisker has such a globule we were not able to ascertain.

According to another hypothesis, put forward in the article [1] already cited, the growth of a whisker might be the result of a lattice defect in the long axis of the crystal. In the material discussed at the time we did not observe any crystals that had such a defect. Very many more whiskers have now been investigated, and a few specimens have indeed been found in which a line is visible in the long axis of the crystal, possibly indicating the presence of an axial defect (*fig. 5a*). The remarkable thing is that in some cases these lines do not continue to the end of the crystal (fig. 5*b*). The number of whiskers showing a line of this kind is extremely small compared with those that seem to grow without an axial lattice defect.

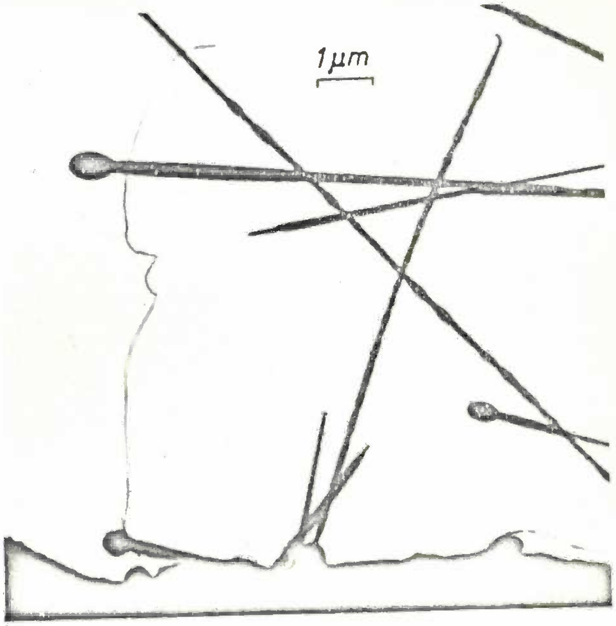


Fig. 4. Whiskers with globules at the end.

The photomicrographs reproduced were selected from a large number made by Mrs. Gijsbers-Dekkers with the Philips electron microscopes EM 100-B and EM 200.

[2] W. F. Knippenberg, Growth phenomena in silicon carbide, thesis Leiden, 1963. This work has also been published in Philips Res. Repts. **18**, 161-274, 1963 (see page 235).
[3] R. S. Wagner and W. C. Ellis, Appl. Phys. Letters **4**, 89, 1964 (No. 5).

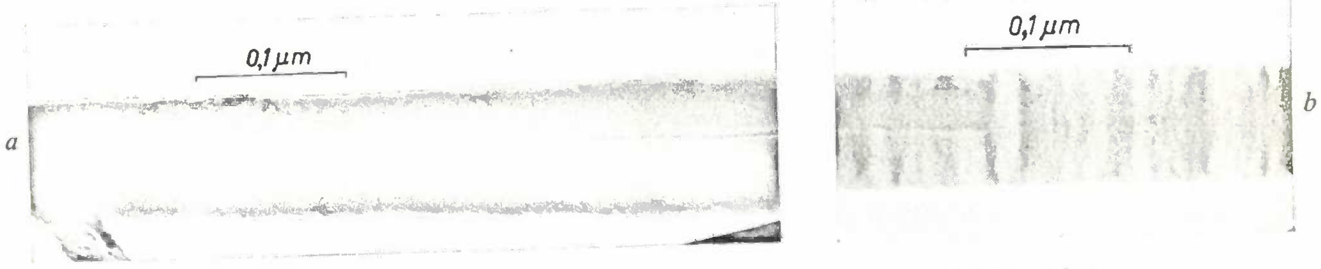


Fig. 5. Electron-photomicrographs of two whiskers showing an axial lattice defect.

A sensitive monitor for X-rays and gamma rays

H. van Ammers and J. Hesselink

539.1.074.2

In hospitals, laboratories and factories there is a growing need for simple, handy instruments for determining the presence and amount of ionizing radiation in various places. These radiation monitors are not required to measure with exceptional accuracy the level of radiation present — or, to be more exact, the exposure rate [1] — but they do have to be highly sensitive, i.e. capable of detecting very low exposure rates. Moreover, in certain cases, for example in X-ray diagnostic rooms, they are required to react to relatively soft radiation, i.e. radiation of relatively low quantum energy.

In the following we shall discuss a radiation monitor, developed by Philips, which is extremely sensitive, which responds to soft X-rays and gamma rays and which can be used not only for measuring exposure rates but also for exposure measurements (*fig. 1*). Apart from measuring radiation levels and checking the effectiveness of radiation screening, the instrument can thus also act as a portable monitor, for example for measuring the exposure to which a person working with radioactive material is necessarily subjected during the course of his work. Both for measuring exposures and exposure rates the upper limit of the measuring range can be set at three values, which are respectively, 1, 10 and 100 mR and 1, 10 and 100 mR/h. The indication is accurate to within 15%.

The detecting element in the new monitor is formed by a cylindrical ionization chamber, part of which is visible on the right in fig. 1. The cylinder wall is of polystyrene and is 4 mm thick. The front face is made of tropic-proof laminated paper 0.5 mm thick (70 mg/cm^2). These walls are thin enough to enable the ionization chamber to detect soft X-rays, but thick enough to ensure sufficient mechanical strength (*fig. 2*). If the detected radiation is not particularly soft (quantum energy higher than 35 keV) a polystyrene cover of the same thickness as the side walls can be fitted over the front face. The inside of the chamber wall is conductive.

The second electrode differs from that found in most ionization chambers in that it is formed by a hollow cylinder of graphitic "Philite". This cylinder, which is also sealed at the rear end, contains the components of the electronic circuit whose cleanliness must be protected; the principal parts concerned are the electrometer tube and the measuring resistance dealt with below.

The meter reading over a wide range depends only very slightly on the quality of the radiation (quantum

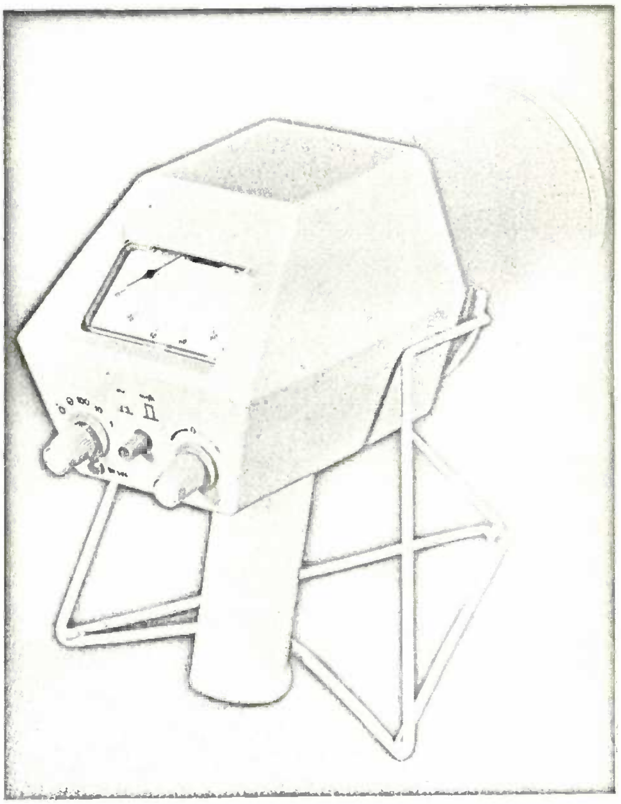


Fig. 1. The radiation monitor XL1000/00, for measuring both exposure and exposure rates. The measuring range for both can be set at three upper limits, viz, 1, 10 and 100 mR or mR/h. The knob on the left is used for selecting the measuring range, the push-button in the middle for changing from "exposure" to "exposure rate". The pointer is set to zero with the knob on the right. The reading is accurate to within 15% in a wide range of radiation qualities. The instrument is specially designed for easy handling and for ease of reading; the weight is about 1.9 kg.

Ir. H. van Ammers and J. Hesselink are on the staff of the Electromedical Laboratory of Philips X-ray and Medical Apparatus Division, Eindhoven.

[1] If it is not a question of measuring the radiant energy absorbed by the body at a particular place but simply of measuring the radiation level at a certain point in space — the measure being the ionization in air — the terms "exposure" and "exposure rate" are nowadays used instead of "dose" and "dose rate".
The definition of the terms "dose "and "dose rate" will be found in the article: J. Hesselink and K. Reinsma, Dosemeters for X-radiation, Philips tech. Rev. 23, 55-66, 1961/62, which also deals with the physical principles of dose measurement.

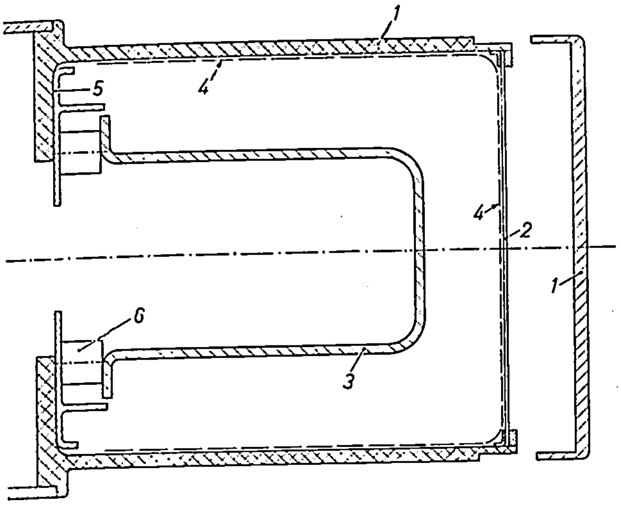


Fig. 2. Axial cross-section of the ionization chamber. The sensitive volume is approximately 850 cm^3. The inner electrode is in the form of a hollow cylinder, which contains the electrometer tube and various other components (R_1, C_1 and C_2 of fig. 4). *1* side-wall and face cover (right) of polystyrene. *2* front face of laminated paper. *3* inner electrode of graphitic "Philite". *4* conductive coating. *5* leak electrode of aluminium. *6* high-insulation methacrylate ring.

energy). It depends rather more markedly on the direction of the incident radiation, but this effect is not significant provided the radiation is not very soft; see *fig. 3*.

The operation of the ionization chamber can be checked by fitting a face cover containing a weak radioactive source (^{137}Cs, radioactivity < 1 microcurie).

The circuit

The circuit (see *fig. 4*) consists of three main parts: 1) the input circuit, 2) the amplifier, and 3) the power supply. In the figure these three sections are shown respectively as *a*, *b* and *c*. We shall start with the input circuit.

When the instrument is switched on, the wall *A* of the ionization chamber is at a potential of about +50 V with respect to the leak electrode *B*. The potential of the inner electrode *C* is roughly equal to that of *B*. When switches D_1 and D_2 are set in the position as shown, irradiation of the chamber causes a current *I* to flow through the resistor R_1 which is almost equal to the ionization current that flows through the chamber from *A* to *C*. In this position of D_1 the exposure *rate* is measured. When the switch D_1 is set to position *2*, the current *I* charges the capacitor C_2, and the *exposure* is then measured. The potential difference V_s produced across R_1 (or C_2) by the current *I* is the input signal to the amplifier. After an exposure measurement, switch D_2 can be used for discharging the capacitor C_2; it can also be used — even in a radiation field — to check the zero setting; see below.

The amplifier (section *b*) consists of two main parts: the actual amplifier circuit (left) and a negative feedback circuit (top right). The first amplifying element is an electrometer tube *E*, and the others are the transis-

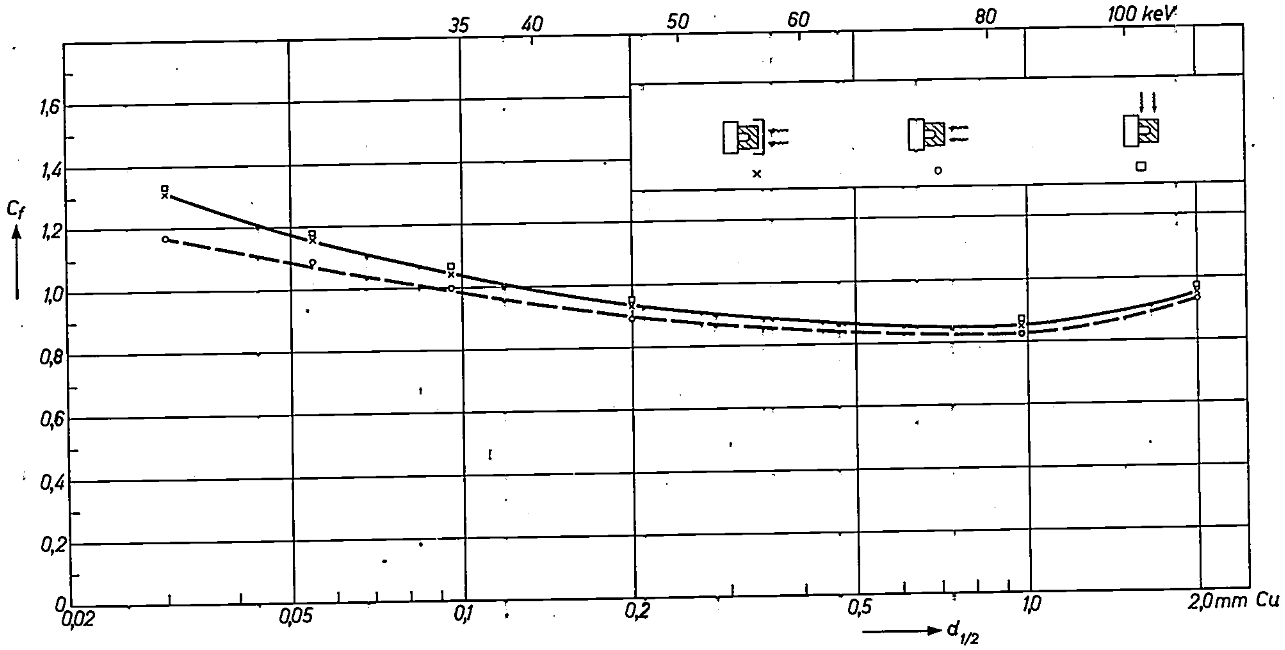


Fig. 3. The meter reading over a wide range depends only very slightly on the quality of the radiation. The figure shows for three situations (see sketches) a plot of the radiation quality (expressed in the half-value thickness $d_{\frac{1}{2}}$ for copper) versus the factor C_f by which the meter reading has to be multiplied in order to find the true exposure rate (or exposure). The solid curve relates to the case where the chamber, with face cover, is irradiated from the front; the dashed line relates to the same case without the cover. Under radiation from the side, values are found which lie roughly on the solid curve. If the radiation enters obliquely, the values lie in between the curves. At $d_{\frac{1}{2}} > 0.1$ mm Cu the direction of incidence is relatively insignificant. The minimum value of C_f is 0.85; in a wide range of radiation qualities the reading thus deviates less than 15%. Also plotted on the abscissa is the quantum energy of monochromatic radiation with the relevant half-value thickness.

tors T_1 to T_4. Without negative feedback the amplifier provides a voltage gain A of 10^4. The anode current of E, which is about 5 μA, is also the base current for T_1. As can be seen, the grid current of the tube E flows through the measuring resistor R_1; this causes the slight difference mentioned above between I and the ionization current (the grid current is at the most 3×10^{-15} A; at 1 mR/h the ionization current is roughly 7×10^{-14} A, which is about 25 times as large).

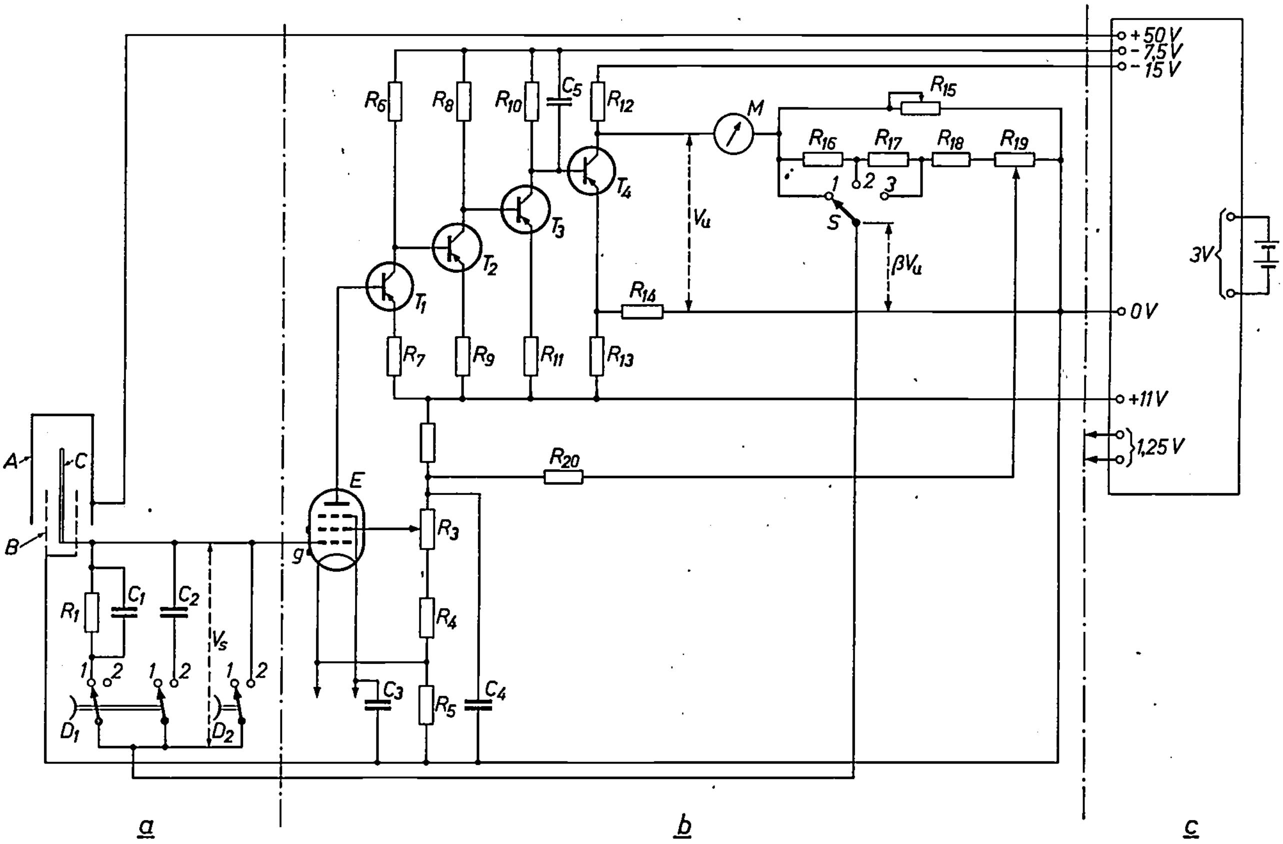


Fig. 4. The circuit; left (section a) the ionization chamber with measuring resistor R_1 and measuring capacitor C_2; centre (section b) the amplifier with meter M and negative feedback circuit (R_{15} to R_{19}); right (section c) the power supply section — the power comes from two 1.5 volt dry cells; the rest of the power pack consists of a converter with a voltage stabilizer.

The wall A of the ionization chamber is connected to the output +50 V of the power pack, the leak electrode B is connected to the output 0 V. The inner electrode is connected to the control grid of the electrometer tube E (type CK 5889; this tube has an external leak electrode g). The connection between C and E is insulated with methacrylate. The value of R_1 is 10^{12} Ω; C_1 consists of two capacitors of 5.6 pF in series and has a leak resistance of more than 10^{13} Ω. Switch D_1 is used for switching from "exposure rate" to "exposure"; D_2 can be used for checking the zero setting in a radiation field and also, after an exposure measurement, for discharging C_2. T_1 to T_4 are silicon transistors (type BCZ 11) which are relatively insensitive to temperature variations. The switch S is used for switching over to a different measuring range, thereby changing the degree of negative feedback. The coarse adjustment of the zero setting is effected by varying the screen grid voltage with R_3; the fine adjustment is effected via the negative feedback by means of R_{19}. Capacitors C_3, C_4 and C_5 prevent the amplifier from going into oscillation.

The negative feedback circuit is formed by the resistors R_{15} to R_{19} and the three-position switch S. This switch serves not only for stabilizing the amplification factor but also for selecting the required measuring range. Setting switch S to a different position changes the part β of the output signal V_u which is fed back to the input. The values of the resistors are such that β, in the three positions of S, has the values 10^{-2}, 10^{-1} and 1, respectively. The amplification factor $A/(1 + A\beta)$ of the amplifier with the feedback switched on, which is almost equal to $1/\beta$, thus has the respective values 100, 10 and 1. The variable resistor R_{15} serves for fine adjustment. The resistor R_3 is used for coarse adjustment of the zero point, and R_{19} for fine adjustment; R_3 varies the screen-grid voltage of the electrometer tube, and R_{19} controls the voltage returned to the input via the feedback circuit.

Adjustment of the zero setting with resistor R_{19}, which is part of the negative feedback circuit, is necessary because in general the meter pointer has to be returned to zero while the switch D_2 is open, in order to eliminate the grid current contribution. However, the time constant of the input circuit is fairly long (3 seconds). This means that any change in V_u is only slowly followed by the corresponding change in the potential V_C of C. In the time that elapses before the new equilibrium state is reached the meter M gives a certain reading. With a direct control V_C would have to be changed just as much as V_u, and the spurious reading would be troublesome. When however the control is effected via the negative feedback, the variation of V_C corresponding to the required variation of V_u is very much smaller, and the unwanted deflection is scarcely noticeable.

The power for the new monitor is supplied by two 1.5 volt dry cells connected in series; cells of this kind are available in most parts of the world. The voltages required for the ionization chamber and the amplifier are obtained from these cells by means of a converter together with a stabilizing circuit. The converter is of a conventional type, consisting of a transformer and two transistors acting as a switch [2]. The output voltage of the transformer is stabilized with a Zener diode, i.e. a solid-state diode which breaks down at a fairly low, accurately defined reverse voltage [3]. This stabilized secondary voltage acts as the input voltage to a second transformer, the output voltage of which is rectified and smoothed. The circuits for —7.5 V and —15 V, required for the transistors, have a very low impedance at low frequencies.

Since transistors are used in the amplifier, the meter deflection cannot be entirely unaffected by room temperature. This effect, however, is minimized by using silicon transistors [4]. For exposure-rate measurements — when the pointer will in any case generally be reset to zero — the temperature effect is insignificant.

Summary. A radiation monitor is described which has an ionization chamber with a thin front face and thus responds to fairly soft radiation. The monitor is also highly sensitive. It can be used for measuring both exposures and exposure rates. For both cases there are three measuring ranges, viz: 1, 10 and 100 mR or mR/h. Under irradiation the voltage which the ionization current produces across the measuring resistor (or capacitor) is first amplified by an electrometer tube and then by four transistors. The amplifier is provided with a negative feedback circuit, which is also used for switching to a different measuring range. The power is supplied by two 1.5 volt dry cells and a converter in conjunction with a voltage stabilizer. A source (^{137}Cs; radioactivity $<$ 1 microcurie) is provided to check the instrument; the pointer can be accurately set to zero and checked at any time, even in a radiation field. The meter reading is only weakly dependent on the quality of the radiation and is accurate to within 15% in a wide range of radiation qualities.

[2] See for example T. Hehenkamp and J. J. Wilting, Philips tech. Rev. 20, 362, 1958/59.

[3] See for example L. P. Hunter, Handbook of semiconductor electronics, chapter 1, McGraw-Hill, New York 1962.

[4] See for example J. P. Beijersbergen, M. Beun and J. te Winkel, Philips tech. Rev. 20, 122, 1958/59: in particular page 134.

Doping methods for the epitaxial growth of silicon and germanium layers

J. Goorissen and H. G. Bruijning

548.4:669.782/.783

It has long been known that single crystals of silicon and germanium can be grown from the vapour as well as from the liquid. The term "epitaxial growth" is used when the crystal grows in the form of a layer on a crystal platelet (the substrate), from which it derives its crystal orientation. The growth from the gas phase can take place at relatively low temperatures, and the crystals thus obtained are comparable in quality to crystals grown from the liquid phase. Epitaxy has become an indispensable aid in the fabrication of semiconductor devices. By the addition of suitable impurities during the process (doping) P- and N-type layers can be formed in any sequence and with any desired concentration gradient. This article describes various doping techniques by means of which multilayers of silicon and germanium with sharp junctions can be achieved.

Introduction

Epitaxial growth methods

It has long been known that single-crystal germanium or silicon can be grown from the liquid phase. By sawing and grinding, pieces of crystal are then obtained of the size needed for the fabrication of semiconductor devices. After a surface treatment, the *P-N* junctions and contacts required for a particular transistor or diode are applied by alloying or diffusion techniques.

Single crystals of this kind can also be grown from the vapour phase, particularly in the form of a layer or film deposited on a crystal substrate. When the lattice planes of the substrate continue in the layer, the process is referred to as "epitaxial" growth. Substrate and layer do not necessarily have to be of the same material; it is known, for example, that gallium arsenide can be grown on germanium, and there are numerous other combinations. By the addition of suitable impurities to the vapour phase, *P*-and *N*-type layers having the desired concentration gradient can be grown. In this way certain characteristics, e.g. electrical conductivity, can be varied more or less at will.

Since about 1959 the technique of producing single-crystal layers from the gas phase has rapidly established itself, on the one hand because already existing types of semiconductor devices can be fabricated more easily in this way, and also because it opened up the possibility of producing novel and more complicated circuit devices.

Many methods are known by means of which single-crystal layers can be grown epitaxially from the gas phase and the required doping effected. Although they differ in many respects, a feature common to these methods is that the vapour flows continuously to the substrate and deposits the element there in one way or another. There are three different procedures.

1) A closed reaction vessel contains the starting material, the substrate, the impurity to be added, and a gas which takes part in the reaction and effects the transport. A thermal gradient is brought about in the vessel, and, since this method depends on an equilibrium state, the establishment of that equilibrium will differ at the two ends of the vessel. Apart from the material to be deposited, only gaseous reaction partners occur, so that if the conditions are properly chosen, this material will be transported by the chemical reaction via the gas phase from one end to the other, and there deposited. This transport takes place by diffusion or convection. The parameters in this process are the temperature at which the growth occurs and the thermal gradient. Schäfer and his associates have investigated many of these transport reactions [1]. Marinace has grown germanium epitaxially by means of the iodide equilibrium [2]:

$$2\,\mathrm{GeI_2} \underset{600\,^\circ\mathrm{C}}{\overset{400\,^\circ\mathrm{C}}{\rightleftarrows}} \mathrm{Ge} + \mathrm{GeI_4} + Q.$$

If an impurity has to be incorporated, the condition then is that its transport, e.g. with an analogous reaction, must take place in the same direction.

2) The second procedure makes use of an *open system*. A gas mixture is produced which consists of a com-

J. Goorissen and Ir. H. G. Bruijning are research workers at Philips Research Laboratories, Eindhoven.

pound of the semiconductor, a compound of the impurity and a gaseous carrier, and this mixture is conducted to the space where the epitaxial layer grows. There it flows over the substrate, the temperature of which ensures that the semiconductor is deposited and grows on the substrate epitaxially. The gaseous reaction products are expelled. In this method the parameters are the concentration of the compounds of semiconductor and impurity, the rate of flow of the gas and the growth temperature. Theuerer [3], Marck and others have produced epitaxial layers of silicon on silicon by the reduction of silicon tetrachloride with hydrogen in accordance with the equations:

$$SiCl_4 + Si \rightleftarrows 2\,SiCl_2, \quad \ldots\ldots \quad (1)$$

$$SiCl_2 + H_2 \rightleftarrows Si + 2\,HCl. \quad \ldots\ldots \quad (2)$$

This procedure in an open system (that is at 1 atm) is very widely employed, and is already used on a production scale. In the doping methods described in this article, both for germanium and for silicon, use is also made of the reduction of the tetrachloride by hydrogen. In recent times epitaxial techniques have been developed based on the thermal decomposition (pyrolysis) of hydrides [4], e.g. silicon hydride: $SiH_4 \rightarrow Si + 2H_2$. This involves no reaction that first affects the substrate itself, as under (1). As a result the concentration boundary between substrate and layer is sharper [5] [6].

3) Of an entirely different nature are the methods in which the element itself is evaporated and condensed. Molten silicon or germanium, to which the impurity may have already been added, is evaporated in a vacuum (10^{-5} to 10^{-6} torr) and epitaxial growth is produced by condensation on a heated substrate. Compared with chemical methods this process is notable for the high rate of growth [7].

Doping methods

In the following we shall confine ourselves to the production of silicon and germanium layers by the reduction of the tetrachloride in an open system. Doping can be carried out by adding to the gas mixture volatile compounds of the appropriate impurity, and, if the concentration of the compound can be varied rapidly enough in the gas, it is possible to obtain successive layers of different conductivity so as to give to the epitaxial structure the electrical characteristics which are desired.

First of all, these compounds can be dissolved in the silicon tetrachloride, which is liquid at room temperature. Theuerer [8] obtained reproducible results by evaporating a mixture of this kind in such a way that the vapour had a constant composition: the special construction of the apparatus he used for this purpose enabled him to give the vapour the same composition as the liquid — in spite of the generally different vapour pressures of the components. This method is particularly suitable for producing large quantities of epitaxial structures with identical properties.

Secondly, a vaporized doping mixture can be injected into the gas mixture, as described by Corrigan [9]. For this purpose the halides of boron and phosphorus, which are liquid at room temperature, have a suitable vapour pressure. The injection is carried out by means of a diffusion mechanism, so that the injected quantity depends only on the temperature of the liquid dope. With this technique it is possible to produce numerous structures consisting of various layers that possess defined properties. Even better — in particular more quickly variable — is the use of volatile hydrides of boron and phosphorus (Cave and Czorny [10]).

In the following sections of this article we shall describe the doping methods which we employ. For the doping of silicon epitaxial layers we have developed the *spark-doping* technique. Between two electrodes consisting of the element to be added, or which contain that element, a spark discharge is generated in the mixture of silicon tetrachloride and hydrogen. This gives rise during the discharge to a narrow zone of high energy density, in which the doping compound is formed by reaction with silicon tetrachloride and hydrogen and transported by the gas mixture. By varying the repetition frequency and the energy of the sparks it is possible to change at any required rate the concentration of the doping compound in the gas. Using this technique it is possible to grow highly reproducible multiple as well as single layers.

For germanium layers we use a *gas-doping* method

[1] H. Schäfer, Chemische Transportreaktionen, Verlag Chemie, Weinheim 1962.

[2] J. C. Marinace, Epitaxial vapor growth of Ge single crystals in a closed-cycle process, IBM J. Res. Devel. 4, 248-255, 1960.

[3] H. C. Theuerer, Epitaxial silicon films by the hydrogen reduction of $SiCl_4$, J. Electrochem. Soc. **108**, 649-653, 1961.

[4] S. E. Mayer and D. E. Shea, Epitaxial deposition of silicon layers by pyrolysis of silane, J. Electrochem. Soc. **111**, 550-556, 1964.

[5] E. A. Roth, H. Gossenberger and J. A. Amick, The growth of germanium epitaxial layers by the pyrolysis of germane, RCA Rev. **24**, 499-510, 1963.

[6] B. A. Joyce and R. R. Bradley, Epitaxial growth of silicon from the pyrolysis of monosilane on silicon substrates, J. Electrochem. Soc. **110**, 1235-1240, 1963.

[7] J. C. Courvoisier, W. Haidinger, P. J. W. Jochems and L. J. Tummers, Evaporation-condensation method for making germanium layers for transistor purposes, Solid-State Elecctronics 6, 265-270, 1963.

[8] H. C. Theuerer, Steady-state evaporation method for composition control of thin films prepared by halide reduction, J. Electrochem. Soc. **109**, 742-743, 1962.

[9] W. J. Corrigan, Doping of silicon epitaxial layers, Conf. "Metallurgy of semiconductor materials", Los Angeles 1961, pp 103-111, Interscience Publ., New York 1962.

[10] E. F. Cave and B. R. Czorny, Epitaxial deposition of silicon and germanium layers by chloride reduction, RCA Rev. **24**, 523-545, 1963.

The apparatus is in principle the same as that used for silicon, except that the doping compound is not produced in the spark but is added in a gaseous form to the mixture of germanium tetrachloride and hydrogen.

Spark-doping of silicon

Apparatus used

Fig. 1 shows schematically the experimental set-up for the spark-doping of silicon. The carrier gas, hydrogen, is purified with a palladium filter, after which it is split into two streams, one of which is saturated with silicon tetrachloride. In this way the required hydrogen/chloride ratio (100 : 1) is easily obtained. In all experiments the rate of gas flow is 1 l/min. The silicon tetrachloride used is the commercial "very pure" grade. If silicon is made from this without doping, it becomes *N*-type with a resistivity of approx. 15 Ωcm. After the stream of pure hydrogen and the stream of the hydrogen and silicon tetrachloride mixture have combined, the doping compound can be added in the next section of the equipment with the aid of the spark discharge. This being done, the entire gas mixture is then fed into the reaction vessel. Contained in this vessel is an inductively heated carrier, consisting of silicon with a resistivity (at room temperature) of about 0.1 Ωcm. The induction coil, fed with 400 kc/s, encloses the reaction vessel. The gaseous reaction products leave the system through an absorption tube. Finally the flow rate of the emergent gas is measured.

On the silicon carrier is placed the substrate, which has previously been lapped and treated with a polishing etchant [11]. This makes the surface so smooth that no surface structure is observable under the microscope. During the etching procedure air is passed through the solution to prevent the adhesion of gas bubbles, which would cause non-uniform etching. In the reaction vessel the substrate is first heated to 1275 °C in pure hydrogen for about half an hour to reduce residual surface oxidation. Immediately afterwards the epitaxial layer is grown. At a temperature of 1225 °C a layer 11 ± 1 μm thick grows in 15 minutes. During this time the substrate rotates at a speed of 50 r.p.m., so that on an average the whole surface comes into contact in the same way with the gas stream.

The doping system

The actual doping system consists of a spark generator and several pairs of electrodes, one pair for each doping element. When the spark discharge is set up across one of the electrode pairs, the relevant compound is formed by the reaction of the electrode material with silicon tetrachloride and hydrogen, and is transported in the stream of gas. To obtain *P*-type layers we use boride electrodes (LaB_6 or B_4C) and for *N*-type layers we use antimony, silicon with 0.1% phosphorus, and also antimony with 1% arsenic. The electrodes are contained in a glass tube with metal leads, the whole assembly measuring only a few centimetres (see *fig. 2*).

The concentration of the doping compound in the gas mixture, and hence in the epitaxially grown layer, depends among other things on the nature of the electrode material, the composition of the negative electrode being in any case the governing factor. Hardly any change is to be noticed if, for example, platinum or silicon is used for the positive electrode. The concentration of the doping compound in the gas mixture further depends on the repetition frequency of the spark, and

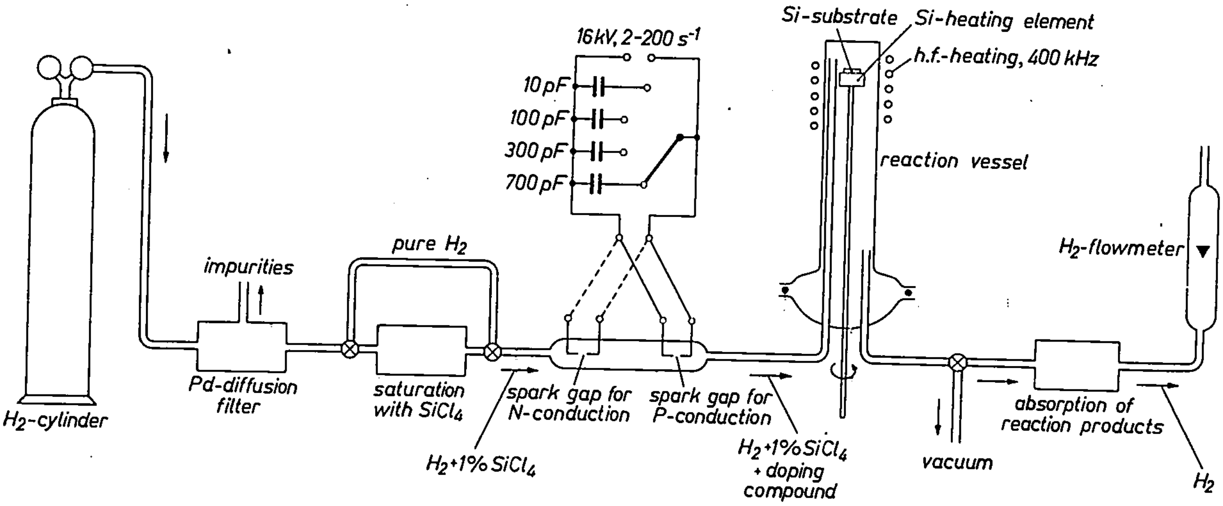


Fig. 1. Experimental arrangement for producing epitaxial silicon layers, using the new doping method (spark-doping). The layer grows on the silicon substrate in the reaction vessel. The substrate is raised to the required temperature by the high-frequency-heated silicon carrier. A spark discharge can be produced across one of two spark gaps having suitable electrodes for doping with elements that give rise to *N*- or *P*- type conductivity. The repetition frequency of the sparks can be varied, and also their energy (by switching-in different capacitances).

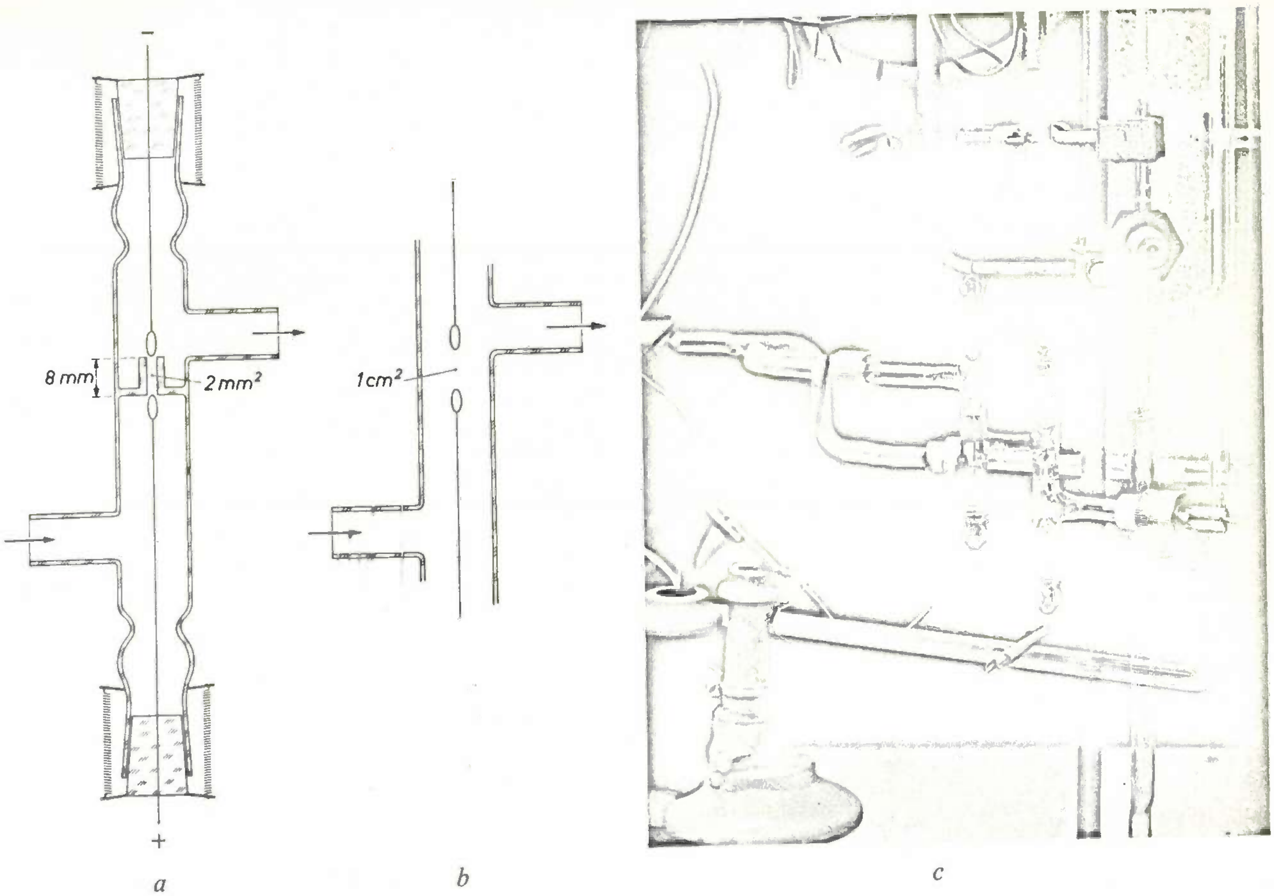


Fig. 2. Configuration of the spark gap. The spark channel has a cross-sectional area of 2 mm² (*a*) or 1 cm² (*b*). In (*c*) a photograph is shown of this part of the equipment (bottom left a part of the capacitors).

on the capacitance of the capacitor which is shunted across the spark and governs its energy. These spark parameters can be varied , the frequency continuously from 2 to 200 per second, the capacitance in steps of 10, 100, 300 or 700 pF. In this way it is possible to vary the energy of the spark by a factor of 7×10^3.

Spark gap and spark generator

Certain requirements which the spark discharge had to satisfy led to an unusual construction of the spark generator. *Fig. 3a* shows the circuit diagram. The spark generator delivers a current pulse of the form shown in fig. 3*b*, which is conducted to the parallel configuration of spark gap and capacitor. Fig. 3*c* gives the form of the corresponding capacitor voltage V, which is likewise the potential between the spark electrodes. The breakdown occurs at a specific voltage, at which an energy $\frac{1}{2}CV^2$ is generated, causing chemical conversions to take place at the electrode surface.

This breakdown voltage is not definitely fixed. It is influenced by a number of factors, such as variations in gas composition, the spark frequency, the shape of the electrodes and the distance between them, so that one can only say that the breakdown occurs at a mean voltage V_m. The ionization produced by the breakdown

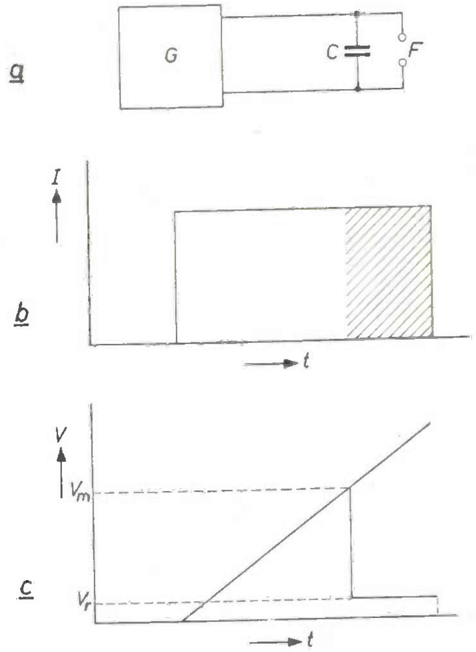


Fig. 3. *a*) Charging circuit. *G* spark generator. *C* capacitor. *F* spark gap. *b*) Current pulse for charging the capacitor. *c*) Voltage on the capacitor and across the spark gap. The breakdown takes place at a mean potential V_m, after which current still flows along the spark channel at a potential V_r.

[11] Saturated $KMnO_4$ solution in 50% HF (room temperature).

enables the rest of the current pulse to flow through the discharge channel thus formed (cross-hatched area in fig. 3*b*). At the pressure and composition of the gas as here employed, this current would flow at the fairly appreciable voltage V_r. This means, however, that after the breakdown, energy will still be supplied to the gas, leading to overheating. This in turn means that an appreciable quantity of silicon tetrachloride will react with hydrogen, in accordance with the bulk equation:

$$SiCl_4 + 2\,H_2 \rightarrow Si + 4\,HCl.$$

As a result the electrodes become gradually coated with silicon and less and less doping compound enters the gas, since the spark no longer encounters the proper electrode composition. What is more, the electrodes grow towards each other.

To avoid these effects a circuit is employed which cuts off the current pulse almost immediately after the breakdown: see *fig. 4*. At the moment of breakdown the sudden change in the potential across the spark gap is used for short-circuiting the current source with the aid of an electronic tube. This occurs about 3 µs after the breakdown, and as a result the deposition of silicon is almost completely suppressed.

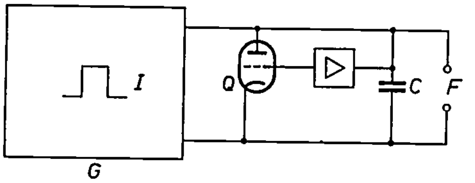


Fig. 4. Circuit for shorting the current source immediately after the spark discharge. *I* current pulse. *Q* transmitting valve type QB 3.5/750. *F* spark gap. *C* capacitor.

The voltage across the spark gap at the moment of breakdown is between 10 and 15 kV in the experiments described here. The valve mentioned must be capable of withstanding this voltage while the capacitor is being charged up. A small transmitting valve, e.g. type QB 3.5/750, meets this requirement. Moreover, the valve must be able to handle the current from the generator circuit at low anode voltage, at which the potential across the spark gap becomes so low that the discharge breaks off. To avoid overloading the valve a generator circuit was chosen that supplies a low current; this current must, of course, flow long enough for the required voltage to be built up across the capacitor.

To meet these conditions the generator must have a high impedance and be able to supply current pulses of relatively long duration. An induction coil of the Ruhmkorff type possesses these properties and is therefore eminently suited to our purpose. The primary current in our case is supplied via a transistor amplifier, using a circuit which generates the repetition frequency required for the discharge.

Results of silicon doping

The choice of the electrode material is governed by a number of factors. For example, there must be no contamination of the mixture of silicon tetrachloride and hydrogen if no spark discharge takes place. Volatile or reactive elements or compounds are therefore ruled out as electrode material. This applies, for example, to arsenic and phosphorus, which are frequently used as doping elements. These elements can, however, be used in the form of a solid solution in an otherwise inert electrode material, as for example phosphorus in silicon. Good results have also been obtained with 1% arsenic in antimony (see *fig. 5*, curve *A*).

In the experiments to which fig. 5 relates, the opposite type of substrate material was in all cases chosen in order to form a *P-N* junction. The conductivity and the layer thickness, from which the concentration is calculated, were measured by previously reported methods [12] [13]. Each point indicated in the figure is the average of at least three experimental values.

P-type layers were obtained using boron. Although the electrical conductivity of this element itself is too low for it to be used as electrode material, it is very useful as such in the form of sintered AlB_{12}, LaB_6 or B_4C. The sintered material must, however, be highly homogeneous. Curves *B* and *C* in fig. 5 give the boron concentration obtained in the epitaxial layer as a function of the spark frequency, with the capacitance as parameter. The figure shows that the expected linear relationship is not found. There are various reasons for this, some of which are bound up with the fact that successive discharges have an increasingly stronger influence on each other as the frequency increases.

Of the factors already mentioned which govern the

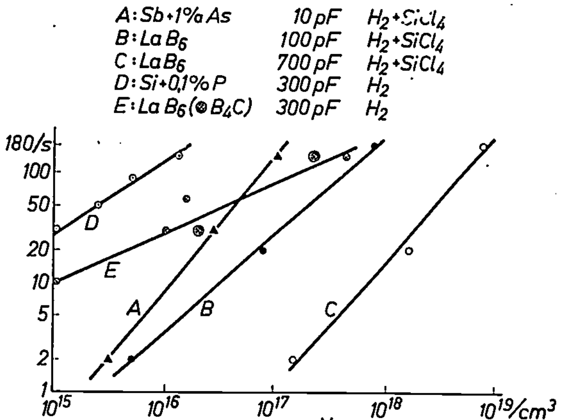


Fig. 5. Concentration of the doping element in the epitaxial layer as a function of the spark frequency, using various electrode materials (curves *A* to *E*). The appropriate capacitance and the gas in which the spark discharge took place are indicated for each curve. In cases *A* and *D* an *N*-type layer is grown, in the other cases a *P*-type layer.

breakdown voltage, and hence the energy generated in each spark, the most important is the $SiCl_4$ concentration. If this concentration is increased by a factor of 2, the spark capacitance and frequency remaining unchanged, four to five times as much doping element enters the epitaxial layer, depending on the electrode material. The presence of hydrogen, incidentally, is essential to the formation of the compound in the spark. If argon is used as carrier gas for the silicon tetrachloride, and the hydrogen is not added until after the spark (hydrogen is in any case necessary for depositing silicon on the substrate in accordance with reactions (1) and (2)), then no doping compound is formed. In the spark discharge the following reactions have apparently taken place:

$$SiCl_4 + 2\,H_2 \rightarrow Si + 4\,HCl, \quad . \; . \; . \quad (3)$$

$$6\,HCl + 2\,D \rightarrow 2\,DCl_3 + 3\,H_2 \quad . \; . \quad (4)$$

the reaction (3) occurring in the spark and (4) at the surface of the electrode (D = doping element). A further argument in support of this statement is the clearly perceptible formation of silicon, which has to be removed after a few experiments by etching the electrodes. The above-mentioned necessity for cutting off the spark is also understandable in this connection. This cut-off is found to benefit considerably both the reproducibility and the quantity of doping compound formed.

The dimensions of the spark channel also influence the reproducibility and the amount of compound formed, so that there must be a direct relation between these two quantities and the spatial extent of the spark discharge. This was investigated with the two configurations shown in fig. 2. It was found that in the narrow tube the amount formed was smaller (about $\frac{1}{6}$) but the reproducibility was better than in the wide tube, given the same electrode spacing and spark parameters.

It is obvious from the foregoing that the use of the spark-doping method involves a large number of factors; it is not difficult in practice, however, to take these into account. The reproducibility is the same as found with other methods (fluctuations of about 25%). The great advantage of the method is that the possibility of fast variation makes it relatively easy to produce very abrupt junctions in the growing layer.

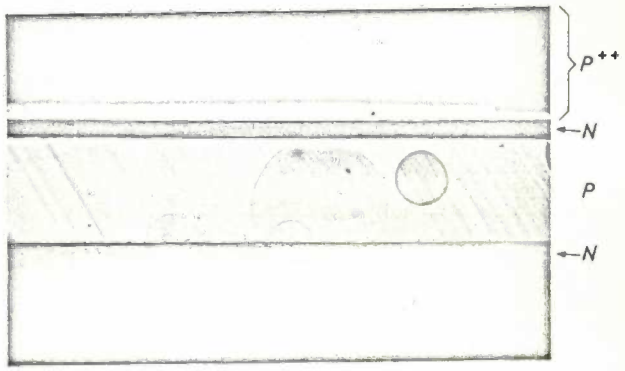


Fig. 6. Cross-section of an *NPN* structure grown on a P^{++} substrate by the spark doping method.

Fig. 6 shows a $P^{++}NPN$ structure whose layers were doped by the spark doping method. On a cleavage face the various layers can be made visible by etching [14]. The current-voltage characteristic that can be measured on the etched surface is illustrated in *fig. 7*.

Finally, it should be mentioned that spark reactions can also be produced in pure hydrogen in a by-pass arm of the system. In that case both with LaB_6 and with Si + 0.1% P the corresponding hydrides are formed instead of chlorides as in reaction (4). Curves *D* and *E* in fig. 5 give the relevant measured results. Since in this case no amorphous silicon is formed and deposited on the electrodes, this method of doping is even more attractive, and is at present the subject of intensive investigation. In this way spark-doping can also be used for germanium epitaxy. The experimental set-up described above for silicon cannot be used for this purpose, because large amounts of amorphous germanium are deposited, in accordance with the reaction:

$$GeCl_4 + 2\,H_2 \rightarrow Ge + 4\,HCl.$$

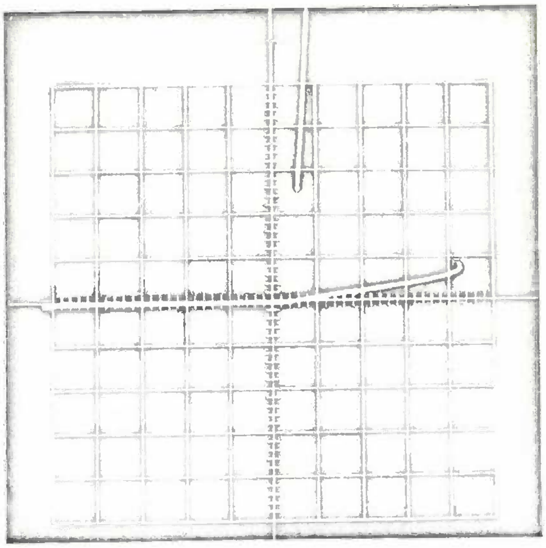


Fig. 7. Current-voltage characteristic of the $P^{++}NPN$ structure shown in fig. 6. Each scale division on the abscissa represents 10 V, on the ordinate 5 mA.

[12] G. Backenstoss, Evaluation of the surface concentration of diffused layers in silicon, and F. M. Smits, Measurement of sheet resistivities with the four-point probe, Bell Syst. tech. J. **37**, 699-710 and 711-718, 1958.

[13] S. Mendelson, Stacking fault nucleation in epitaxial silicon on variously oriented silicon substrates, J. appl. Phys. **35**, 1570-1581, 1964.

[14] The method used was devised by J. Appels of this laboratory. A cleavage plane is etched for about 10 seconds with a solution of 0.1% concentrated nitric acid in 50% HF. The *P*-type conducting parts turn a dark colour.

Gas-doping of germanium

Experimental

As mentioned at the end of the last section, complications arise in the doping of germanium by the spark method. For the doping of epitaxial layers of germanium we have adopted a gas-doping method, the principle of which is illustrated in *fig. 8.* Apart from the actual system of doping, the set-up is basically the same as used for silicon. The high-frequency-heated carrier on which the substrate is placed consists here of molybdenum with a closely fitting quartz sleeve which is open at the bottom end. Reaction of the molybdenum with germanium tetrachloride or with reaction products is negligible in this arrangement.

The liquid germanium tetrachloride is purified by extraction for 24 hours with aqua regia followed by phase separation and distillation. An epitaxial germanium layer grown from this germanium tetrachloride is found to possess the intrinsic conductivity of Ge, i.e. it contains no foreign atoms, provided the substrate only comes into contact with quartz or with polytetrafluorethylene during handling. The pre-treatment of the germanium substrate is similar to that described in the case of silicon, except that the composition of the polishing etchant is different [15]. The reduction temperature is now 875 °C. The gas containing the dope enters the system via stopcocks *A* or *B* in fig. 8. The epitaxial growth takes place at 840 °C; at a $GeCl_4$ concentration of 0.2% a layer of about 10 μm thick is then obtained in 30 minutes.

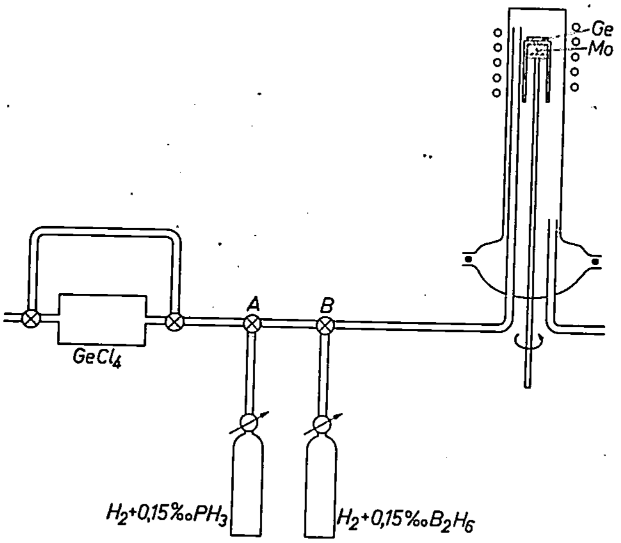


Fig. 8. Experimental arrangement for growing epitaxial germanium layers, using a gas-doping method. Only those parts are shown in which the set-up differs essentially from that in fig. 1. The layer grows on the germanium substrate *Ge*, which in this case is raised to the appropriate temperature by an indirectly heated carrier *Mo* of molybdenum, which is enclosed in a quartz sleeve.

The doping system

The doping of the germanium layers is based on a gas mixture of hydrogen and 0.015% PH_3 or B_2H_6. This mixture is contained in a normal gas cylinder at an initial pressure of 120 atm; the supply is regulated with a reduction valve.

The arrangement of the actual doping system is shown in *fig. 9.* The pressures p_1 and p_2 on the capillaries C_1 and C_2 make it possible to adjust accurately a prescribed concentration of the doping compound in the gas. This is done as follows. The mixture of hydrogen and hydride can be fed in either through T_1 or T_2. Let us first assume that the mixture is supplied via T_1. The mixture then flows only through capillary C_1. The dimensions of the latter are such that, by adjusting the pressure p_1, which can be 100 g/cm² maximum, the flow through C_1 is varied from about 5 to 100 cm³/min. The amount of doping compound supplied varies correspondingly. Now this amount can be reduced by admitting the mixture via T_2 while introducing pure hydrogen through T_1. This decreases the concentration of the doping compound in the hydrogen-hydride mixture in a ratio which can be calculated as follows. Upon variation of the pressure p_2, which can also be 100 g/cm² maximum, the capillary C_2 delivers a gas stream of 0.5 to 10 cm³/min. If we now pass the smallest possible stream through the capillary C_2 (0.5 cm³/min) and allow hydrogen to pass via T_1 until the flow meter *S* gives a reading of 150 cm³/min, at the same time making the gas flow through C_1 as small as possible (5 cm³/min),

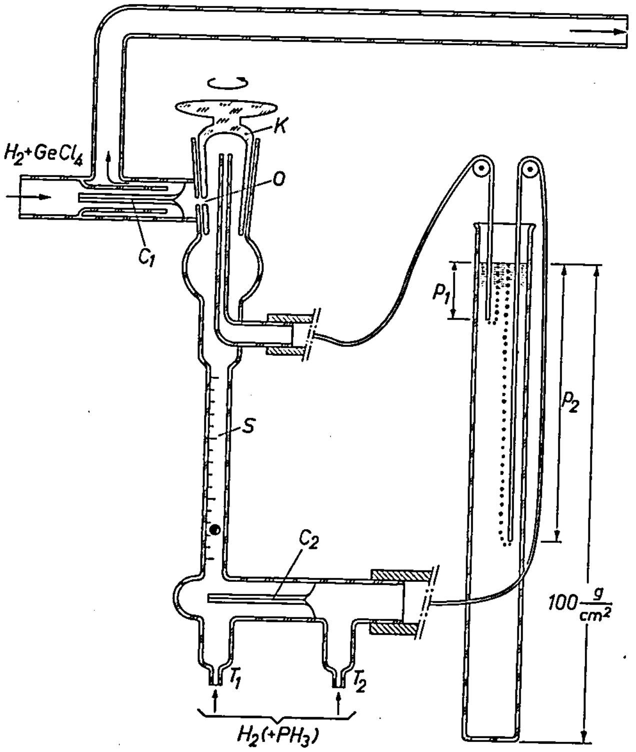


Fig. 9. One of the two doping systems for Ge, which, in the arrangement in fig. 8, are connected at *A* and *B*.

then we have brought down the lower limit of the doping compound supplied to 5×0.5/150 = 0.016 cm³/min. We can thus change the supply of doping compound from 100 to 0.016 cm³/min, that is to say by a factor of 5×10³. For most purposes this is amply sufficient. The dual system illustrated here makes it possible to give the epitaxial layer almost any required resistivity, with *P*- and *N*-type layers following one another in any arbitrary sequence.

In order to obtain a sufficiently sharp junction between two layers, the gas mixture must previously have been set to the appropriate value. This is done as follows (see fig. 9). Before the actual experiment, the hydrogen-hydride mixture is allowed to flow without it being able to reach capillary C_1. For this purpose stopcock *K* is set in a position at which the opening *O* is turned through 180° with respect to the position shown in the figure. After a few minutes the gas mixture has reached the appropriate composition inside the stopcock, and it can then be admitted into capillary C_1 by simply turning the stopcock through 180°. The mixture of germanium tetrachloride and hydrogen now immediately takes up the doping compound with the required concentration.

Results of measurement for germanium

The concentration of the doping element in the epitaxial layer can be calculated from the measured resistivity. In *fig. 10* this concentration N_L (number of charge carriers per cm³ in the layer) is plotted versus the quotient [D] [Ge]/[$GeCl_4$], where [D] and [$GeCl_4$] are the respective concentrations of the doping compound and of $GeCl_4$ in the gas, and [Ge] represents thenumber of germanium atoms per cm³ (4.4×10²²) [16]. It can be seen from the graph that a linear relation exists, so that:

$$\frac{N_L}{[Ge]} = K\frac{[D]}{[GeCl_4]}.$$

If the doping element could be incorporated in the same ratio as the germanium in the layer, the proportionality factor *K* would be equal to unity. In fact, however, one finds *K* = 36, both for boron from B_2H_6 and for phosphorus from PH_3. Plainly, then, *K* is unexpectedly large. If we add these elements with the aid of other compounds (BBr_3 and PCl_3), then *K* has the same values as if the hydrides had been used [17].

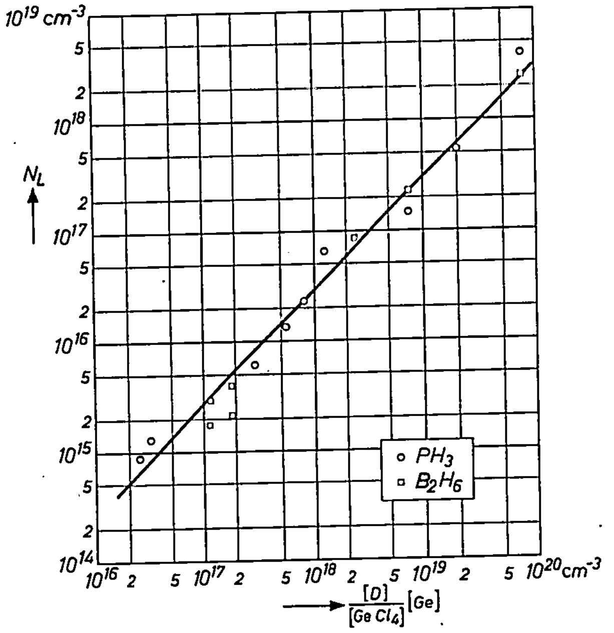


Fig. 10. Concentration N_L of the doping element (i.e. of the charge carriers) in the layer, as a function of the concentration of the doping compound in the $GeCl_4$, expressed as the number of PH_3 molecules per 4.4×10²² molecules of $GeCl_4$.

The proportionality factor also differs from 1 in the case of silicon. Corrigan [9] reports *K* = 5 for PCl_3 and *K* = 10 when BBr_3 is used. In the circumstances described here, values of 5 and 10 are also obtained using PH_3 and B_2H_6.

From these observations it may be concluded that the take-up of foreign substances involves unexpected factors. These do not depend on the equilibrium state for the corresponding chlorides, because this differs considerably for phosphorus and boron. The fact that the value of *K* differs from 1 is attributed by Corrigan to circumstances of geometry. At the present stage, however, attempts to explain this effect would seem to be premature.

Summary. By the reduction of silicon and germanium tetrachloride with hydrogen in an open system, epitaxial layers can be grown and the doping element incorporated during growth. For the doping of silicon layers the *spark-doping method* has been developed in the Philips Research Laboratories, Eindhoven. By means of a spark discharge between two electrodes that contain the doping element, accurately defined quantities of the dope can be added to the mixture of $SiCl_4$ and hydrogen. By changing the spark parameters (capacitance and frequency) it is possible to change with immediate effect the concentration of the doping element in the growing layer. In this way one can obtain sharply defined layers with the required values of resistivity, or structures with very accurately controlled concentration gradients. The doping of germanium layers is done by a gas method, the dope being added in a gaseous state to the mixture of $GeCl_4$ and hydrogen. This addition can be accurately regulated within wide limits. The concentration of the doping element in the layer can be calculated from the measured resistivity of the epitaxial layer. In the case of silicon the authors examine the dependence of this concentration on the spark capacitance and frequency, and in the case of germanium its dependence on the concentration of the doping compound in the gas.

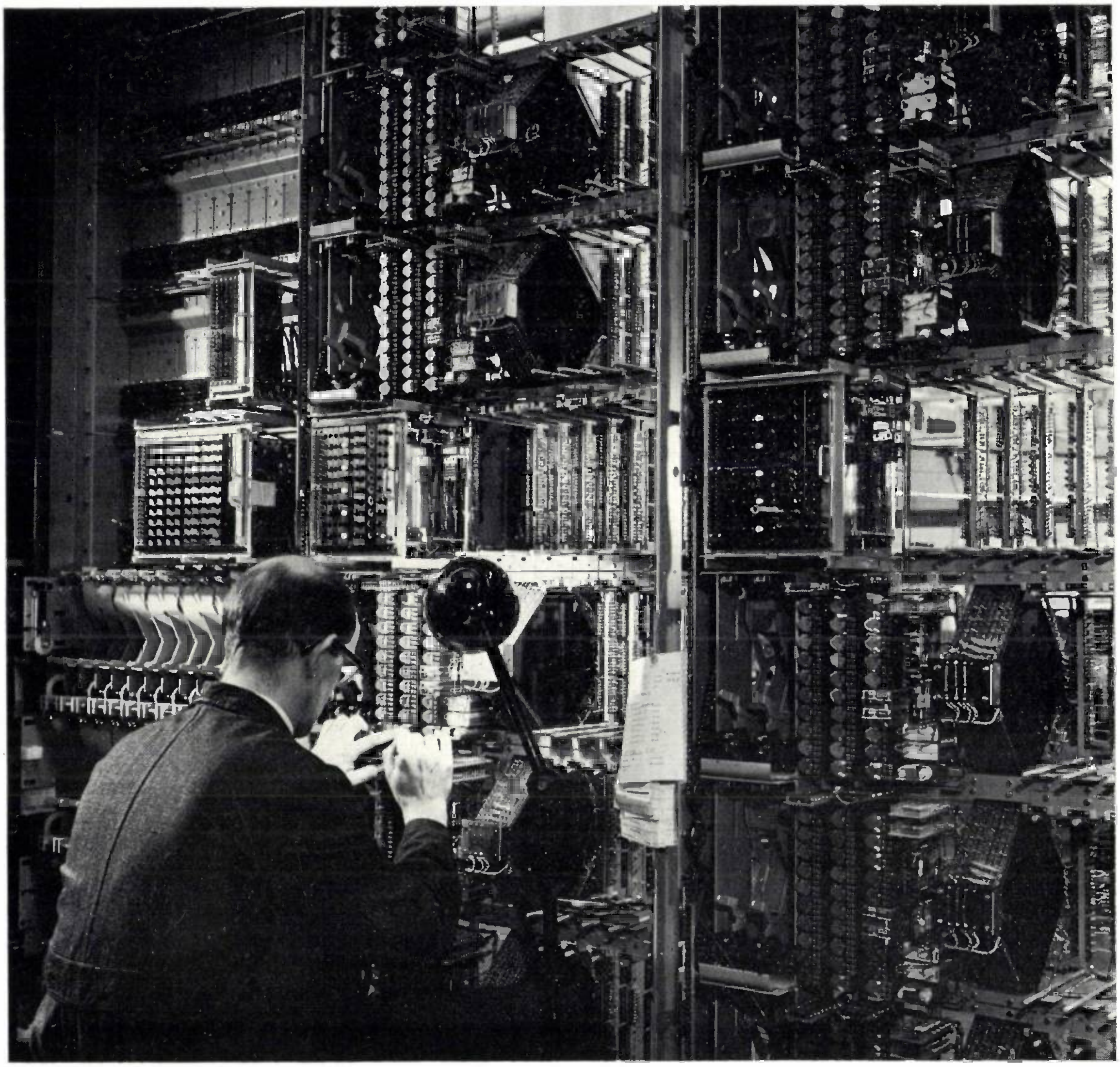
[15] 15 cm³ fuming nitric acid and 1 cm³ 50% HF (at boiling point).

[16] This graphical method, which might seem somewhat complicated, takes account of the fact that the germanium present in the gas arrives only as a minute fraction in the grown layer.

[17] Communication from J. Bloem, Philips Semiconductor Works, Nijmegen.

PHILIPS TECHNICAL REVIEW

VOLUME 26, 1965, No. 8/9



Research and production in telecommunication form the business of many departments of the Philips Group in various countries. The centre of these activities, so far as they concern actual telecommunication equipment, is N.V. Philips' Telecommunicatie Industrie, with its main establishments at Hilversum and Huizen and others at The Hague, Hoorn and Amersfoort. Work on defence systems, which are closely related to telecommunication, is chiefly concentrated in the N.V. Hollandse Signaalapparaten plant at Hengelo.

The present number of Philips Technical Review presents a collection of articles giving a survey of the work done at these establishments. Such a survey must naturally be incomplete. A more complete picture is given in the columns of the quarterly "Philips Telecommunication Review", published by N.V. Philips' Telecommunicatie Industrie. We have chosen for this number the following representative subjects: the development of carrier telephone systems; the design of automatic telegraph switching centres and an interesting aid for such centres, a magnetic tape store with static read-out; the problem of traffic control at large airports; and the design of television transmitters for the new bands IV and V. Appended to these articles is an account of the fundamental research being done at the Research Laboratories, Eindhoven, on companders for the better adaptation of speech-level variations to the telecommunication system.

These articles have been prepared for publication by Ir. F. Westerveld, former editor of "Philips Telecommunication Review", whom we were pleased to have on our editorial board for the occasion.

The reader will also find in this issue a historical introduction appropriate to the general theme, which describes some important episodes in the early days of modern telecommunication engineering.

Modern carrier telephone systems

H. N. Hansen

621.395.44

Since the last review of carrier telephone system techniques in this journal [1], many important developments have taken place, largely owing to the new possibilities opened up by the use of transistors. In the present article we shall briefly review the developments which Philips have carried out in this field.

As the number of telephone conversations carried on simultaneously via a single transmission channel — e.g. a cable pair — increases, the economic advantages of carrier telephone systems also increase. This explains the tendency to widen the available frequency band on the transmission channel further and further by raising its upper limit to progressively higher frequencies.

To make progress in this direction possible, a number of limitations of a technical nature have to be overcome. In the first place, for transmission by cable, this must be suitable for the transmission of the desired band of frequencies. In cables containing symmetrical pairs the upper frequency limit is determined by crosstalk, i.e. by the interference resulting from the inevitable inductive and capacitive couplings between the various pairs of the same cable. In cables with coaxial pairs no such limitation exists, since in this type of cable crosstalk diminishes as the frequency increases.

If, however, the cable does not limit the widening of the frequency band, the limitations inherent in the line equipment are very real. The function of this equipment is to compensate for the attenuation of the cable, and so it is chiefly made up of amplifiers. For all cables the attenuation per kilometre increases with frequency, and the higher must be the upper frequency of the band to be transmitted, the closer the spacing between successive repeaters. Spacings of the order of 5 km are nowadays by no means uncommon.

It is entirely feasible to design valve repeaters for such short spacings. Economically, however, and also from an operating point of view, they are at a disadvantage. The economic disadvantage of these repeaters lies not so much in their purchase price, for this is not so high as to have an unfavourable effect on the cost of large systems, but rather in the fact that they have to be accessible for maintenance. Valves, for instance, must be replaced at regular intervals, so that the repeaters have to be housed in simple buildings or street cabinets. Repeater spacings of 5 km soon make the cost of installation excessive in comparison with the cost of the equipment. In addition, they render the connections undesirably vulnerable.

From an operating point of view the main problem is that of supplying power to the repeaters. The requirement of absolute reliability which the connection has to meet means that there must never be any interruption of the power supply, even in the event of a mains failure. Consequently, emergency power supplies are necessary, and it would be uneconomic to repeat these at distances of a few kilometres. For systems with a limited number of channels per cable pair, where the upper frequency limit is not too high and repeater spacing still reasonably wide, the power supply stations can be sited at distances of 25 to 100 km from one another, and the intermediate repeaters can be fed over the cable pairs. This is an attractive solution but its use is, of course, limited by the fact that as the number of repeaters increases the power-carrying capacity of the cable is eventually exceeded.

In carrier systems using valve repeaters, therefore, there is an inevitable conflict between the requirements of logical technical development, leading to large numbers of repeaters spaced at short intervals along the cable route, and the requirements of economic operation, such as minimum routine maintenance and the simplest possible form of accommodation for repeaters, the ideal being to bury them in the ground, just as is done with the cable. Only the introduction of the transistor has enabled us to find a way out of this deadlock. As we shall see, the change has not been limited to the line equipment, but also involves the carrier terminals.

Introduction of transistors

The development of carrier telephone systems first began about forty years ago, when the amplifier valve had progressed beyond its initial evolutionary stage. Logically enough, the design principles followed were entirely adapted to the properties of valves. Consequently, when we came to the conclusion some twelve years ago, that further logical development of carrier telephone systems made the introduction of transistors imperative, we also realised that our designs would require fundamental revision.

The first useful property of the transistor to be

Ir. H. N. Hansen is on the staff of N.V. Philips' Telecommunicatie Industrie, Hilversum.

exploited in this redesign was its durability, which under the right conditions, is practically unlimited. In this respect it is on a par with the passive components of the repeater such as resistors, capacitors, coils, etc. Now, at last, it is possible to bury repeaters with the cable in a manner very similar to that adopted for loading coils.

The transistor has other welcome properties: it produces very little heat — for one thing, it has no heater supply — and needs only a very low supply voltage, e.g. 10 V. Its very low heat dissipation makes a very compact repeater design possible without any risk of excessively high temperatures. Its low supply voltage not only lengthens the useful life of the other components; in combination with the very low power consumption of the transistor, it also enables the power for the amplifiers to be brought over the cable in a very simple manner. Such power supply equipment as is needed can be buried with the amplifiers and is reduced to a bare minimum.

Until now we have discussed only the effects of the introduction of transistors on line equipment, because the results obtained have been most striking in this sector. We shall now dwell briefly on their effects on terminal equipment.

In the terminal equipment a modulation process of some complexity is used to transpose the band of voice frequencies from its natural position to the required position in the frequency band of the carrier system, and vice versa. In contrast with the line equipment, where there is only collective channel amplification, the terminal equipment gives much more individual treatment to every channel. As the name implies, terminal equipment is normally found at the ends of a connection, in the repeater stations, which in larger towns, are often equipped for many thousands of channels. Large as these numbers already are, they show a regular growth of 7 to 15 per cent a year and this readily explains the present tendency to keep apparatus dimensions down to a minimum in order to make maximum use of the available accommodation.

In valve equipment heat dissipation very soon puts a limit to reduction of size. Continued reduction in size would eventually give rise to excessive temperatures, not only inside the amplifier units themselves, but also in the space where the equipment for so many channels has to be housed. Here again the transistor, with its small size and low dissipation, permits further reduction of equipment dimensions. In addition, the reliability of the equipment is enhanced, thus reducing the periodic maintenance needed by each channel and helping to solve the problem created by the continuous growth of repeater stations and the increasing scarcity of maintenance personnel.

New carrier systems

We have already stated that the introduction of transistors necessitates thorough redesign of carrier equipment. When the valve is replaced by the transistor, all the other components also need replacement. Carrier telephone equipment has to meet extremely stringent requirements of reliability and no system can be put into regular service before all the teething troubles have been overcome. This means that all new components have to be thoroughly tested in this respect and, as a result, the transistorization of carrier equipment calls for intensive and time-consuming development. After twelve years of development work there is now available a complete "family" of transistorized carrier systems of considerably reduced dimensions.

We have been speaking of a family, because a harmonious production programme for carrier telephone equipment should comprise systems with channel capacities ranging from twelve to several thousands. Every manufacturer naturally attempst to form these systems from a minimum number of equipment units.

Carrier telephony is used not only in transmission systems using metallic conductors, but also in radio links. As a result of international co-operation between telephone administrations and manufacturers, standards have been established for the entire field, and systems with greatly varying numbers of channels can now be assembled from a minimum number of different apparatus units. The advantages of such an arrangement will be obvious when it is realized that the telephone connections at the junction points of a network may make use of transmission media of all descriptions, such as symmetrical or coaxial cables, open-wire lines or radio links. Interconnection of carrier equipment for all these media must be possible without difficulty.

General design details

The standardization rules to which we have just referred apply mainly to the external electrical characteristics of carrier systems. In the matter of circuits and of mechanical design — apart from some main dimensions — manufacturers have a free hand, and their designs therefore differ mainly in these respects.

Philips have developed their own and very distinctive design, in which all electrical components forming a functional entity, such as an amplifier, are combined into a hermetically sealed unit. This procedure has come to be known as the "conclave" technique. Its

[1] H. N. Hansen and H. Feiner, Coaxial cable as a transmission medium for carrier telephony, Philips tech. Rev. **14**, 141-150, 1952/53.

principles were first introduced in 1950, when a new version of the carrier current system with valves was being designed. Then as now, the designers were aiming at low first cost, small dimensions and minimum maintenance.

This method of sealing off components ensures that all harmful external influences, such as those of moisture and dust, are excluded. Many units contain filters whose frequency responses show extremely steep flanks that cannot be maintained unless the characteristics of the components are kept within very narrow tolerances. If the entire unit is not sealed off as in the conclave technique, it is usually necessary to use the much more complex method of soldering the filters into individual containers. In a conclave unit all components are protected, while accessibility leaves nothing to be desired. *Fig. 1* shows a conclave channel unit made in 1954. Since then many tens of thousands of these units have been manufactured.

In 1953 application of the conclave principle was still restricted to a single carrier telephone system. Experience with it proved so favourable that the principle has since been applied to all Philips transmission equipment. The key to its success lay in the simplicity of the method of sealing which gave all the advantages of a hermetically closed unit without adding to its cost. The principle is now applied quite generally, even where climatic conditions do not strictly require its use, because it is an advantage that each unit should, as it were, carry its own packing with it, thereby simplifying the storage of spare units.

Fig. 1 also shows that, with the method used in 1954, components were mounted on pins anchored in SRBP sheets or strips. This method permitted partial mechanisation of production and had the great advantage of making the wiring highly reproducible. Errors in assembly and damage to components were also reduced. This technique, now no longer used, remained in continuous use until a printed wiring technique had been developed which was considered sufficiently reliable to be applied to carrier systems.

It will come as no surprise to the reader to learn that, since changing over to transistor techniques, we have kept to the conclave principle, even though this meant completely redesigning all mechanical details. On investigating whether or not hermetic sealing was still the most suitable method, we found it to be even more suitable than before. This is due on the one hand, to the very low heat dissipation of transistors, which makes their drying action on surrounding components almost negligible; on the other hand, the tracks of a printed wiring board lie so close together that very effective protection against dust and moisture is essential.

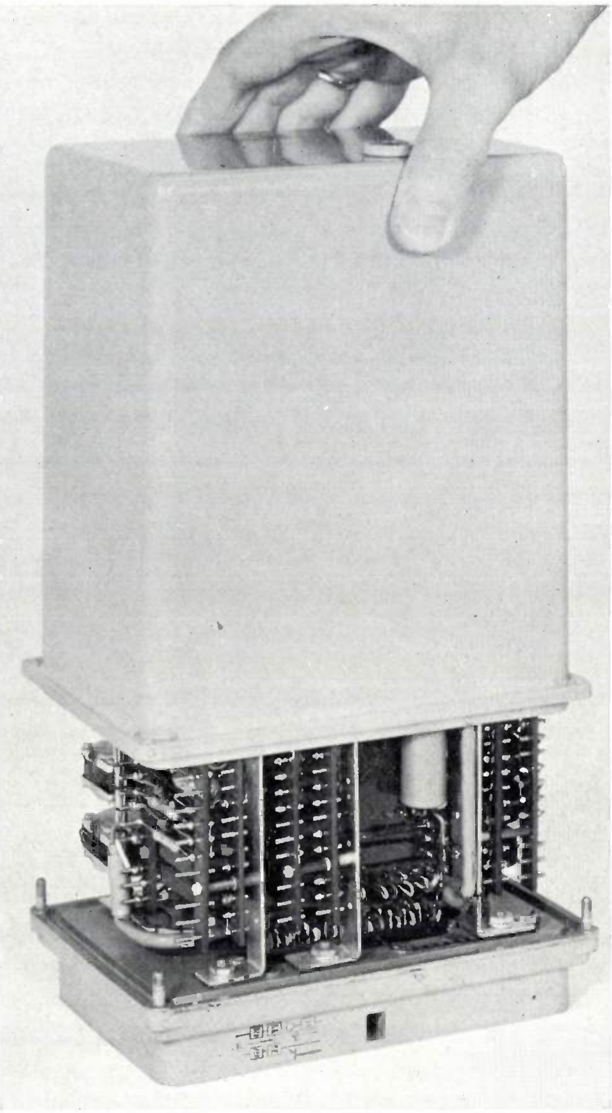


Fig. 1. Channel unit, fitted with valves, as manufactured in 1954. The unit contains the first stage of modulation and the last stage of demodulation for a single channel. A rubber gasket makes an airtight joint between cover and base plate, so that all components, including the valves, are completely sealed off from the outside.

Without going into details of the operation of the equipment, we shall now give a number of examples to bring out the features of our technique.

Line equipment

We have already mentioned that, in line equipment in the first place, the introduction of transistors led to an important advance by making it possible to use buried repeaters. This can best be illustrated by the concrete example of carrier systems for paper-insulated symmetrical cables. Systems then in current use permitted a maximum of 48 or 60 channels per symmetrical pair; the maximum frequencies in use were 204 and 252 kc/s and repeater spacings varied between 15 and 25 km.

Cables of the type mentioned are not very suitable for the transportation of the power needed by valve repeaters. For this reason each repeater station is provided with emergency power supply equipment which comes into action upon a mains failure. However, the Dutch PTT administration had found [2] that relatively simple means sufficed to improve the crosstalk properties of these cables to a point where frequencies of some 550 kc/s could be used. At such a maximum frequency the channel capacity of a cable pair can be raised to 120, i.e. the existing capacity can be at least doubled. If valve amplifiers continued to be used, repeater spacings would have to be halved and the number of emergency power plants doubled. As we have pointed out already, this would be unacceptable from the point of view of both economy and operating conditions. Only with transistorized amplifiers is it possible to take full advantage of this widening of the frequency band.

Once this has been decided, the next problem is to determine the optimum repeater spacing. With valve amplifiers the tendency has always been to make this spacing as wide as possible with the available valves, since the number of repeater stations, which not only were vulnerable but also required maintenance, was thereby reduced to a minimum. With buried transistorized repeaters this is, of course, no problem and other considerations apply. All one has to decide is the most economical solution to the transmission problem as a whole bearing in mind also the surface stations where the power supply and automatic gain control equipment are installed. Looking at the problem in this way, we find that the best possible use of the properties of the transistor is made by designing the system in such a way that the repeater output level is kept well below the available output power of the transistor, rather than by aiming at maximum repeater spacings. This then means that small conservatively rated amplifiers are used at short distances apart.

This principle of fairly short repeater sections is valid, not only for the systems of 120 channels per pair on symmetrical cables that we have just mentioned, but also for the coaxial cable systems that will be discussed below. Its value can be demonstrated by examining the power supply problem, for example. At a frequency of 550 kc/s the symmetrical cables that are generally used in the Netherlands have an attenuation of 4.5 dB per km. Every 700 metres added to the length of a repeater section therefore means that the available output power of the amplifiers has to be doubled. For a repeater spacing of 8 km, 25 mW of output power would be necessary; if that distance were raised to 12 km, i.e. increased by 50 %, the output power would have to be increased to 1 W, i.e. multiplied by 40.

For the types of amplifier used in practice, the power drain is very nearly proportional to the output power of the last stage, so that 12 km repeater spacing requires 40 times as much power per repeater as 8 km spacing; per km of cable route the multiplication factor is $8/12 \times 40 \approx 25$. Ultimately, of course, it is only the absolute value of the supply power that is important, but the example clearly shows the great influence of repeater spacing on the question of the design of the line equipment.

These and other considerations have led to the choice of an 8 km repeater spacing for the 120-channel system. Since the existing surface repeater stations are spaced about 25 km apart, two underground repeaters have to be added between each existing pair. We may add that the power drain of each repeater is only 0.2 W; the total power dissipation of the 24 repeaters installed in one underground case is no more than 5 W. *Fig. 2* shows the case used to house the 24 repeaters, with the equipment required for patching service trunks, programme channels, etc. These cases are installed in concrete pits; for each cable such a case is the equivalent of a repeater station [3].

The application of this technique, which is the result of very close collaboration between the Netherlands PTT administration and N.V. Philips' Telecommunicatie Industrie, thus enables the traffic-carrying capacity of an existing route to be at least doubled without necessitating the construction of new repeater stations or the laying of new cable. As the cable usually represents the highest single item of capital expenditure on the connection, there is no doubt that the method just described is a very economical one. The cables used in the network of carrier telephone circuits in the Netherlands are almost exclusively of a type that lends itself to the application of this technique. The capacity of the network is now in the process of being multiplied by a factor of 2.5 by a change-over from the 48- to the 120-channel system [4].

[2] L. J. E. Kolk, Crosstalk problems in balanced carrier cables at frequencies up to 552 kc/s, Philips Telecomm. Rev. **23**, 167-178, 1962.

[3] J. Metz, B. H. Wijnen and A. Timmer, Underground cases containing repeaters for multiquad carrier cables, Philips Telecomm. Rev. **23**, 179-185, 1962.

[4] D. van den Berg and A. P. Bolle, Recent and future developments in the field of line transmission, Het PTT-bedrijf **10**, 172-178, 1960/61;
G. H. Bast, Widening the frequency band of carrier cables, Philips Telecomm. Rev. **23**, 100-102, 1962;
A. P. Bolle, Sie Swan An, J. H. Duimelaar and J. F. Lansu, Transistorized line equipment for a 120-channel carrier telephone system, Philips Telecomm. Rev. **23**, 103-121, 1962.

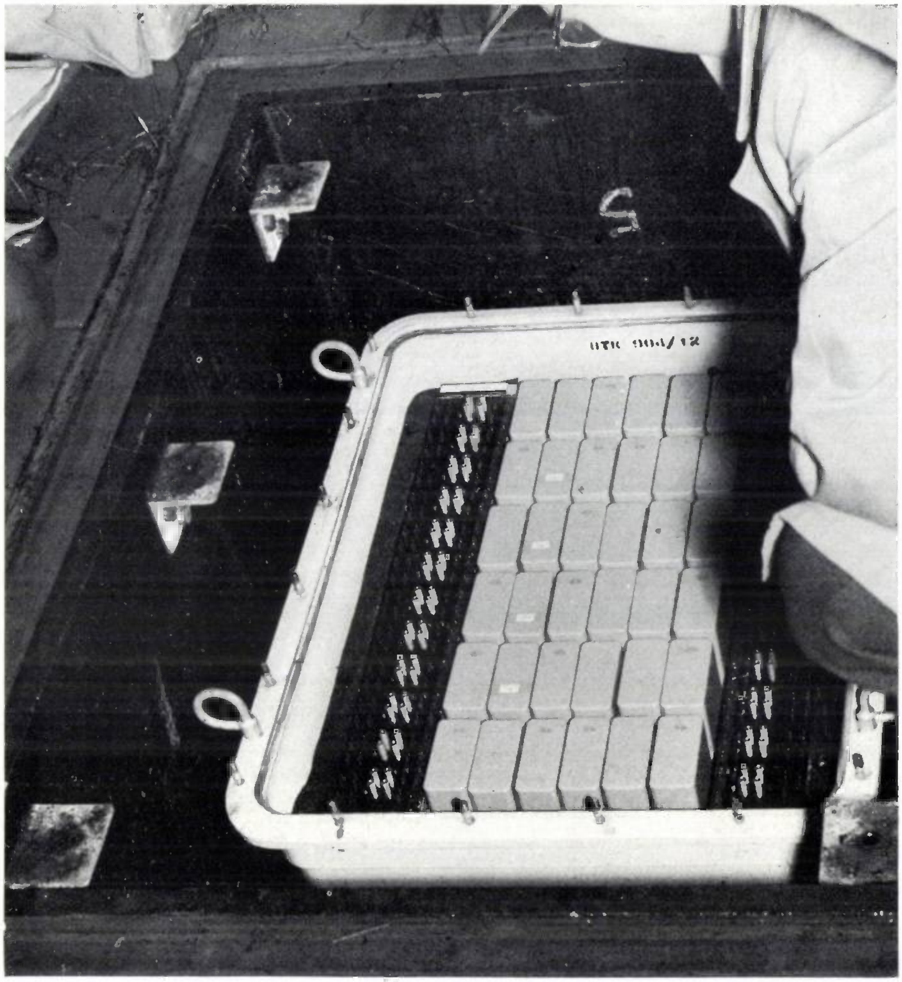


Fig. 2. Transistorized line amplifiers can be safely installed in underground cases. The case shown holds 24 amplifiers, for use with a 24-pair carrier cable.

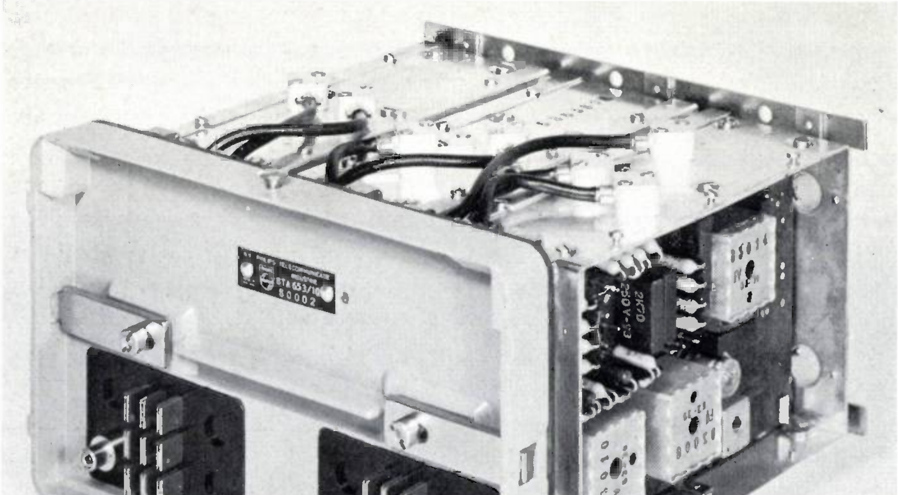


Fig. 3. Supermastergroup modulator in conclave technique, with the protective cover and the lid of one of the silver-plated brass boxes enclosing the components removed. These boxes are a continuation of the outer conductor of the coaxial cable.

Terminal equipment, with special reference to coaxial systems

Each individual telephone channel is brought to its final position in the frequency band by a process of modulation in several consecutive steps; during this process the channels are brought together in groups that increase in size with each step and are manipulated as single units. The following groups are distinguished: 12-channelgroups; supergroups of 60 channels or five 12-channelgroups; mastergroups of 300 channels or five supergroups; and supermastergroups of 900 channels or three mastergroups.

This system of modulation in stages is necessary, not only for technical reasons determined by filter design considerations, but also to keep to a minimum the number of types of unit needed in a system. Likewise, it reduces the number of unit types required to build up a whole family of systems such as we have already mentioned.

The CCIR (Comité Consultatif International de Radiocommunications) and the CCITT (Comité Consultatif International pour la Télégraphie et la Téléphonie) have, between them, standardized the frequency bands for carrier telephone systems, and these are now identical for cable and for radio transmission systems. Input levels have also been standardized. As a result, a given type of terminal equipment can be used for transmission by cables or by radio links as required and equipments of different makes may be used together without difficulty. It is fortunate that the channel capacities for carrier systems on cables and on radio links have developed on largely parallel lines.

We have already mentioned that for each channel of a carrier terminal there is an individual unit, the channel unit. These units are therefore very numerous; the Philips modulation procedure never requires them to handle any frequency higher than 36 kc/s. At the other end of the scale we have the supermastergroup modulator. This modulates a group of 900 channels and therefore occurs only once for 900 channels per supermastergroup. The band of frequencies it transposes has a width of 3.8 Mc/s and the transposition takes place in the range from 300 kc/s to 12.4 Mc/s. There is obviously a great difference between the electrical requirements to be met by these two units but the conclave design mentioned above has been found to suitable for both. Its compactness is in fact extremely suitable for the high frequencies mentioned. Externally the supermastergroup modulator can be distinguished from the channel unit by its connecting block, on which there are a number of coaxial connectors. As can be seen from *fig. 3*, the printed-wiring boards and their components are fitted in silver-plated brass boxes. These boxes should be regarded as the logical extension of the structure of the coaxial cable pair. They effectively suppress crosstalk which, at these high frequencies, would be liable to occur in a less systematic type of design.

The fact that the same design can be used for both high and low frequencies results in a uniformity of appearance which is not only aesthetically pleasing, but which has been found to be very convenient during installation, for maintenance, and when expansion becomes necessary.

The channel unit

The first stage of modulation and the last stage of demodulation in the chain that makes up the entire modulation procedure, are situated in the channel unit. Since these units are employed in such large numbers, very close attention to every detail of their design is necessary. *Fig. 4* shows the latest design of channel unit, with the cover partly removed. Since 1946 the volume of the channel unit has been greatly reduced, as *fig. 5* demonstrates. In 1946 one side of a modern carrier system rack would take 12 channel

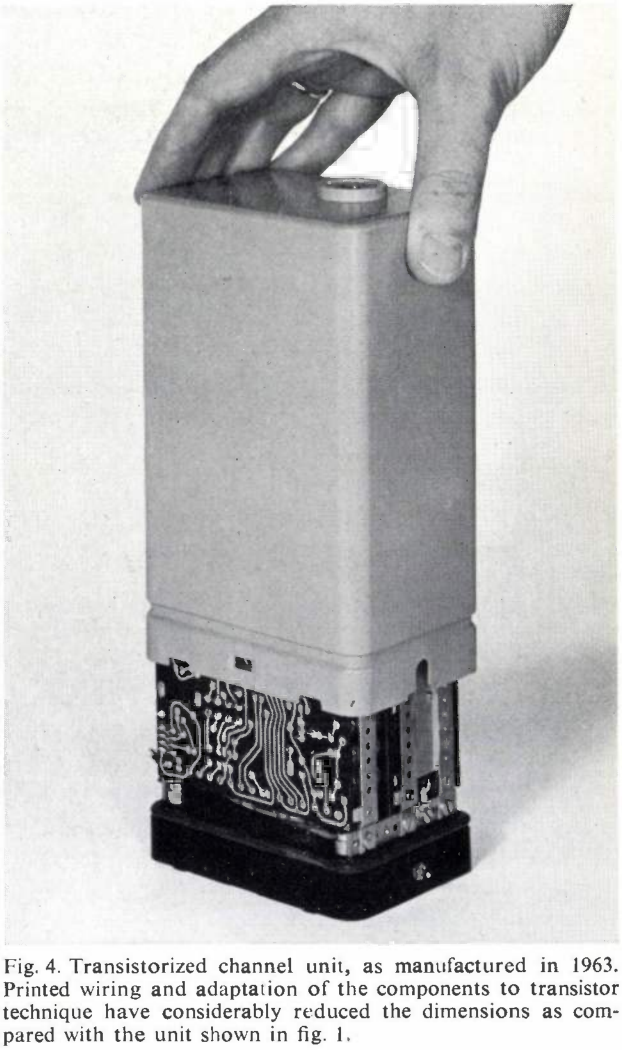


Fig. 4. Transistorized channel unit, as manufactured in 1963. Printed wiring and adaptation of the components to transistor technique have considerably reduced the dimensions as compared with the unit shown in fig. 1.

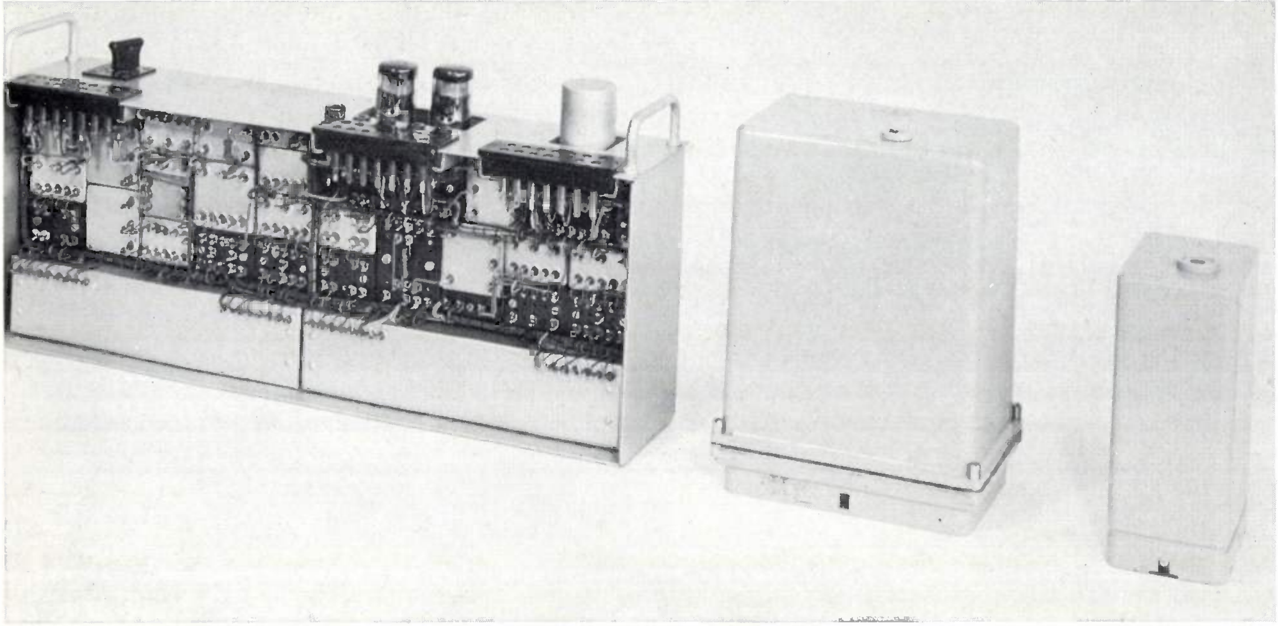


Fig. 5. Channel units dating from 1946, 1954 and 1963, respectively. The first two are equipped with valves, the third with transistors. In the 1946 unit the open method of assembly was still in use, while the last two units follow conclave principles.

units, but this number has now gone up to 96 or 120.

One reason why such a compact design as that of fig. 4 has become possible is that almost all components have been given the same uniform height. Without creating any assembly difficulties, this principle permits very efficient use of the available space. The type of printed wiring that has been specially developed for our carrier telephone systems has contributed greatly to their increased reliability. Correct connection between components is automatically made and the percentage of defective soldered joints is much smaller than can be attained on even the best of assembly lines.

Two channel modulation racks are shown in *fig. 6*, each with a full complement of 96 channel units and some auxiliary equipment units, such as a measuring panel and a few units for stabilized voltage supplies.

Systems for coaxial cables

As we have already pointed out, the crosstalk properties of coaxial cables improve as the frequency increases and the number of channels per pair is therefore limited only by the amplifiers. Ten years ago the recommendations of the CCIF (now the CCITT) were concerned only with carrier telephone systems having a maximum of 960 channels on a standardized type of coaxial cable. Nominal repeater spacing for these systems was 9.6 km. Later the recommendations were extended to a 2700-channel system with 4.8 km repeater spacing. Such short spacings are barely acceptable with valve repeaters, mainly because, as we have said, it is technically very difficult to feed the repeaters via the cable. However, the economic way in which they can be operated makes these 2700-channel systems very attractive for main traffic arteries between large cities.

Transistorization of line amplifiers has also opened up new prospects for coaxial cables. By putting most of the repeater stations into underground cases, it is possible to reduce the price per kilometre of the line equipment to less than half and to achieve a notable reduction in the total cost of transmission.

This important change in the relative costs of the cable and of the line equipment also justified the development of a second type of coaxial cable pair, in order to achieve minimum transmission costs per kilometre. The "small" coaxial cable that resulted has "pipes" whose outer conductor has an inner diameter of 4.4 mm, and whose attenuation is 2.2 times as high as that of the standard coaxial cable with its corresponding diameter of 9.5 mm. Transistorized carrier telephone systems for 300 channels are already in operation on the new type of cable.

In the meantime the CCITT has worked out recommendations for 1200-channel systems on the new cable. Philips are now developing a system of this kind which will shortly be put into operation on an experimental route. In determining the repeater spacings the same considerations have prevailed that we mentioned in connection with the 120-channel system, i.e. the aim has been to arrive, not at maximum spacing, but at the optimum overall solution.

This brings us to a problem to which no reference has so far been made, viz the problem of automatic gain control. The reader will have no difficulty in realizing the importance of this problem if he considers the enormous attenuations for which the amplifiers have to compensate. At an attenuation of 10 dB per km of cable, the overall loss on a route of 100 km length is 1000 dB, which corresponds to a power ratio of 1 in 10^{100}. If we assume that this route has 33 repeaters with a gain of 30 dB each and spaced 3 km apart, even a systematic error per repeater of only 0.5 dB will result in a deviation of 15 dB in the overall gain. In addition, the cable attenuation has a temperature coefficient of 2.10^{-3} per degree Celsius. Temperature variations of $\pm$ 10 °C in the course of a year are a normal occurrence, and if appropriate steps were not taken, they would cause an attenuation variation of $\pm$ 20 dB on a 100 km route.

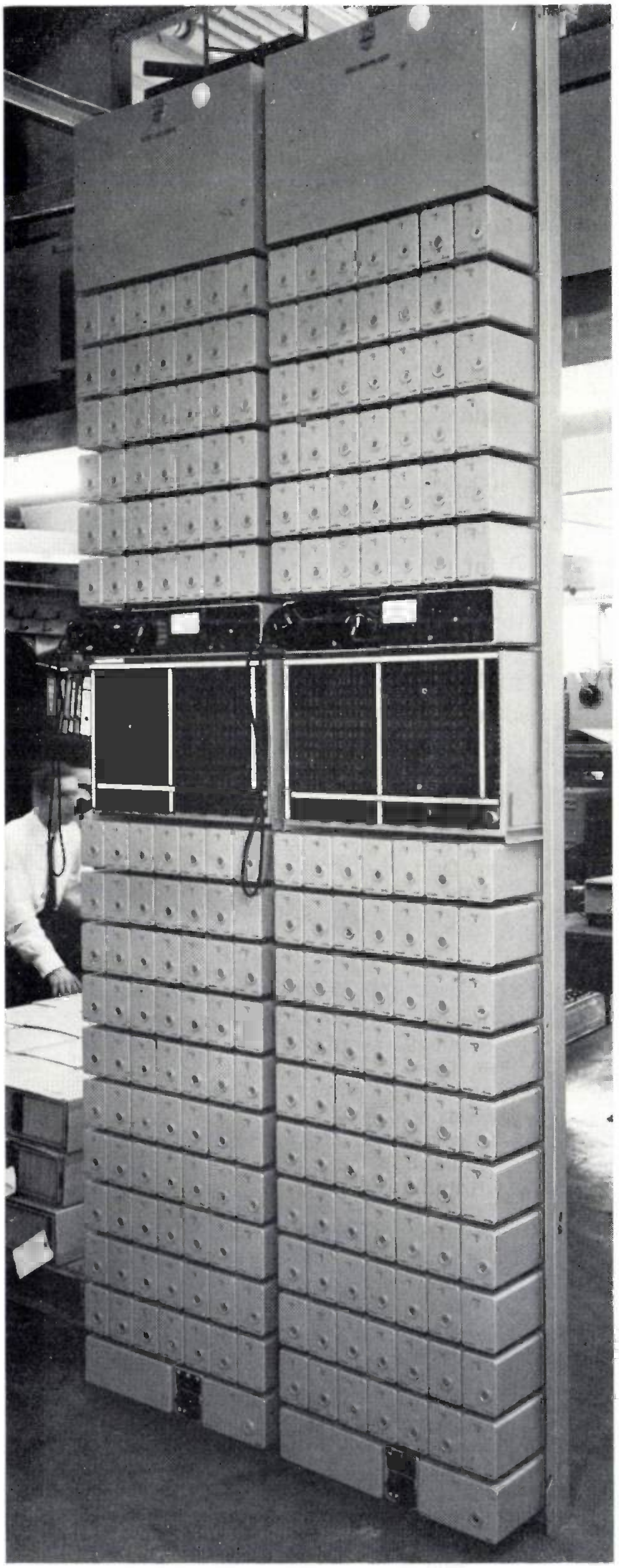


Fig. 6. Two channel modulation racks, mounting 96 transistorized channel units each. A measuring panel and some power supply units complete the rack.

If, in addition, we consider that CCITT standards require a carrier telephone system to be so designed that it will bridge distances of up to 2500 km, we shall appreciate the stringency of the requirements that the reproducibility and the stability of the repeaters have to meet. On top of all this, measures must be taken to compensate for the variations in cable attenuation and in repeater gain. The sum total of these measures is referred to collectively as automatic gain control.

Returning to our problem of determining the most suitable repeater spacing, we must, of course, take automatic gain control into account. Leaving out technical details, we may state that for 1200-channel systems on small coaxial cable the optimum repeater spacing was found to be 3 km. For this system the distance between two surface repeater stations supplying intermediate underground repeaters can be up to 100 km. This arrangement permits an entirely new solution to the problem of automatic gain control: it means that complicated equipment can be limited entirely to the surface stations and that the design of the underground stations can be greatly simplified. Because the attenuation between repeaters is kept low, a certain margin remains available for automatic control and the number of controlled repeaters per route can be reduced. The solution adopted is independent of the type of coaxial cable and also of the channel capacity of the system. It has brought a fundamental improvement to coaxial techniques, which had hitherto not given a completely satisfactory answer to the question of automatic gain control. The transistor can be credited with having made the new solution possible. Following the above, it will come as no surprise that the development of a 2700-channel system using 9.5 mm coaxial cable has now also been initiated.

Final observations

Will it be possible or desirable, one may ask, to operate carrier systems with even more than 2700 channels per pair of conductors? So far, technical possibilities have fixed the upper limit. The maximum of 2700 channels for valve systems was determined by the minimum repeater spacing that appeared acceptable in practice. Basically, transistor systems require somewhat smaller spacings than valve systems, although here the spacing problem is of no more than secondary importance. There is no doubt that improvement in transistor techniques will make further extension of the bandwidth of amplifiers possible and we may confidently expect that 10 000-channel systems will become technically feasible in a not-too-distant future. Traffic requirements will ultimately be the only limit. As the demand for long-distance telephone circuits goes up 10-20 % each year, and the transmission of television programmes and data will require a great deal of additional traffic capacity, the upper limit of the frequency band may be expected to be raised very considerably. Some thought has already been given to the possibilities of using waveguides, since these offer the basic possibility of transmitting several hundreds of thousands of channels per waveguide. So far, no means has been found of overcoming the practical difficulties in the application of this technique, but for the time being there is still sufficient latitude in the possibilities offered by the coaxial cable and the radio link.

Besides the total traffic demand, the distribution of traffic is of great importance. Geographical conditions may have considerable influence on the traffic distribution. In countries such as the Netherlands and Belgium these are such as to lead to a fairly close-meshed telephone network with traffic groups of moderate size. In such a configuration the need for systems with very large numbers of channels per pair of conductors will not be felt so readily as in a country like Denmark, where the main network is more nearly star-shaped.

We have just mentioned the possibility of using coaxial cables for data transmission and television programmes as well as telephone traffic. Of the two, data transmission is still in the initial stages of development, but this cannot be said of television transmission. Until now, the coaxial cable has been relatively little used for television transmission. This is due partly to the fact that transmission of television programmes by cable creates some special problems that have only recently been completely solved, and partly to the fact that working out a coaxial cable route and putting it into operation take much more time than is required for a radio link system. As the coaxial cable network continues to expand, however, a point will be reached where combined operation will become profitable for both television and telephone traffic. If, as a result of this combination, more "pipes" are required in a cable, this will improve the profitability of its operation still further. To make such a combination possible, the line equipment must of course be suitable for both types of transmission. This necessity has been fully reckoned with in the design of the 1200-channel system for the small type of coaxial cable we have already mentioned.

In concluding our final observations we would draw the reader's attention to carrier telephone systems for very short distances. The systems previously discussed are economical only for distances upwards of some tens of kilometres; the cost of the terminal equipment, averaged per km of cable, would otherwise be prohibitive. However, by far the greatest number of telephone conversations are conducted over distances of less than 20 km and a carrier system that would be competitive with an audio connection at distances of 10 km or so has been sought after for many years. Such a system would appear to require a different modulation method from that now in use. In the present state of the art pulse modulation systems such as pulse code modulation or delta modulation [5] seem to hold most promise. Although systems of this kind are already in operation in a few places, mainly in the U.S.A., it cannot yet be said that the technique has advanced far enough to make its general application economically justified. It may be that new technologies, permitting complete circuits with all their components to be manufactured as a single compact assembly, will bring the solution.

[5] J. F. Schouten, F. de Jager and J. A. Greefkes, Delta modulation, a new modulation system for telecommunication, Philips tech. Rev. **13**, 237-245, 1951/52;
F. de Jager, Delta modulation, a method of PCM transmission using the 1-unit code, Philips Res. Repts. **7**, 442-466, 1952.

Summary. When valves in a carrier telephone system are replaced by transistors, all the components require adaptation to the change. Upon transistorization, Philips carrier systems were redesigned completely, but the conclave principle of construction has been retained. This technique, originally developed for valve systems, consists in enclosing all components of a functional unit in an airtight container.

Upon transistorization, carrier equipment becomes more compact. In addition, transistors have practically unlimited life and their characteristics vary little with time: as a result they meet the requirement of reduced maintenance. The advantages of transistor application are most apparent in line equipment for symmetrical and coaxial cables. Transistors are so reliable that the amplifiers can be laid underground with the cable. Their low power consumption and the low supply voltage required for these amplifiers make it possible fo feed them via the cable and to concentrate the emergency power equipment in a few surface stations. It thereby becomes technically feasible to space the amplifiers very closely and so compensate for the very high attenuations that result when the upper frequency limit of the carrier system is raised higher and higher in order to accommodate the necessary number of channels per cable pair. The low cost of transistorized repeaters makes it economic to keep their gain at a lower value than in valve systems. Some of the margin gained in this way has been used to improve and simplify the automatic gain control circuits.

Companders with a high degree of compression of speech level variations

J. A. Greefkes, P. J. van Gerwen and F. de Jager

621.395.665.1

There is no technical difficulty nowadays in connecting any two points in the world by telephone. A wide variety of transmission systems is available for this purpose. They can be classified under the two main headings of radio and cable transmission systems.

In the microphone the acoustic oscillations generated by the vocal cords are converted into electrical oscillations. Owing to the natural dynamics of the spoken word, and also because many speakers do not maintain a constant distance from the microphone, the amplitude of these electrical oscillations can vary between wide limits. In addition, the attenuation of the lines connecting individual telephone stations to their exchange, shows considerable variation. The result is that the speech level on transmission circuits will obviously be anything but constant.

Large variations of the speech level may have a very unfavourable influence on the quality of a telephone connexion. On radio circuits, for example, the noise level may be high at times and when this happens, a low speech level can soon reduce intelligibility to zero. On the other hand, unusually high speech levels can cause operation of the limiter that is normally connected to the input of a radio transmitter for its protection, and this can then cause serious distortion. In multi-channel carrier telephone systems high speech levels can give rise to crosstalk between the various channels. Crosstalk may be either intelligible or unintelligible; in the latter case it gives the effect of an increase in noise level.

Even from the above summary examination it will be obvious that means have been sought to reduce the variations in level on telephone circuits. These include the use of companders. This word is a contraction of the words compressor and expander and indicates a combination of two electrical circuits, one of which is connected to the input of a transmission circuit in order to reduce variations in level, i.e. to compress them, whilst the other is connected to the output in order to restore these variations to their original values, i.e. to expand them.

One of the reasons why the compressor is followed by an expander is that speech sounds more natural if its original dynamics are restored. In addition, the stability of telephone circuits that include four-wire sections terminated by hybrids requires the amplification in each of the branches of the four-wire circuit to be as constant as possible. In essence, however, a compressor is a circuit with variable amplification and its use would inevitably cause instability if it were not followed by a circuit whose amplification varies in a complementary sense, i.e. by an expander.

In our survey all companders will be described as if they worked in the band of natural speech frequencies. This method has been followed because it greatly facilitates comparison of the various principles involved. On the other hand, it precludes the possibility of comparing the practical value of the various solutions. In actual fact, the specific advantages of some of the solutions discussed become apparent only if they are applied to a speech signal which has been shifted upwards in frequency. This is to some extent the case with, for example, the pilot compander by Ensink and Verhagen which we shall describe presently and which is already familiar from patents. Space, however, does not permit us to enlarge on the specific advantages of this compander when used with a transposed speech signal.

Compander properties in general

Both the compressor and the expander show a nonlinear relation between their input and output voltages. The most desirable relation between the input voltage x and the output voltage y for a compressor may in many cases take the form $y = \sqrt[n]{x}$, where $n > 1$. *Fig. 1* shows those sections of the curves $y = \sqrt{x}$ and $y = \sqrt[3]{x}$ that lie in the first quadrant, together with the linear relation $y = x$. The graph clearly shows that for low values of the input voltage the output voltage exceeds that for the linear relation, whilst for high values the reverse is true. The higher the value of n, the more pronounced this effect becomes.

If, for a curve of the type $y = \sqrt[n]{x}$, y_1 and y_2 are the output voltages corresponding to the input voltages x_1 and x_2, then $\log y_2 - \log y_1 = (\log x_2 - \log x_1)/n$. In other words, the variations in output level, expressed in decibels, are always smaller than the variations in input level by a factor n. Compressors for which $n = 2$ are therefore generally called factor-2 compressors. For practical reasons compressors in actual use hitherto have not been made with n larger than 2, but it is obvious that higher values would make the compander much more successful in reducing the effects of variations in level.

J. A. Greefkes, P. J. van Gerwen and Ir. F. de Jager are research workers at Philips Research Laboratories, Eindhoven.

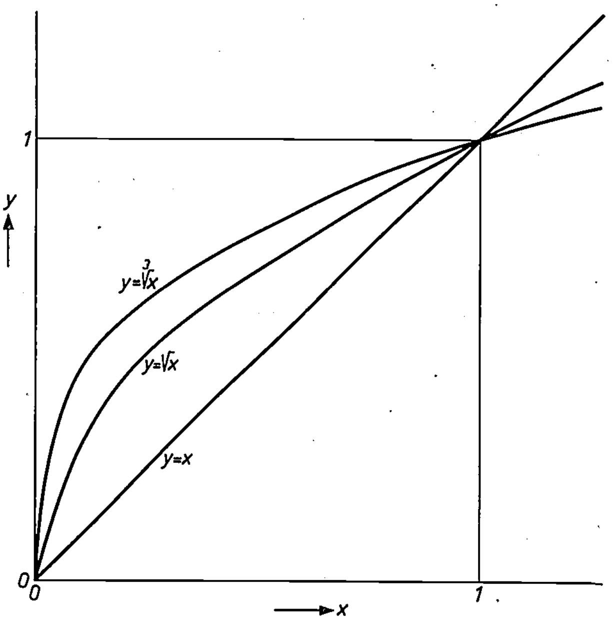


Fig. 1. Comparison of two compression curves of the type $y = \sqrt[n]{x}$ with the linear characteristic $y = x$.

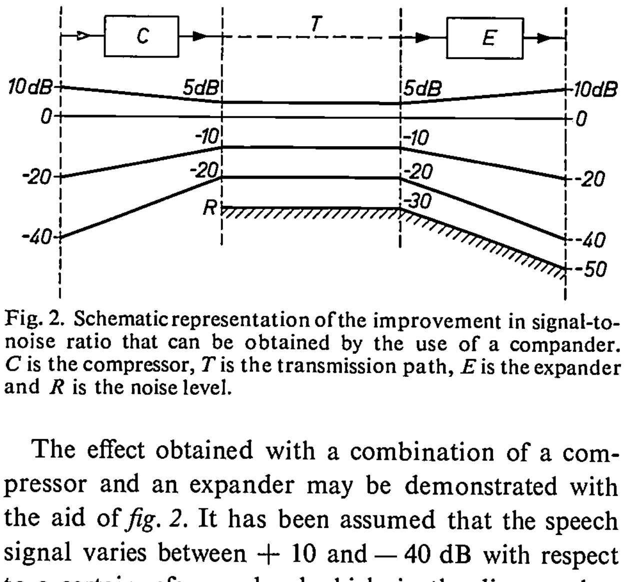


Fig. 2. Schematic representation of the improvement in signal-to-noise ratio that can be obtained by the use of a compander. *C* is the compressor, *T* is the transmission path, *E* is the expander and *R* is the noise level.

The effect obtained with a combination of a compressor and an expander may be demonstrated with the aid of *fig. 2*. It has been assumed that the speech signal varies between + 10 and — 40 dB with respect to a certain reference level which, in the diagram, has been taken as zero level. At zero level, both the compressor and the expander have zero attenuation (or zero gain), and no change in level occurs. For other levels, however, the compressor, which is assumed to be a factor-2 compressor, reduces the variations to half their original values.

A total variation of 50 dB at the input is usually the maximum for which a compressor can be guaranteed to function properly. The compressor then reduces such a variation to 25 dB. At the top end the maximum level becomes + 5 dB, which is of importance where distortion or crosstalk have to be avoided, while at the lower end the level is raised from — 40 to — 20 dB, and thus brought 10 dB above the noise level of — 30 dB of the transmission circuit.

The original speech levels are reconstituted in the expander *E*. As long as there is no speech, the noise level will be reduced in the manner indicated. In the presence of speech the gain of the expander is determined by the volume of the speech signal and the relative levels of speech and noise remain unaltered. However, the difference in level between speech and noise has already been improved by the compressor, so that the noise will not be unduly noticeable at the receiving end. At moments of "silence" the expander reduces the noise level still further, so that the listener has the impression that the circuit is extremely quiet.

Had a factor-3 compressor been used the lowest input level of —40 dB would have been raised to —13.3 dB at the output, or 16.7 dB above the noise level instead of 10 dB.

Level control in both the compressor and the expander is effected by applying a control signal to a circuit called a variolosser on account of its variable attenuation. The control signal is obtained by tapping some of the speech power at a suitable point of the circuit and then amplifying, rectifying and smoothing it. Before it is used as a control signal, the rectified signal is passed through a low-pass filter. The higher the cut-off frequency of this filter, the faster the control circuit will be able to follow the variations in level of the speech signal.

Level control in the compressor produces harmonics of the speech frequencies and though the expander eliminates these harmonics, they nevertheless have to pass through the transmission circuit. The choice of the cut-off frequency of the low-pass filter just mentioned is therefore important. With a high cut-off frequency, i.e. with a fast compander, the total band of frequencies to be transmitted may become considerably wider than that occupied by the original speech signal. In multi-channel carrier telephone systems, where the bandwidth available to each channel is limited, a slow type of compander will be selected and the cut-off frequency situated below the lowest speech frequency to be transmitted. If, however, ample frequency space is available, as in pulse code modulation systems, a fast type of control may lead to simplification of the equipment [1].

Compander system design generally aims at distortion-free transmission. This aim, however, limits the extent to which interference can be eliminated. If the requirement of distortion-free transmission is waived, as in the Frena system [2], a still higher degree of suppression is possible.

It is an inevitable corollary to the function of the expander — of increasing variations in level — that it will also increase those spontaneous variations in level due to variable attenuation in the transmission

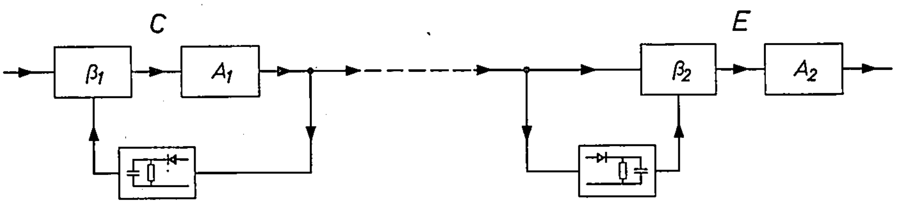


Fig. 3. Circuit diagrams of a compressor C and an expander E as used in practice. The characteristics of the two variable attenuators β_1 and β_2 must be complementary to one another.

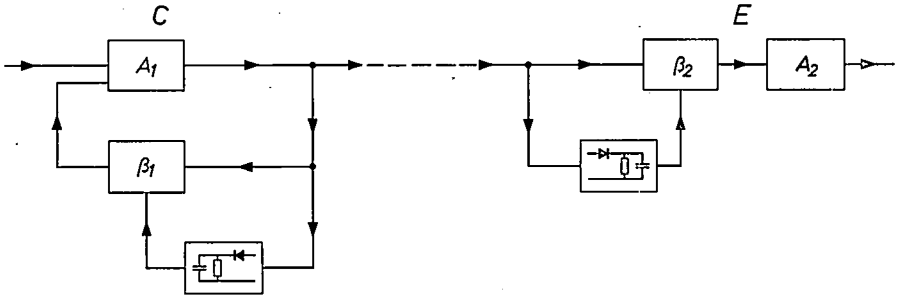


Fig. 4. Schematic diagram of a compander in which the transfer factors β_1 and β_2 of the variolossers in the compressor C and the expander E are identical. The signal at the output of the variolossers is exactly proportional both to the input signal and to the control signal. By negative feedback of the output signal from the variolosser into amplifier A_1 of the compressor pure quadratic compression is obtained. Expansion in E (which includes amplifier A_2) is also quadratic.

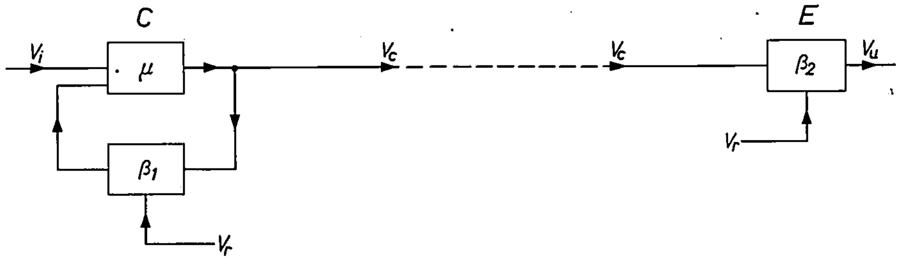


Fig. 5. A simplification of fig. 4. V_i is the input signal to the compressor, V_c is the compressor output signal and also the input signal to the expander, V_u is the output signal of the expander, μ is the gain factor of the compressor amplifier and β_1 and β_2 are the transfer factors of the variolossers in the compressor and the expander.

system between compressor and expander. Some of the compander designs discussed in this article make use of an auxiliary signal added to the speech signal with the object of overcoming this difficulty.

Companders without an auxiliary signal

Fig. 3 shows a combination of a compressor and an expander circuit such as may be used in practice. In the compressor C the speech signal passes through a variolosser whose transfer factor is indicated as β_1, and an amplifier A_1. At the output of the amplifier part of the signal is branched off, rectified, filtered and then applied to the variolosser. In the expander the speech signal also passes via a variolosser whose transfer factor is β_2, and an amplifier A_2. In this case the control signal is obtained by tapping the input to the variolosser. From these details it will be appreciated that the characteristic of β_1 must be such that attenuation increases with the input speech level, whilst for β_2 the reverse must be true.

If overall distortion is to be zero, the characteristics of β_1 and β_2 must be exactly complementary. This requirement is all the more imperative as every error will be magnified by the expanding action of E. The higher the factor of compression and expansion, the more difficult it becomes to make the characteristics of β_1 and β_2 exactly complementary, and this explains why with the type of circuit shown in fig. 3 compression is generally limited to a factor of 2 [3].

Compared with that of fig. 3, the circuit of *fig. 4* has the advantage that the characteristics of β_1 and β_2 can be identical. In compressor C the output signal of amplifier A_1 is applied both to the input of a variolosser with transfer factor β_1 and to a rectifier circuit. The output from the rectifier is used to control the attenuation of the variolosser, and the output from the latter is then applied to the input of A_1 as a negative feedback signal.

Using the circuit shown in fig. 4, the authors have been able to obtain a very low overall distortion figure by employing variolossers based upon application of the Hall effect. The underlying physical principle ensures that the output signal from the variolosser always remains truly proportional to both the control signal and the input signal. This proportionality is independent of the absolute signal level (this is not so in variolossers using diodes). As a result, signal control may be effected at a point of high speech level and amplification in the control loop kept down.

Fig. 5 is a simplified version of fig. 4. Amplification (not counting feedback) in A_1 is by a factor μ, whilst the ratio between the output and input voltages of the variolosser is equal to β_1. This ratio depends on the

[1] H. Mann, H. M. Straube and C. P. Villars, A companded coder for an experimental PCM terminal, Bell Syst. tech. J. **41**, 173-226, 1962.

[2] F. de Jager and J. A. Greefkes, "Frena", a system of speech transmission at high noise levels, Philips tech. Rev. **19**, 73-83, 1957/58.

[3] N. Valentini, The dynamics compressor-expandor (compandor) in telephony, Telettra **2**, 12-22, 1954.

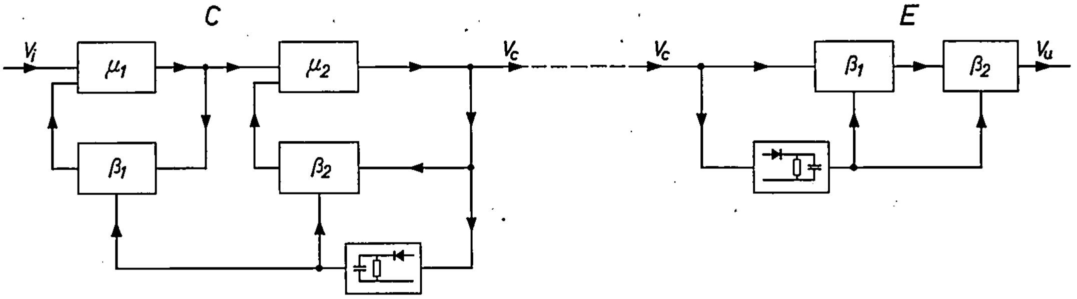


Fig. 6. If two control loops of the type shown in fig. 4 are connected in tandem, it is possible to obtain a factor-3 compander without having to introduce excessive gain into the control loops. C is the compressor, E is the expander, V_i is the input signal to the compressor, V_c is the compressed signal at the output of the compressor and also the input signal to the expander, V_u is the expander output signal, β_1 and β_2 are the transfer factors of the variolossers in the compressor and the expander and μ_1 and μ_2 are the gain factors of the amplifiers in the control loops of the compander.

magnitude of the control voltage V_r. For a given frequency the relation between the output voltage V_c and the input voltage V_i to the negative feedback amplifier may be written as:

$$V_c = \mu\, V_i/(1 + \mu\,\beta_1)\,. \quad \ldots\ldots \quad (1)$$

For high overall amplification in the control loop ($\mu\,\beta_1 \gg 1$), this may be simplified to:

$$V_c = V_i/\beta_1. \quad \ldots\ldots\ldots\ldots \quad (2)$$

If we assume the attenuation of the transmission path to be constant and, for convenience, zero, the input voltage to the expander is also equal to V_c. If β_1 and β_2 are given the same characteristics, we may write:

$$V_u = \beta_2 V_c = \beta_1 V_c\,. \quad \ldots\ldots \quad (3)$$

Combining (2) and (3) we obtain:

$$V_u = V_i \quad \ldots\ldots\ldots\ldots \quad (4)$$

which shows that the output voltage is identical with the input voltage.

The control voltage V_r varies linearly with the compressed signal V_c. If the transfer factor β_1 is made a linear function of V_r, it will also be a linear function of V_c and we may write:

$$\beta_1 = k_1\, V_c. \quad \ldots\ldots\ldots\ldots \quad (5)$$

Substitution of this value in (1) yields:

$$V_i = V_c/\mu + k_1\, V_c^2\,. \quad \ldots\ldots \quad (6)$$

For $\mu\,\beta_1 \gg 1$, the linear term on the right-hand side may be neglected with respect to the quadratic term, leading to the simplified relation:

$$V_c = \sqrt{V_i/k_1}\,. \quad \ldots\ldots\ldots\ldots \quad (7)$$

This equation shows that we have obtained a factor-2 compressor.

Expansion is also quadratic, since combination of (3) and (5) gives:

$$V_u = k_1 V_c^2\,. \quad \ldots\ldots\ldots \quad (8)$$

Expansion is quadratic under all circumstances, but a square-law compression characteristic is obtained with sufficient approximation only if in the compressor $\mu\,\beta_1 \gg 1$. If this condition cannot be satisfied to a sufficiently high degree, for example because stability would otherwise become critical, then the linear term in (6) cannot be entirely neglected. Under such circumstances the expansion can be made exactly complementary to the compression by adding a small constant fraction to the variable control voltage that is obtained from the speech signal and applied to the variolosser in the expander.

If the transfer factor β is made proportional to the square of the control voltage V_r, we obtain the relation:

$$V_c \propto \sqrt[3]{V_i}. \quad \ldots\ldots\ldots \quad (9)$$

In other words, the result would be a factor-3 compressor. However, to realize a variolosser whose transfer factor varies with the square of the control voltage is by no means simple. In addition, the varying gain in the control loop would, at such a high degree of compression, make it difficult to maintain stability. It is possible to overcome both these problems by connecting two control loops in tandem, as shown in *fig. 6*. In the compressor and expander alike, the transfer factors β_1 and β_2 of the variolossers in the two control loops are varied simultaneously by one and the same control signal obtained by rectification and smoothing of the compressed signal.

For the compressor we may take the relation:

$$V_c = \frac{\mu_1}{1 + \mu_1\beta_1}\,\frac{\mu_2}{1 + \mu_2\beta_2}\,V_i. \quad \ldots \quad (10)$$

As in the previous case, the amplification in the two control loops can be made large enough to satisfy the relations $\mu_1\beta_1 \gg 1$ and $\mu_2\beta_2 \gg 1$, which causes (10) to simplify to:

$$V_c = V_i/\beta_1\beta_2 . \quad (11)$$

It will be seen that, if we make β_1 and β_2 equal to one another, and cause them to vary linearly with the control voltage a relation such as (9) is obtained or, in other words, the compression factor becomes equal to 3. Similarly, it can be shown that expansion will be cubic.

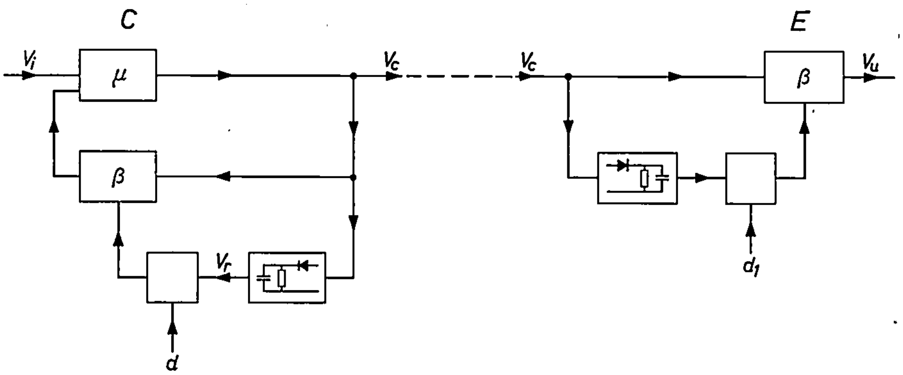


Fig. 7. If a threshold voltage d is introduced into the control voltage circuit of the compressor of fig. 4, the latter, although its compression is in principle quadratic, will yield higher compression factors than 2. By suitable adjustment of the threshold voltage d_1 in the expander the characteristics of compressor and expander can be made complementary. V_r is the control voltage.

A compander based on the principle of fig. 6 has been built in this laboratory and factors of compression and expansion equal to 3 were obtained without difficulty. Here again, it proved possible to keep the overall distortion low by using variolossers based on the Hall principle. Simultaneous variation of the attenuation of the two variolossers was effected by putting two Hall plates inside a single coil whose magnetic field was generated by the control current.

Increasing the degree of compression with a single control loop

In discussing the general properties of companders we have shown that, if a relation of the type $y = \sqrt[n]{x}$ exists between the output and input voltages of a compressor, the variations in level at its output, measured in decibels, are always a constant fraction of those at its input. If we confine ourselves to the factor-2 compressor, which is easy to realize with the aid of the circuit in fig. 4, constant compression is obtained for every interval in the parabola representing the equation $y = \sqrt{x}$.

If, however, the top of the parabola is moved from the origin to the second quadrant, the equation representing the curve takes the form $y = b + \sqrt{x + a}$. For an equation of this form the ratio

$$\frac{b + \sqrt{x_2 + a}}{b + \sqrt{x_1 + a}}$$

is, for positive values of the roots, always smaller than the ratio

$$\frac{\sqrt{x_2}}{\sqrt{x_1}},$$

as long as $x_2 > x_1 > 0$. This means that the compression factor always exceeds 2, although its actual value depends on the interval chosen on the curve.

Such a shift of the top of the parabola can be obtained by introducing a threshold voltage in the manner shown in *fig. 7*. The transfer factor β of the variolosser is equal to zero as long as the control voltage V_r, obtained from the output voltage V_c of the compressor, does not exceed the threshold voltage d. Beyond this point β increases linearly with the amount by which V_r exceeds d.

Until this threshold value has been reached, V_c increases linearly with the input signal. Thereafter it is easy to show that the relation between the two values can be represented by a parabolic curve defined by an equation of the form $y = b + \sqrt{x + a}$. *Fig. 8* may be helpful in demonstrating the effect of introducing a threshold voltage. In this diagram the parabola shown is represented by the equation $y = \sqrt{x}$ with

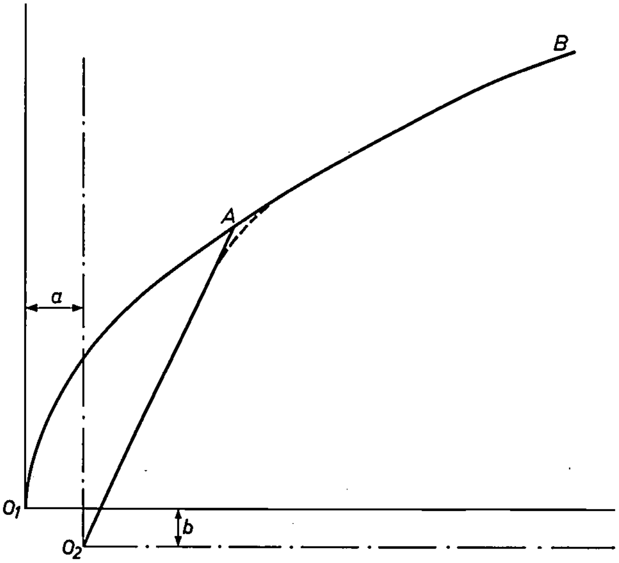


Fig. 8. The introduction of a threshold voltage in fig. 7 is equivalent to shifting to O_2 the origin O_1 of the system of coordinates for which the parabola shown in the diagram satisfies the equation $y = \sqrt{x}$. With respect to the new system of coordinates the curve is represented by $y = b + \sqrt{x + a}$, an equation which yields a compression factor greater than 2 for the first quadrant.

respect to the system of co-ordinates with its origin at O_1, while equation $y = b + \sqrt{x + a}$ applies to the system with its origin at O_2. Owing to the presence of the threshold voltage the control curve now includes a linear section O_2A and a quadratic section AB. In reality the transition between the two sections of the curve will not be as abrupt as the diagram suggests, but will follow the dotted connecting line.

Using the principle described, it was found possible to reduce a variation in level of 30 dB at the compressor input to less than 10 dB at its output, i.e. to considerably less than half its original value, while avoiding the complication of the circuit in fig. 6. In fig. 8 the slope of section O_2A is greater than that of the line $y = x$ in fig. 1, indicating that at the lower end of the scale the speech level has been raised a fixed amount by adding extra gain.

In a similar manner, a threshold voltage is introduced into the expander, so that there is no expansion at low levels, whilst beyond a certain input level, expansion by a factor greater than 2 takes place. By suitable adjustment of the threshold voltages the system can be made linear.

Companders with an auxiliary signal

In the companders described so far, the level of the speech signal applied to the input is in itself a measure of the degree of expansion to be applied. The design of the expander must of necessity be based on the assumption that the envelopes of the transmitted and the received signals are identical in shape. As soon as this ceases to be true, expansion will become incorrect and the expander will tend to make the deviations worse.

On the transmission path the speech signal is subject to two forms of distortion. In the first place, the attenuation of the transmission circuit may show spontaneous variations, and in the second place there may be amplitude distortion due to noise. In the latter case it is unsatisfactory that the expander has to read the amplitude variations from the speech signal itself, for the entire width of the speech band is then required to transmit the amplitude information. Consequently, a large amount of noise energy will enter the control circuit.

In the three circuits that we shall now describe, an auxiliary signal has been introduced with the object of eliminating the effects of one or both of the two sources of distortion we have mentioned. The average level of a speech signal varies comparatively slowly; therefore, only a narrow band of frequencies is occupied if these variations are modulated upon an auxiliary signal. This band of frequencies can then be placed either below or above the speech band, so that the two bands can be separated with ease. As the bandwidth occupied by the auxiliary signal is limited, little noise energy can enter this information channel.

In the first two companders described below the auxiliary signal is instrumental in eliminating the influence of spontaneous variations in the attenuation of the transmission path. The third compander has been designed for transmission systems whose method of modulation itself eliminates the effects of spontaneous variations in attenuation; we mention as examples frequency modulation and code modulation.

The Ensink and Verhagen pilot compander

The first example we shall describe is a compander developed by J. Ensink and J. Verhagen of N.V. Philips' Telecommunicatie Industrie, Hilversum. A circuit diagram is given in *fig. 9*. The speech signal with amplitude V_i is applied to a variolosser with transfer factor β_1. At the input to this variolosser is added an auxiliary signal having the character of a pilot signal: it has constant amplitude and frequency f_p. As this compander was planned for application in multichannel carrier telephone systems in which the band of speech frequencies transmitted extends from 300 to 3400 c/s, a pilot frequency of 3700 c/s was chosen. For convenience we shall consider the amplitude of the pilot signal as unity. On leaving the variolosser the combined signal undergoes amplification by a factor μ_1, and a fraction of the signal is tapped off and rectified. After passing through a low-pass filter with a cut-off frequency of approx. 100-125 c/s, the signal has to overcome a threshold voltage E_1. The threshold value is so adjusted that in the absence of a speech signal the rectified output voltage (at pilot frequency) of the compander exactly equals E_1, V_{r_1} the control voltage, is then obviously zero.

After amplification by a factor μ_2 the control signal is applied to the variolosser. In the present compressor control is effected at a point of low signal level, and this permits the use of diodes in the variolosser. The circuit is such that the following relation holds between β_1 and V_{r_1}:

$$\beta_1 \doteqdot (1 + k_1\mu_2 V_{r_1})^{-1}.$$

Therefore $\beta_1 = 1$ when there is no speech ($V_{r_1} = 0$) and the pilot signal amplitude at the output of the variolosser will then also be 1. At the output of the amplifier following the variolosser the pilot signal amplitude will be μ_1, and for this value the rectified output voltage just matches the threshold voltage E_1.

When there is a speech signal, a control voltage V_{r_1} is generated. The factor μ_2 by which this control voltage is amplified is chosen such that the voltage at the output of the variolosser will never exceed the

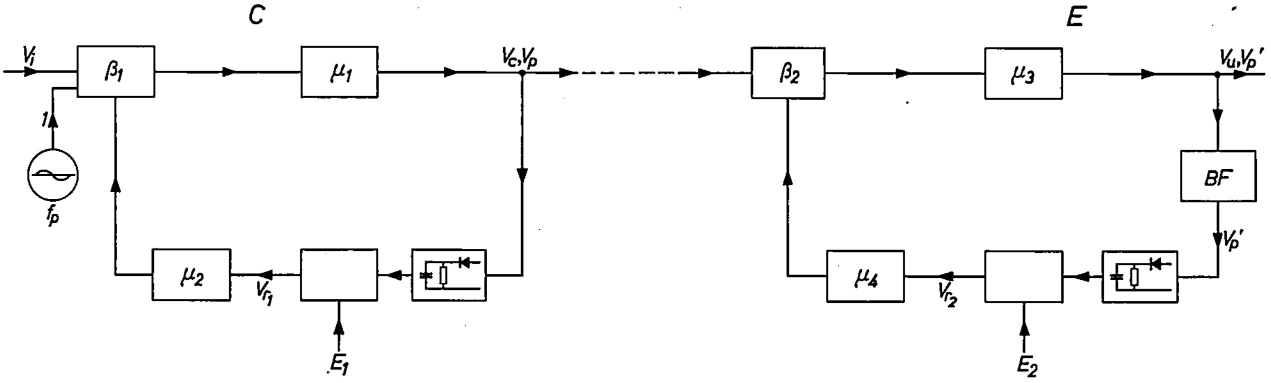


Fig. 9. Diagram of a pilot compander after Ensink and Verhagen. As the information giving the amplitude variations of the speech signal is now conveyed over a narrow band of frequencies by the pilot signal, the influence of noise is reduced. In addition, the circuit arrangement yields protection against the effect of spontaneous variations of the attenuation on the transmission path.

quiescent value 1 by a significant amount. As the input voltage to the variolosser has the value $1 + V_i$, the compression ratio obtained is obviously $1/(1 + V_i)$. In the output voltage the speech component has the value $V_c = \mu_1 V_i/(1 + V_i)$, while the pilot frequency component has the value $V_p = \mu_1/(1 + V_i)$. These equations show that the speech component at the compressor output approaches asymptotically the value μ_1, while at the same time the pilot component approaches zero. *Fig. 10* is a graphical representation of the variation of the two components as functions of the speech signal amplitude at the input to the compressor.

In the expander, the diagram of which is shown on the right hand side of fig. 9, expansion takes place in a variolosser circuit with transfer factor β_2, to which the incoming signal is applied. After expansion the signal is amplified by a factor μ_3. The pilot component (amplitude V_p') is then separated from the speech component (amplitude V_u) by means of a low-pass filter not shown in the diagram.

Part of the combined signal is tapped off and applied to a band-pass filter *BF* that passes only the pilot and its sideband frequencies. After filtering, the signal is rectified, filtered a second time, and then meets a threshold with a voltage E_2. The latter is so adjusted that for the minimum expected value of the pilot signal amplitude V_p' it exactly equals the rectified voltage. Whenever the rectified pilot signal exceeds this minimum a control voltage V_{r_2} is generated. After amplification by a very large factor μ_4 this control voltage is used to vary the transfer factor β_2.

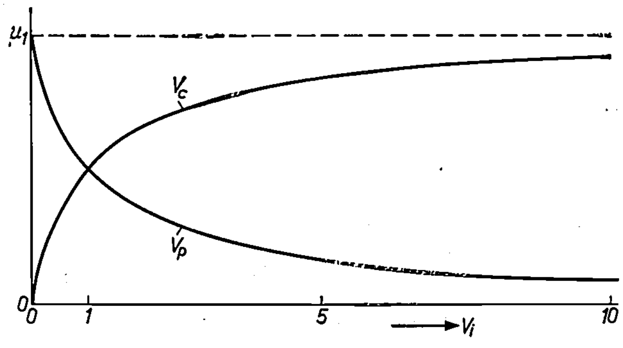


Fig. 10. The speech amplitude V_c and the pilot signal amplitude V_p at the output of the compressor of fig. 9, as a function of the incoming speech signal V_i.

The transfer factor β_2 depends on the control voltage V_{r_2} in exactly the same way as β_1 depends on V_{r_1}; therefore, when $V_{r_2} = 0$, $\beta_2 = 1$. For the corresponding value of the pilot voltage — it must be kept in mind that this coincides with the maximum expected value of the speech signal — the gain between the input and the output of the expander will therefore be a maximum. Since the amplification factor μ_4 is very high, the value of β_2 will drop sharply for even a slight increase in pilot voltage. It follows that, even for large variations of the incoming pilot signal, its amplitude variations at the output will remain insignificant or, in other words, the pilot signal voltage at the output will be maintained at a constant value of 1. Since the ratio between the speech and pilot frequency amplitudes has remained invariable at $V_i : 1$, the speech amplitude must of necessity assume the value V_i, identical to that at the compressor input. Moreover, this result is independent of any spontaneous variations of the attenuation on the transmission path.

Limiter-compander

Like the preceding compander arrangement, the one which we shall now discuss adds a pilot signal of constant frequency f_p and constant amplitude 1 to the speech signal of variable amplitude V_i. Here again, the amplitude $(1 + V_i)$ of the composite signal is reduced to the fixed value 1 by the compressor; compression, in other words, is in the ratio of $1 : (1 + V_i)$. In the expander the pilot signal amplitude is held at the constant value 1, causing the speech component to assume the value V_i automatically.

The control circuit employed in the compander now being discussed is somewhat unusual for this type of application in that it makes use of a limiter. We shall begin by showing that linear control of signal amplitude can be effected by means of a limiter.

To this end we shall consider a single frequency component of an information signal having an angular frequency of ω_i and an amplitude of V_i. To this information signal we add an auxiliary signal with an angular frequency ω_h and an amplitude V_h, taking care that $V_h \gg V_i$. As shown in *fig. 11a* this composite signal can be represented in a vector diagram; the time vector is assumed to rotate at the angular frequency ω_h. Vector *OB*, representing the composite signal, is made up of a stationary vector *OM*, representing V_h, and a vector *MB* rotating at an angular frequency $\omega_h - \omega_i$, representing V_i. The amplitude variation of the composite signal is equal to 2 V_i, its phase variation to 2 φ_m.

If we apply such a signal to a limiter, the phase variation will remain small as long as $V_h \gg V_i$. Although the limiting process generates harmonics of the auxiliary frequency ω_h, these are eliminated with the aid of a low-pass filter, as we shall see later. This permits the output signal from the limiter to be represented by a vector *OR* of constant magnitude — we take this as being equal to 1 — whose end point moves along arc *PQRS*. For small values of φ_m we may consider arc *PQRS* as a straight line and then simplify the diagram for the limited voltage to that of fig. 11*b*. The end point *R* of vector *OR* is then seen to oscillate between points *P* and *S* at a frequency $(\omega_h - \omega_i)/2\pi$. The amplitude *PQ* can now be found: fig. 11*a* shows that $\sin \varphi_m = V_i/V_h \approx \varphi_m$, and from fig. 11*b* $\tan \varphi_m = \mathrm{PQ} \approx \varphi_m$; whence $PQ = V_i/V_h$.

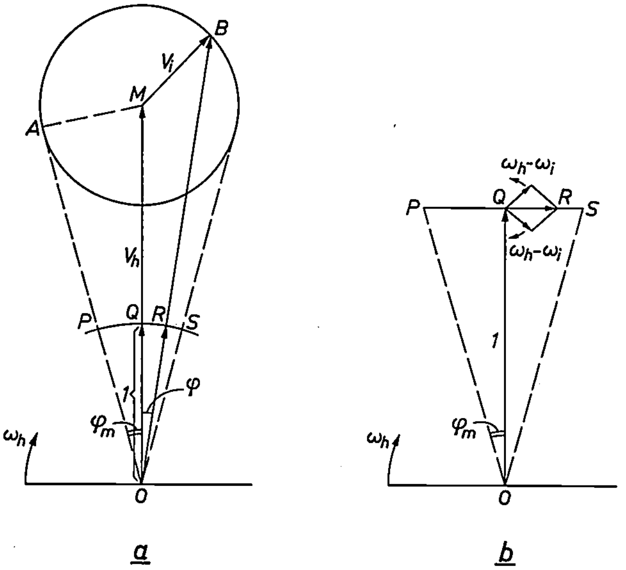


Fig. 11. Vector diagram illustrating the operation of a limiter as a control element.

The oscillating vector *QR* in fig. 11*b* may be considered as being composed of two rotating vectors with the same amplitude $\frac{1}{2}PQ$ and angular frequencies of $(\omega_h - \omega_i)$ and $-(\omega_h - \omega_i)$. The limited signal thus has three components: one of magnitude 1 and angular frequency ω_h, one of magnitude $\frac{1}{2}$ V_i/V_h and angular frequency $\omega_h - (\omega_h - \omega_i) = \omega_i$, and one of magnitude $\frac{1}{2}V_i/V_h$ and angular frequency $\omega_h + (\omega_h - \omega_i) = 2\omega_h - \omega_i$. If sufficient difference is maintained between ω_h and ω_i, the first and the third component can be easily separated from the second and suppressed. The amplitude of the signal remaining after limiting and filtering will then be proportional to V_i, and inversely proportional to V_h. Since a linear relation is found to exist between the incoming information signal and the limited signal, it follows that for a speech signal containing a complex of frequencies and having a variable amplitude V_i, the limited signal will also depend linearly upon the incoming signal.

Let us now consider the compressor and expander circuits of *fig. 12*. Upon entering the compressor, the speech signal with variable amplitude V_i is first applied to a filter BF_1 with a pass band ranging from 300 to 3400 c/s. At the output of this filter a pilot signal with frequency f_p is added. This frequency is situated above the speech band, for example at 3700 c/s. It again has a constant amplitude which we will take as 1.

The composite signal, of amplitude $1 + V_i$, is applied both to a limiter CL_1 and to a rectifier circuit. At the output of the rectifier the signal amplitude is equal to $1 + V_i$, which only varies slowly. After filtering in F_1 the rectified signal modulates the amplitude of an auxiliary signal with a frequency of $f_{h_1} = \omega_{h_1}/2\pi$ in such a way that the amplitude of this signal is again equal to $1 + V_i$.

Upon leaving the modulator M_1 the auxiliary signal is amplified by a factor μ_1 in A_1, so that its amplitude becomes $\mu_1 (1 + V_i)$, and then applied to limiter CL_1, together with the combined speech and pilot signal. 20 000 c/s was found to be a suitable value for f_{h1}; the value of μ_1 does not need to be much higher than 4 in order to guarantee a sufficiently linear relation between the output signal from the limiter and the combined speech and pilot signal.

It follows from our theoretical considerations that the wanted component of the limiter output signal (unwanted components are suppressed by filter BF_2) is directly proportional to the amplitude $(1 + V_i)$ of the speech-plus-pilot signal and inversely proportional to the amplitude $\mu_1 (1 + V_i)$ of the auxiliary signal. In other words, the variable amplitude $(1 + V_i)$ of the speech-plus-pilot signal is reduced to the constant

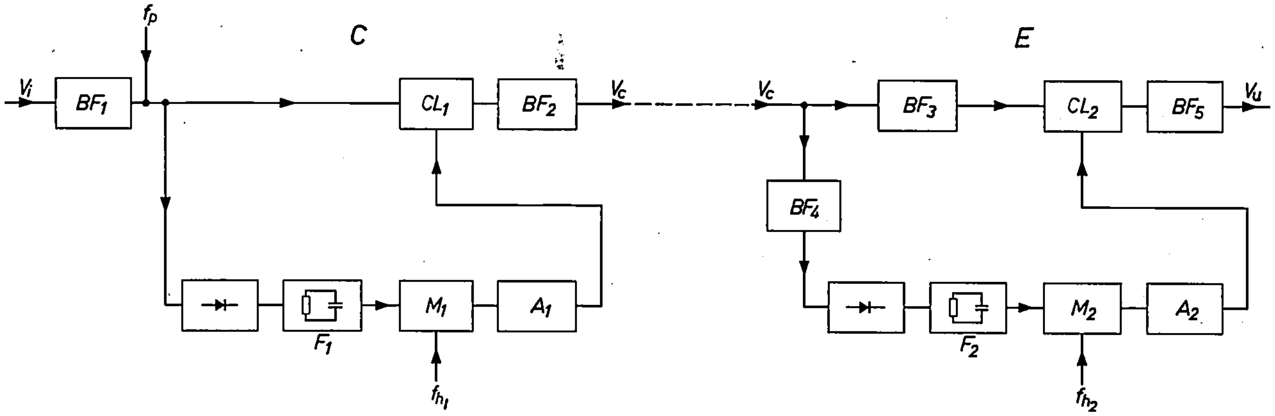


Fig. 12. Diagram of a compander arrangement in which limiters CL_1 and CL_2 are used as control elements. Like that of the compander in fig. 9, the compression characteristic of this compander can be represented by a curve of the form of fig. 10.

value $1/(2\,\mu_1)$. In this constant signal the amplitude of the pilot frequency component is proportional to $1/(1+V_i)$, while that of the speech component is proportional to $V_i/(1+V_i)$. The shape of the compression characteristic is the same as that given in fig. 10 for the Ensink-Verhagen compander.

The amplitudes of the pilot and speech components in the transmitted signal vary simultaneously, but in opposite directions. As these amplitudes only vary slowly, the pilot signal takes up only a limited bandwidth: 50 c/s to either side of the 3700 c/s pilot frequency is sufficient. Two filters BF_3 and BF_4 in the expander (see fig. 12, on the right) can therefore be used to separate the pilot and speech components without difficulty. Limiter CL_2 receives the band of speech frequencies from 300 to 3400 c/s, while the pilot and its sideband frequencies are rectified, filtered and then used to modulate the auxiliary voltage, which has a frequency $f_{h_2} = 20\,000$ c/s, in M_2. After amplification in A_2 the modulated auxiliary voltage is also applied to CL_2.

Leaving aside the unwanted components of the output signal from CL_2 that are suppressed by filter BF_5, the wanted component, as our theoretical considerations have shown, will be directly proportional to the signal from BF_3 and inversely proportional to that from A_2. We also know that the signal from A_2 is proportional to $1/(1+V_i)$, while the signal from BF_3 is proportional to $V_i/(1+V_i)$. The wanted component at the output of CL_2 is therefore proportional to:

$$\left(\frac{V_i}{1+V_i}\right)\Big/\left(\frac{1}{1+V_i}\right) = V_i\,.$$

Hence, the output signal is a true copy of the input signal to the compressor; once again, this result is independent of any spontaneous variations of the attenuation on the transmission path.

Compander with a d.c. auxiliary signal

In the third compander of the group we are discussing, the auxiliary signal is introduced in such a form as to give no protection against spontaneous variations of the attenuation on the transmission path. It is, therefore, best to use this compander in combination with modulation systems that give this protection naturally, such as frequency modulation or code modulation systems. The facility, offered by these modulation systems, of permitting the transmission of a d.c. component in the information signal, has been exploited in order to maintain the d.c. character of the auxiliary signal during transmission.

The sole purpose of the auxiliary signal in this compander is to transmit the information concerning the amplitude variations in a narrow band of frequencies situated outside the speech band, and thus reduce the interference due to noise generated on the transmission path. This narrow band of frequencies is most conveniently located below the standard speech-frequency-band (300 to 3400 c/s), e.g. from 0 to 125 c/s, an arrangement which permits separation of the two bands at the receiving end by means of a very simple filter.

As *fig. 13* shows the speech signal is put through a high pass filter F_1. This suppresses all low frequencies

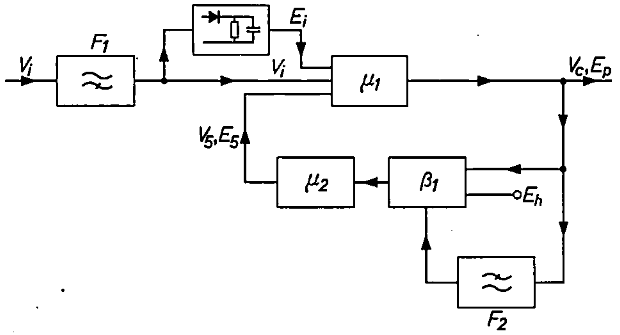


Fig. 13. Diagram of a compressor with a d.c. auxiliary signal, yielding a high degree of compression.

that might interfere with the auxilliary signal. Beyond this filter the speech signal, with amplitude V_i, is applied both to the input of a d.c. amplifier with amplification factor μ_1 and to a rectifier circuit. The d.c. output signal from the latter, of variable amplitude E_i, is then applied to the input of the same amplifier.

Part of the combined signal at the output of the amplifier is tapped off and applied both to the input of the variolosser with transfer factor β_1 and to the low-pass filter F_2. The latter filter passes only the band of frequencies making up the auxiliary signal, which is then used to control the transfer factor of the variolosser.

After amplification by a factor μ_2 the output signal from the variolosser is fed back into the d.c. amplifier in phase opposition to V_i and E_i. The control characteristic of the variolosser is so chosen that the output voltage is proportional to both the input voltage and the control voltage.

From fig. 13 and what has been said above, it is not difficult to arrive at the following equation for the compressed speech component V_c:

$$V_c = \mu_1 (V_i - V_5) \quad \ldots\ldots \quad (12)$$

in which

$$V_5 \propto \mu_2 E_p V_c . \quad \ldots\ldots\ldots \quad (13)$$

The following relation holds for the auxiliary signal:

$$E_p = \mu_1 (E_i - E_5) \quad \ldots\ldots \quad (14)$$

in which

$$E_5 \propto \mu_2 E_p (E_p + E_h) . \quad \ldots\ldots \quad (15)$$

In the last expression E_h represents an auxiliary d.c. voltage applied to the variolosser. It is introduced with the same object as the threshold voltage d in fig. 7, viz to raise the compression factor at high speech levels.

Relations (12) and (13) indicate that V_c depends not only on the input signal V_i but also on the amplification factors μ_1 and μ_2 and on the auxiliary signal E_p. As (14) and (15) show, the auxiliary signal in its turn depends on the auxiliary d.c. signal E_h applied to the variolosser, and again on μ_1 and μ_2. From (12), (13), (14) and (15) mathematical relations can be derived between the compressed voltages V_c and E_p on one hand, and the input voltage V_i, the amplification factors μ_1 and μ_2 and the auxiliary voltage E_h on the other. However, these relations are intricate and difficult to evaluate, and we prefer to represent them graphically. *Fig. 14* shows the compressed speech component V_c — indicated as a fraction of its maximum expected value $V_{c\,max}$ — as a function of the input voltage V_i, the latter being likewise shown as a fraction of its maximum expected amplitude $V_{i\,max}$. A number of curves are given, corresponding to various values of the product $\mu_1 \mu_2$ as a parameter. The auxiliary d.c. voltage E_h to which these curves apply has been taken equal to the maximum expected value of E_p. It will be seen that the shape of the various curves depends closely on the value of $\mu_1 \mu_2$. For the maximum value indicated $\mu_1 \mu_2 = 500$, a high degree of compression is obtained; the actual value found depends on the selected interval on the curve.

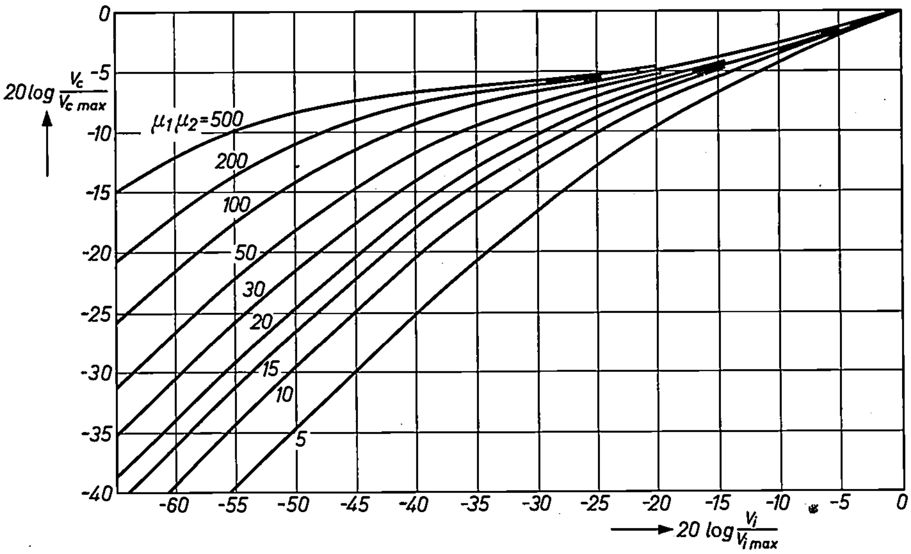


Fig. 14. Curves showing the compression of the speech signal obtained with a compander of the type represented in fig. 13. The overall amplification $\mu_1 \mu_2$ in the control loop has been used as a parameter.

As to the auxiliary signal E_p at the output of the compressor, this is shown in *fig. 15* as a function of the amplitude V_i of the incoming speech signal. Both magnitudes are shown as fractions of their maximum expected values. Curve a holds for a d.c. auxiliary voltage $E_h = E_{p\,max}$. It varies little in shape with the value of the product $\mu_1 \mu_2$ and comes close to a linear relation between the two variables.

Comparison of figs. 13 and 4 shows them to be equivalent if we omit E_h from fig. 13. This means that for $E_h = 0$ the circuit of fig. 13 yields a compression factor of 2, both for the speech component and for the auxiliary signal. Consequently, curve b in fig. 15, for which $E_h \doteq 0$, is a parabola.

If we ignore for a moment the auxiliary voltage E_{h_1} in *fig. 16* the expander circuit shown there is the same as that in fig. 4, and gives quadratic expansion. It is worth pointing out that the auxiliary voltage E_{h_1} in

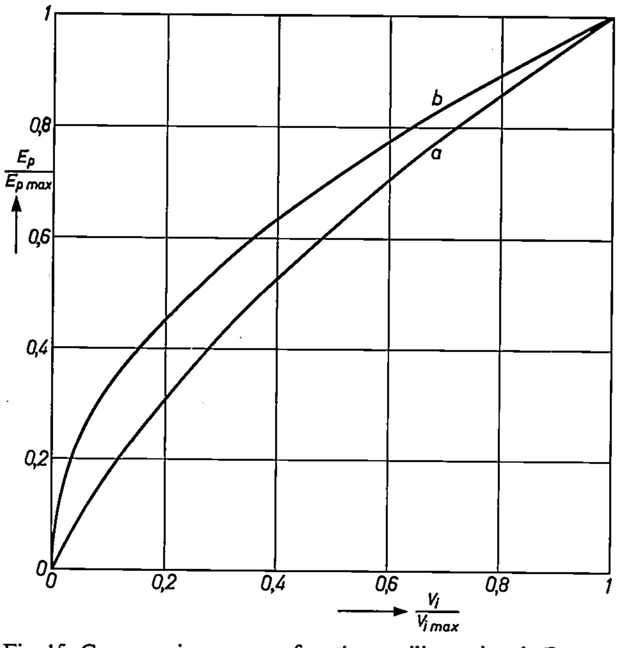


Fig. 15. Compression curves for the auxiliary signal E_p as a function of the incoming speech signal V_i, for the compressor shown in fig. 13. Curve *a* is for the case where the auxiliary d.c. voltage E_h in fig. 13 is equal to the maximum expected value of the compressed signal E_p; curve *b* is for $E_h = 0$.

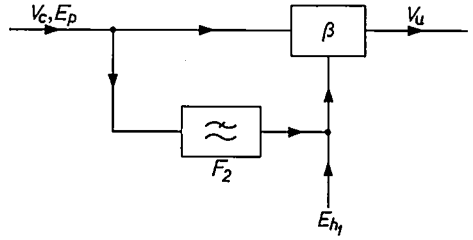


Fig. 16. Diagram of the expander used in conjunction with the compressor of fig. 13.

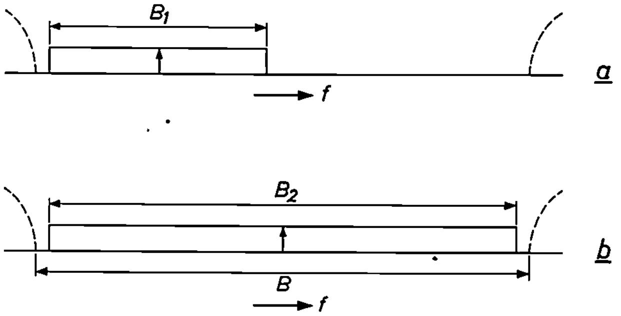


Fig. 17. Diagram showing how the auxiliary signal E_p of the compressor in fig. 13 may be used to control the average frequency in a frequency modulation system. In *a*, only section B_1 of the total available frequency band B is occupied. At full modulation, as shown in *b*, the band of frequencies occupied is extended to a width B_2, and the average frequency now coincides with the centre of the available band. The information contained in the auxiliary signal is thus transmitted without requiring extra bandwidth.

fig. 16 must not be viewed as a compensation for the presence of the auxiliary voltage E_h in the compressor of fig. 13. E_{h_1} is introduced with the same object as the auxiliary voltage in the expander of fig. 4. It only becomes necessary to introduce E_{h1} when the compression curve departs from the quadratic law because of insufficient overall gain.

If the compander with a d.c. auxiliary signal which we have just discussed is used in a frequency modulation system, the method set out in *fig. 17* may be followed. Fig. 17*a* indicates how for low modulation depths, corresponding to a limited frequency devation, the band of frequencies occupied is situated eccentrically with respect to the total band of frequencies allotted to the channel. When the modulation level is raised, the average frequency of the band occupied is shifted, under the control of the auxiliary signal, towards the centre of the available band, until, at full modulation, the entire band is occupied. This condition is shown in fig. 17*b*. The advantage of this procedure lies in the fact that hardly any additional power or bandwidth is required for the transmission of the auxiliary signal.

For the first two in our last group of three companders the amplitudes of the auxiliary signal and of the speech signal are always complementary. It must be realised, therefore, that very high speech levels may result in extremely low amplitudes of the auxiliary (pilot) signal. Under such conditions even a very small spurious voltage at pilot frequency introduced into the channel, by noise or crosstalk, may mean serious misrepresentation of the pilot signal level and result in incorrect expansion. In our last compander the auxiliary signal continues to increase as the speech voltage increases: hence its lower sensitivity to interference. A point on which space does not permit us to enlarge any further, but which we would mention in conclusion, is that in our last compander the influence of spontaneous variations of the attenuation on the transmission path is no greater than in a factor-2 compander, even though considerably higher compression factors can be obtained.

Mr. K. Riemens has assisted in the experimental and theoretical examination of the companders described in the present article.

Summary. After stating the reasons for the possible use of companders, the article discusses their general properties. Maximum values of the compression factor, i.e. the numerical ratio between the variations of the input and output levels of the compressor, each measured in decibels, are always aimed at. For practical reasons greater compression factors than 2 cannot usually be obtained. The article describes a number of circuit arrangements yielding considerably higher compression factors. These circuits fall under two headings. In the companders of the first group the information concerning the variations in level is conveyed by the speech signal itself. In those of the second group a separate auxiliary signal is used for this purpose and this results in less sensitivity to noise. Two of the three companders in this second group furthermore give protection against the effect of spontaneous variations of the attenuation on the transmission path between compressor and expander. In the third compander of the second group this protection is lacking in principle, but the influence of variations in transmission path attenuation is no greater than with a factor-2 compander, even though the compression factor can be made much higher than 2.

An electronic computer for air traffic control

R. A. Grijseels

681.14-523.8:656.7

Introduction

New equipment which will soon be put into regular service at the air traffic control centre of Schiphol Airport near Amsterdam, will perform a number of necessary control operations entirely automatically. Designed by N.V. Hollandse Signaalapparaten, the system adopted has been named *Signaal Automatic Air Traffic Control*, or *Satco* for short.

One of the most important equipment items is an electronic computer which is duplicated in the equipment. Before giving a detailed description of this computer or of the other equipment, it seems desirable to survey the control routines that are at present in force. When describing the new equipment, we shall then be able to indicate to what extent this routine will be modified.

The reader should bear in mind that the description applies to civil air traffic control.

Civil air traffic regulations

In air traffic control it is customary to use a number of standardized abbreviations that have by now been accepted internationally. We shall follow this custom wherever convenient.

International convention has defined a number of contiguous *Flight Information Regions* (*FIR*) which between them cover the entire surface of the earth. Inside each FIR air traffic control is centralized at one point which usually coincides with the main airport of the region. The Dutch FIR comprises the entire territory of the Netherlands, plus part of the North Sea, and Schiphol is its control centre.

The various FIR's are interconnected by a number of well-defined air lanes, each of which is characterized by a colour, followed by a number (e.g. Red 1, Green 3, etc.). Outside these air lanes aircraft may move freely, although pilots remain entitled to receive information regarding their own position and that of other aircraft. Inside the air lanes, however, pilots are required to follow the instructions given by the central air traffic control.

Air lanes have a width of 10 nautical miles; they extend in a vertical direction from 3000 feet to 25 000 feet. At various points along the air lanes pilots fly over radio beacons that enable them to determine their position. The time and height of flying over each radio beacon must be reported to the control centre. Each aircraft has a *Flight Level* (*FL*) assigned to it; in view of the limited precision of the altimeters used, the flight levels of any two aircraft following the same air lane must differ by at least 1000 feet. In a horizontal direction there is also a minimum separation which, by international agreement, has been fixed at 10 minutes' flying time. It is customary to reduce this separation in the vicinity of airports, where radar equipment is usually available.

In the air traffic control organization different functions have been assigned to the *Area Control Centre* (*ACC*), the *Approach Control* (*APP*) and the *Tower* (*TWR*). The ACC controls traffic in the FIR in general, and in the air lanes in particular. Landing procedures of aircraft approaching the airport are directed by APP, usually with the aid of radar equipment, while the Tower, which observes only visually, controls all traffic on, and in the immediate vicinity of, the airport.

Manual methods of air traffic control

Fig. 1 shows the room housing the ACC as it looked before the introduction of the new equipment. The left-hand side of the stand on the table in the centre is used for the control of traffic in the air lanes west of Schiphol; the right-hand side serves the lanes east of the airport. Each section has its own control officer — at the moment the photograph was taken, each position was for special reasons being temporarily manned by two officers. Opposite each control officer, an assistant is seated on the other side of the table.

The stand is divided into a number of sections, one for each beacon in the Dutch FIR. In each section there is a metal holder carrying a paper flight progress strip for each aircraft that is shortly to pass over the respective beacon. A number of details of the flights are marked on these strips, such as flight number, type of aircraft, air speed, flight level, route to be followed, and the expected time of flying over the beacon. By comparing the various strips the control officer can estimate whether conflict situations are likely to arise, i.e. situations in which aircraft no longer observe the minimum prescribed separations.

Each pilot of an aircraft wishing to leave Schiphol submits his flight plan to the *Flight Information Office* (*FIO*), whence the details are passed on to the assistant

Ir. R. A. Grijseels is on the staff of N.V. Hollandse Signaalapparaten, Hengelo.

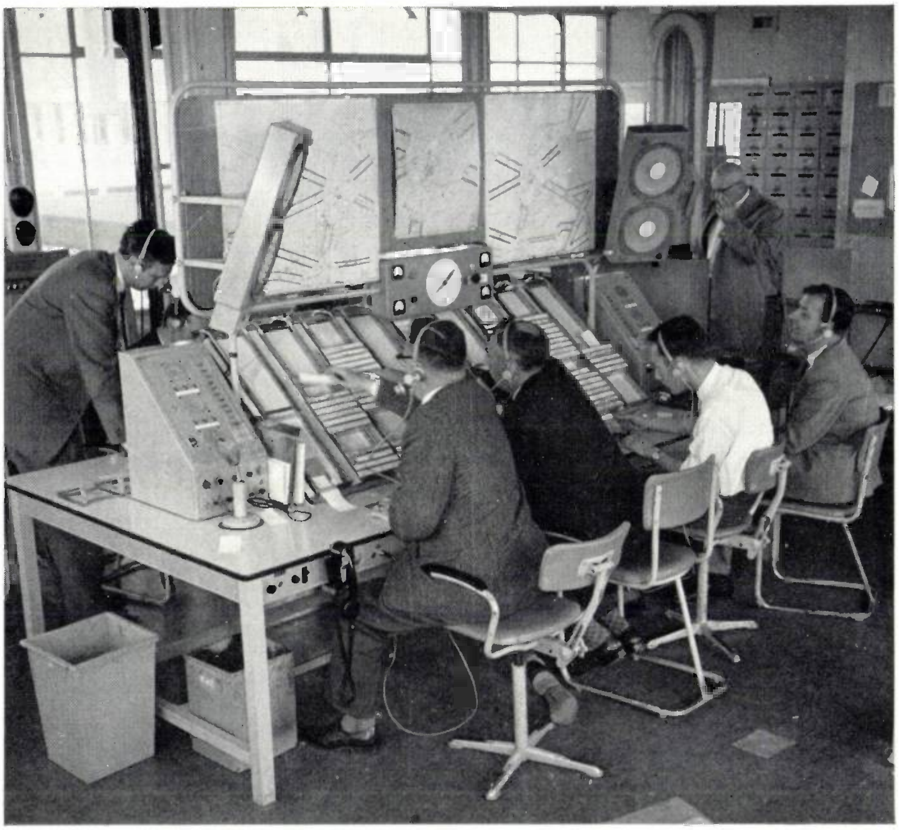


Fig. 1. The air traffic control room before the introduction of the new equipment. The stand on the table is divided into a number of sections, each of which corresponds with one of the reporting positions in the Flight Information Region. Each section contains a flight progress strip for every aircraft that is shortly to pass over the reporting position. By comparing the progress strips for each reporting position the control officer can evaluate the risk of possible conflict situations.

control officer by telephone. Information about aircraft which will enter the Dutch FIR from adjoining FIR's is also passed to the assistant by telephone. On the basis of the information received the flight path of the aircraft is calculated in order to determine at what time, speed and height it will fly over the various beacons on its way. For each of these beacons a flight progress strip is made out and passed to the control officer.

Usually the data on the flight progress strips have to undergo many modifications: first of all, departures or arrivals are often delayed and, secondly, the control officer may consider it necessary to modify the flight schedule in order to avoid conflict situations. For each modification the flight path has to be calculated afresh, and the data on the flight progress strips relating to the flight concerned have to be brought up to date. In the case of an aircraft that is not going to land at Schiphol airport but is merely flying over, both the control officer West and the control officer East have to modify their strips.

Air traffic is constantly growing in density. Moreover, the aircraft now comprising it fly at greatly varying speeds and heights and there are propeller aircraft with piston engines as well as jet and turbo-jet aircraft, so that the risk of flight paths crossing one another has greatly increased. As a result it is becoming increasingly difficult to combine the control of traffic proper with the work of making the flight path calculations, preparing flight progress strips and keeping them up to date. The task of the new equipment is the automation of these operations.

The new equipment

From the operating point of view two of the most important items of the new equipment are the automatic flight progress boards that will be put at the disposal of the control officer West and the control officer East. They are the boards on the left and right of the group of three shown in *fig. 2*. The front panels of each board are subdivided into a number of sections, each corresponding to a beacon in the FIR. Inside each section, the data relating to the various flights are shown on automatic display units arranged in horizontal rows. These display units (to be described in more detail later), have an electromechanical drive. This has the advantage that their indication is not affected by temporary interruptions to the mains supply, so that no flight data are lost in such an event.

For each flight progress strip the display units are arranged in two horizontal rows, one above the other, but for the sake of convenience we shall always refer to such a strip as a "line". Data can only be fed to the lines under the control of the computer. The flight progress boards are the main output devices of the computer; in addition there are three teleprinters, which print out the

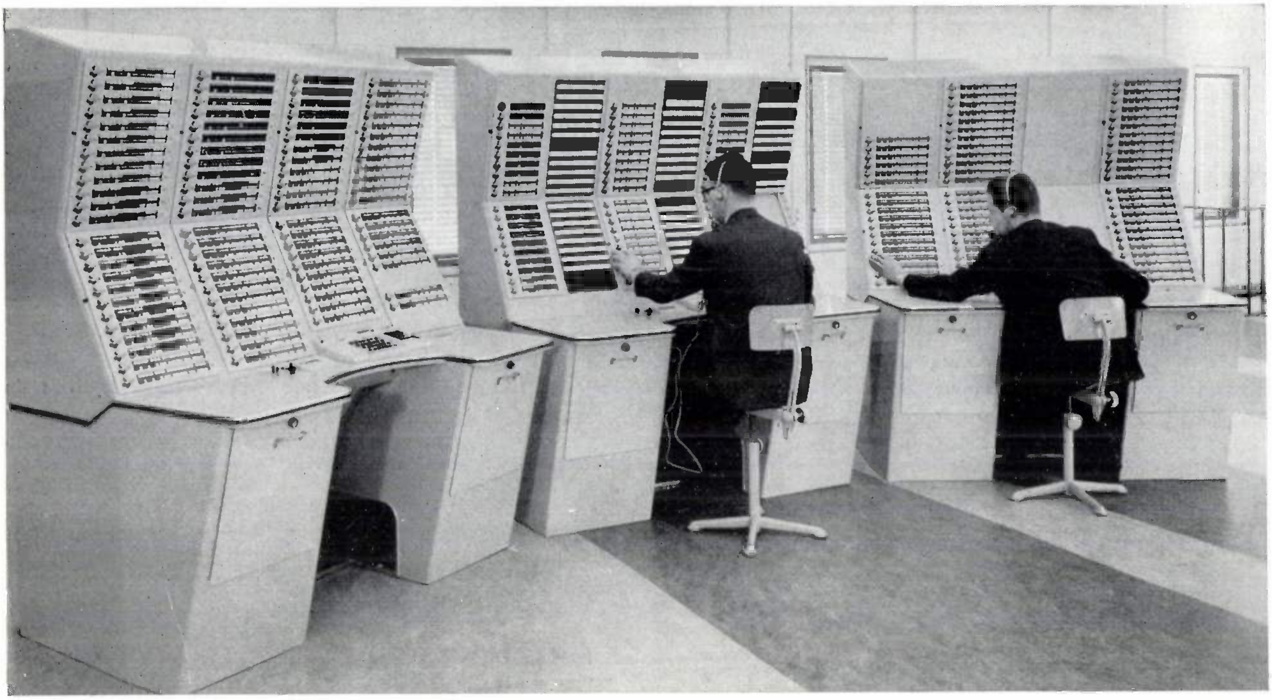


Fig. 2. The new equipment provides the central air traffic control service with three automatic flight progress boards. The boards on the left and on the right permit supervision of the traffic in the air lanes, whilst one of the tasks assigned to the boards in the centre is to control traffic in the vicinity of the airport. Like the stands on the table in fig. 1, the boards on the left and right are subdivided into as many sections as there are reporting positions. The flight data on the progress strips inside these sections are displayed automatically under the control of the computer. These data are automatically kept up to date with all modifications due to changes in the estimated times of arrival or departure, and with cleared and reported flight levels.

flight data on strips at APP, TWR and FIO.

No data will be put out by the computer without a preceding input. Data can be put into the computer by means of one of the teleprinters situated at the assistant control officers' positions, i.e. at ACC, TWR or FIO, or by means of the keyboard on each of the control officers' progress boards.

Fig. 2 shows a third progress board between the East and West boards mentioned so far: this is the *TMA* (*Terminal Area*) board. Adjacent to the lines of automatic display units on this board it is possible to mount flight progress strips produced by two teleprinters which are controlled by the computer. The control officer at the TMA board covers the same area as APP, but the division of duties is such that the TMA officer determines the time schedules and the flight levels for the approach area, while the radar means at the disposal of the APP officer enable the latter to guide the aircraft through the approach area, giving instructions by radio telephone. (The TMA officer only takes over this latter function in the event of radar failure.) However, since these details are of little relevance to a technical description of the SATCO equipment, we shall not enlarge on them any further.

The last equipment item to be mentioned at this point is the real heart of the system, the computer. It is installed in duplicate in a separate technical room.

Before entering into a more detailed description of the various parts of the system and of their interrelation, we shall briefly survey the operational situation. In doing so we must distinguish between aircraft making their departure from Schiphol and aircraft entering the Dutch FIR, either to land at Schiphol or to fly over it and leave the FIR at another point.

The FIO, upon receiving the flight schedule of an outgoing aircraft, will in general feed the details directly into the computer, using the teleprinter provided for that purpose. An assistant control officer at either the East or the West board who receives details of an incoming flight from the traffic control officer of an adjoining FIR by telephone, will also feed these directly into the computer by teleprinter. In both cases we are concerned with a first, or basic, input of flight data.

In feeding basic input into the computer, the control officer always adheres strictly to the following sequence:

a. call sign of the aircraft
b. the letter B, indicating basic input
c. the type of aircraft
d. airport (for an outgoing flight) or first reporting position (for an incoming flight)
e. estimated time of departure or estimated time of entrance into FIR

f. flight level according to flight plan
g. air speed
h. origin of flight
i. air lanes to be followed
j. destination of flight
k. reporting positions
l. closure signal

Upon receiving the basic input, it is the task of the computer to make the necessary data appear on the automatic progress boards and on the strip printers at the TMA board, and at APP, TWR and FIO at the proper time. In many cases the basic input is made rather early, at a time when the flight is still of no direct interest to the control officer, and for this reason the computer waits until 20 minutes before the estimated time of departure or of entrance before starting to process the input data.

When this time is reached, the computer starts making a flight path calculation, determining at what time, speed and height the aircraft will pass its various reporting positions. In order to do this the computer uses the data relating to the type of aircraft (input item c) that are recorded in one of its permanent stores. These data include the speeds of ascent and of descent. After concluding the flight path calculation, the computer selects a free "line" in each of the sections of the progress boards that correspond with the reporting positions the aircraft has to pass and sets up the display units so that the calculated data appear before the traffic control officer.

After calculating the flight path, the computer also makes a search for possible conflict situations, i.e. for every path section it compares the aircraft's flight data with those of all other aircraft in the section and checks whether there is any danger of non-observance of the minimum horizontal and vertical separations. If a conflict situation is found, a red lamp starts flashing alongside the display line associated with the reporting position at the beginning of the path section on which the conflict situation has arisen. By pressing a button provided for the purpose, the control officer can cause the call sign of the second aircraft involved in the conflict situation to appear on a special line on his progress board.

The control officer then examines the situation and decides whether there is a real conflict. If so, he modifies the flight schedules of one or both of the aircraft. The necessary corrections are fed into the computer via the keyboard at the officer's control position.

Corrections must also be introduced if the aircraft is found to be deviating from the original flight plan; this is by no means uncommon especially in the case of times of departure or entrance. For every correction of the flight data the computer makes a new flight path calculation and a new conflict search and displays the modified data on the flight progress boards.

System layout

As we have mentioned before, the main equipment items are the two computers on one hand, and the input and output equipment on the other. The latter comprises the sending and receiving sections of the flight progress boards and the input and output teleprinters. The directions "in" and "out" must always be understood as referring to the computer.

The choice of the manner of interconnecting the various parts of the equipment came from two basic necessities, first that the amount of input and output equipment should be adaptable to any changes in traffic volume without significant changes in the station wiring, and second, that failure of one of the computers should not cause any interruption in service. The result was the connecting diagram shown in *fig. 3.* It will be seen that the exchange of all messages takes place over a single channel, called the main channel. All input and output equipment, the two computers and, finally, a central synchronization circuit have access to this main channel.

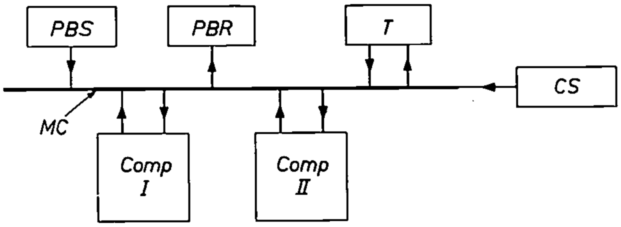


Fig. 3. Schematic diagram of the connections used for the exchange of data between the computers, the flight progress boards and the teleprinters. All messages pass via a single main channel that can be used by only one computer at a time. *Comp* computer; *MC* main channel; *PBS* flight progress board, sending section; *PBR* flight progress board, receiving section; *T* teleprinter connector; *CS* central synchronization circuit.

The main channel comprises fifteen conductors which take the form of coaxial cables. Only one message at a time can be sent from or to one computer over the main channel. The following functions have been allocated to these fifteen conductors:

1 to 7:	message exchange
8 to 11:	synchronization
12 :	calling and answering
13 :	failure reporting
14 :	spare
15 :	"occupied" conductor

Organization of message exchange

For a proper understanding of the manner in which the exchange of messages over the main channel is organized, it is useful to take as our starting point the

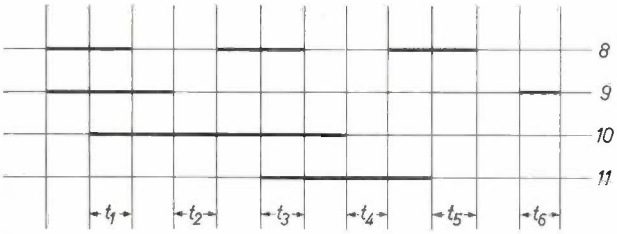


Fig. 4. The main traffic channel in fig. 3 comprises 15 conductros. The impulse pattern shown, which has a repetition frequency of 10^4 c/s, is applied to conductors 8 to 11 inclusive by the central synchronization circuit. It has six characteristic intervals t_1 to t_6. Each of the two computers has its own characteristic interval allotted to it, during which it may attempt to take over the main channel. As a result, the two computers can never do so simultaneously.

moment the equipment is first put into service. When the computers are switched on, they start a programme cycle. During the first part of this cycle each computer checks whether there is any message awaiting transmission to any of the output devices connected to the main channel. As no such message can be waiting at the moment in question, the second part of the programme cycle is initiated automatically. This second part of the programme cycle is used to check whether any of the input devices has a message waiting for the computer.

This check is made via the main channel. A pulse pattern of the kind shown in *fig. 4* is generated in the central synchronization circuit and applied to the synchronization conductors 8 to 11 at a repetition rate of 10^4 c/s. In this pattern there are six characteristic intervals t_1 to t_6. The two computers receive this pattern, and each has a characteristic interval of its own allotted to it for testing the potential on the "occupied" conductor 15. The potential on this conductor indicates whether the main channel is free or occupied, and if, during its own test interval, a computer finds the main channel free, it occupies it immediately by changing the potential on conductor 15, making the main channel inaccessible to the other computer.

Having occupied the main channel, the computer now sends out the call numbers of all the input devices, one after the other, over conductors 1 to 7 to find out whether any of them has a message waiting. Seven conductors are needed for the exchange of messages because the international five-unit code is used for the transmission of letters and figures. Each combination of this code can be read either as a "letter" or as a "figure" and for this reason a sixth bit is added to indicate whether the combination of the first five should be interpreted as a letter or a figure. Finally, the seventh bit is added as a so-called parity bit to provide a check on the correct transmission of the character. The seven bits of a character are not transmitted sequentially over one conductor as is done by the teleprinter, but simultaneously over seven conductors in order to save time.

Call numbers for input devices are composed of the same characters as are used for the messages proper. To ensure that the input devices recognize call numbers as such, the computer therefore applies a short pulse to calling conductor 12 at the beginning of the impulse pattern shown in fig. 4. This pulse is given for every character included in the call number. If an input device which is holding a message ready for the computer recognizes its own call number, it gives an answering pulse on conductor 12 before the pulse pattern ends. On receipt of this answering pulse the computer immediately stops sending any further call numbers. Without further delay the message is then fed into the computer, which disconnects itself from the main channel at the end of the message and starts processing the message at once. This gives the second computer the opportunity to take over the main channel and to continue its own programme.

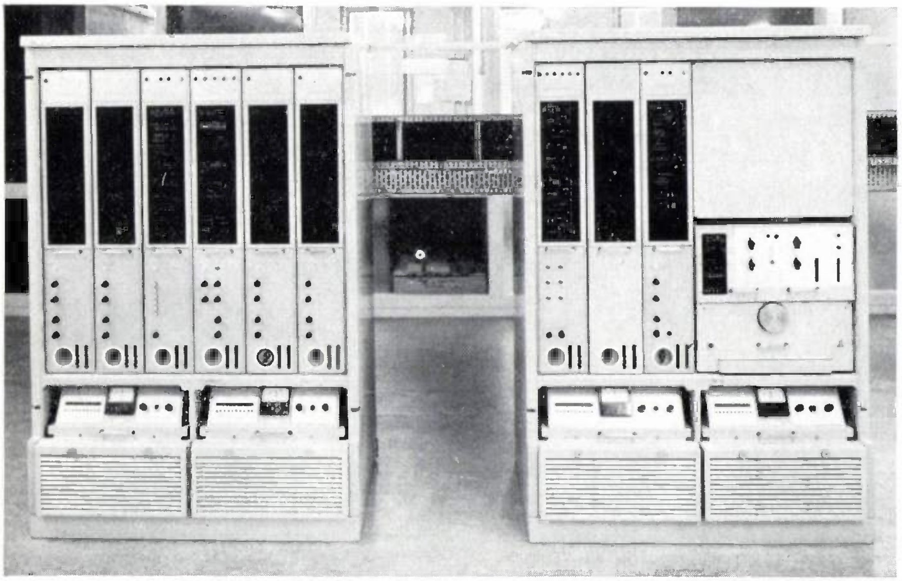


Fig. 5. Each of the two computers is mounted in two cabinets. The cabinet on the left houses the control circuit and the arithmetic unit and also the arithmetic and programme stores. The magnetic drum store is located in the right-hand cabinet, which also contains the buffer circuits for connecting the teleprinters.

On the front panels of both cabinets there are a large number of signal lamps, by means of which it is possible to follow the progress of operations. Under normal operating conditions progress is much too fast for the eye to follow. However, if there is a failure, the computer will stop and the pattern of lamps is frozen. There is a further possibility of using a very low clock frequency for test purposes. This is low enough to permit operations to be followed visually.

If a computer has a message for a certain output device it takes over the main channel in the manner described and sends out the call number of that device in order to find out whether it is free to accept the message. If it is, the device returns a suitable signal and transmission follows. If the device is inaccessible, the computer disconnects itself from the main channel and makes another attempt a little later.

The computers

The two computers, shown in *fig. 5* are identical. Everything we say below about "the computer" must therefore be taken to apply to both alike. *Table I* sums up the main data for the machine.

As indicated in the heading of table I, the computer uses the decimal-binary notation, which means that numbers received are converted to the binary code decimal by decimal, and not in their entirety. For a computer of the type under consideration, which continually has to feed in and out considerable amounts of data comprising both letters and figures, the binary-decimal notation has the advantage that the conversion from the notation used by the machine to that used by the input and output equipment can take place more quickly. Speed of operation is of vital importance for a traffic control machine if it is to carry out all instructions without delay, even under peak traffic conditions. A considerably higher computing speed has been chosen than would be required for the traffic expected at present, in order to avoid having to expand the computer capacity for only a slight increase in the number of input and output devices.

Four binary units or bits are required to represent the ten possible values of a decimal, but four bits, each of which can have either the value 0 or the value 1, can be combined in 16 different ways as *table II* shows, and there is, therefore, considerable freedom in the selection of the combination used. For the Satco-computer the so-called "excess-three" code [1] has been chosen. This code is characterized by the fact that the ordinal of the combination used always exceeds the value of the decimal represented by 3. It has the advantage of being symmetrical with respect to the centre, so that the nine-complement of any decimal can be obtained by substituting 0's for 1's and vice versa. It is easy to arrange bistable registers for bi-polar read-out and with the excess-three code this means that no conversion is required to read out either the decimal itself or its nine-complement. Not only do we gain time in this way, but we can now also use the same adder circuit for both adding and subtracting. Another incidental advantage

Table I

The computer is of the one-plus-one address type, uses decimal-binary notation and addition is parallel-series.

Word length	
Numbers (fixed point)	22 bits; 5 decimals + 1 sign bit + 1 parity bit

Instructions	32 bits;
	4 (+ 2) operation bits;
	10 (+3) address bits;
	7 modification bits;
	10 (+ 4) programme address bits;
	1 parity bit

Clock pulse frequency	660 000 c/s (basic time 1.5 μs)
Cycle times (including time to prepare next instruction)	
Addition and subtraction	24 μs
Multiplication	approximately 300 μs
Division	,, 425 μs
Number of operations per second for a representative programme	30 000
Stores	
Variable store	normally 2048 words of 22 bits
Constants store	normally 1024 words of 22 bits; maximum for the two memories together 8192 words of 22 bits
Programme store	normally 3072 words of 32 bits; maximum 16 384 words
Drum store	60 000 characters of 7 bits each
Buffer stores for input and output devices	Input and output at a rate of 10 000 7-bit characters/s
Input	
Punched tape (5 track)	50 or 75 bauds
Teleprinters Clock (real time) Input keyboards	50 or 75 bauds
Output	
Punched tape	50 or 75 bauds
Teleprinters	50 or 75 bauds
Automatic flight progress boards	for letters and numerals: 40 7-bit characters/s

Table II

Ordinal	Combination	Ordinal	Combination
0	0 0 0 0	8	1 0 0 0
1	0 0 0 1	9	1 0 0 1
2	0 0 1 0	10	1 0 1 0
3	0 0 1 1	11	1 0 1 1
4	0 1 0 0	12	1 1 0 0
5	0 1 0 1	13	1 1 0 1
6	0 1 1 0	14	1 1 1 0
7	0 1 1 1	15	1 1 1 1

[1] For this and similar codes see: R. K. Richards, Arithmetic operations in digital computers, Van Nostrand, New York 1956, ch. 6.

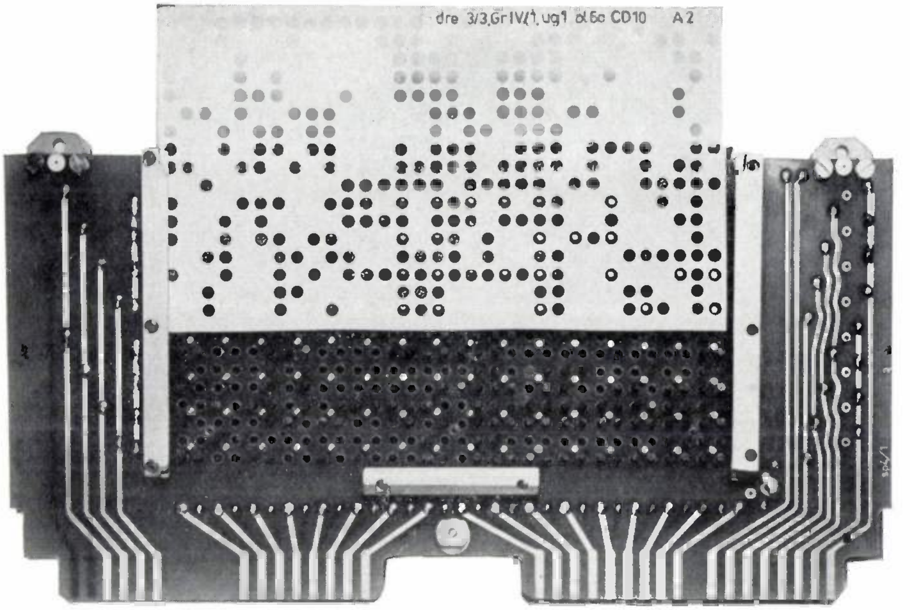


Fig. 6. The programme store is subdivided into matrices of 16 programme lines each. Such a matrix is mounted at the rear of the synthetic resin bonded paper board shown in the picture. A hole has been punched in the board over each ring of the matrix and a small cylindrical permanent magnet can be placed in each hole. Each magnet is magnetized in such a direction that the ferrite core below it is permanently kept in the "0" condition. If an energizing current is sent through the write-in wire threaded through all rings of the matrix, only the non-polarized rings can be brought to the "1" condition. Upon read-out all cores are restored to "0", but the information remains present in latent form. The positions in which the permanent magnets must be inserted are found with the punched card also shown in the picture.

of the excess-three code is that the first five combinations used all have a 0 in the first place, whereas the second five all have a 1. This can be put to advantage in rounding off.

Table I shows that the computer can handle numbers of up to 5 decimals. This means that the accuracy is not particularly high, but as, in air traffic, speeds are measured in miles per hour, heights in hundreds of feet and time in whole minutes, it is perfectly satisfactory for the present purpose.

Like all computers this one includes a number of stores, a control section and an arithmetic section. These will be discussed in some detail in that order.

The stores

Each computer has two high-speed ferrite-core stores, viz an arithmetic store and a programme store. In addition, there are two slower magnetic drum stores for common use by the two computers for the storage of flight data fed into the system.

Contrary to the normal practice in general-purpose computers such as PASCAL [2] the arithmetic and programme memories have here been kept separate. Before an operation in a general-purpose computer can start, the programme must be fed in first, and this is normally stored in the arithmetic store. For a computer like the one used here the programme — made up, as it is, of a large number of sub-programmes — is invariable, although operational requirements may necessitate modifications from time to time. During normal operation all sub-programmes are cycled very many times indeed. With the usual types of ferrite-core store, read-out is destructive, and if the information is not to be lost, it must be written back into the store. If, as is the case here, reading out and writing back are repeated many times, there is a real risk that, at some time or another, a bit will get lost in the process — which would invalidate the programme.

These considerations led to the decision to design the programme store in such a manner that read-out does not cause the information to disappear from the store, but leaves it present in a latent form, so that it can easily be written back. The store is composed of 16-line units. Each unit therefore has 16×32 rings in a matrix which is mounted on a plate of insulating material, a hole having been punched in the plate opposite each ring. As *fig. 6* shows, cylindrical ferroxdure pins can be put into these holes in any desired pattern. After assembly of the unit, these pins are permanently magnetized in such a way that the rings on which they rest are always kept in the "0" condition, even if a current tending to set all rings to the "1" condition is sent through the write-in wire running through all the rings of a matrix — or even of a number of matrices. In this way the programme information can be written back in a simple fashion at any time and without risk of error.

First, the 16 lines of each unit in the programme memory are punched in 8-unit code on paper tape with a *Flexowriter* perforating typewriter. Using this tape, a specially designed translating machine and associated perforator then prepare a punched card like the one shown in fig. 6. This card is put on top of the perforated plate of insulating material and the holes in the card then determine the positions into which ferroxdure pins have to be put. With this procedure any unit of the programme store can be changed in a simple and comparatively rapid manner each time operational requirements call for programme modifications.

The special design chosen for the programme store-units was not however the reason for separating the arithmetic and programme stores. If it had been, the same design would not have been used for the part of the arithmetic store reserved for permanent storage of a number of constants used over and over again in the various calculations. The real reason for the separation was the gain it yields in operating speed.

To appreciate this clearly, one must remember that each programme line contains not only the instruction relative to the operation to be performed, but also the address of the number to be subjected to the operation. To permit the operation to be carried out, this number therefore still has to be read. However, considerations of electrical stability put a limit to the speed at which two successive read-out operations can follow each other in one and the same store. They must, in fact, be separated by a quiescent period which is a multiple of the time required for the read-out itself. If the two successive read-out operations are made to take place in two separate stores then there is no need for this quiescent period.

The two magnetic drum stores we have mentioned are identical in design and in their functions. They are located outside the computer cabinets proper and use the same type of floating heads as the drums in the PASCAL computer [2]. *Fig. 7* shows that the diagram of connections between the drums and the computers is similar to that of fig. 3. From this it can be seen that there is a drum channel which permits the connection of one computer to one drum at a time. The basic data and many values calculated from these data by the computer, relating to all the flights fed into the system, are written into the magnetic drums. A complete track on the drum is available for each flight, and a total of 100 tracks are reserved for this purpose. Next there are 20 tracks in use for the writing in of data which facilitate the speedy recovery of certain flight information. There are also four more tracks used for synchronization purposes. Into each track — i.e. for each flight — 500 characters can be written. As mentioned above, 7 bits are used in the writing of each character, so that each track has a capacity of 3500 bits. One revolution of the drum takes 13 ms.

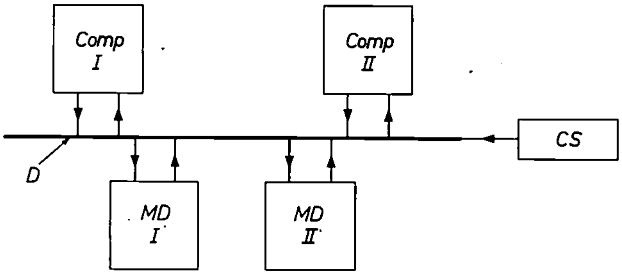


Fig. 7. The data relating to all flights introduced into the system are written into two magnetic drums. The connections for the exchange of data between the two computers and the two drums are similar to those shown in fig. 3 and include a common drum channel. *Comp* computer; *MD* magnetic drum; *CS* central synchronization circuit; *D* drum channel.

Each computer writes the flight data into both drums in succession. In reading out the computer alternates between the two drums. If the drum chosen is out of order, the second one is chosen in its place and a failure-alarm given for the first. This arrangement makes each drum a complete stand-by for the other, and breakdown of either causes no delay.

The control section

In addition to carrying out the computing operations proper, the computer fulfils a number of functions of an organizational character, each of which is laid down in a subroutine. As a result, the machine must very often take logical decisions affecting the continuation of the programme. Programme jumps are therefore quite frequent during the various operations and the machine has consequently been arranged to make jumping easy.

Arrangement as a one-plus-one-address computer facilitates the making of programme jumps. This arrangement means that each instruction read from the programme store contains both the address in the arithmetic store to which the operation indicated applies and the address in the programme store from which the next instruction is to be read. Consequently, successive instructions need not be stored in successive programme lines. It is therefore a simple matter to keep certain programme lines open so that if a jump requires the modification of a programme address, such a modification can easily be made, e.g. by changing one address bit in a fixed location.

The diagram in *fig. 8* shows the principal parts of the control section. One of the 32 bits of each programme line serves as a parity bit for checking whether the line has been correctly written in. On reading out, the remaining 31 bits are transferred to four separate sets of bistable registers. The first four bits determine the operation to be performed, and the corresponding register is for that reason to be found in the arithmetic unit. Next come 10 bits determining the address in the arithmetic store to which the operation refers. These are followed by 7 modification bits, to which we shall return presently. Finally, the last 10 bits, indicating the address of the next programme line to be read out, are stored in the *pa* register.

The contents of the *pa* register are transferred to the *pβ* register. This is necessary because in reading out the programme line determined by these contents, the *pa* register contents are modified. Fig. 8 shows the

[2] W. Nijenhuis, The "Pascal", a fast digital electronic computer for the Philips Computing Centre, Philips tech. Rev. 23, 1-18, 1961/62.

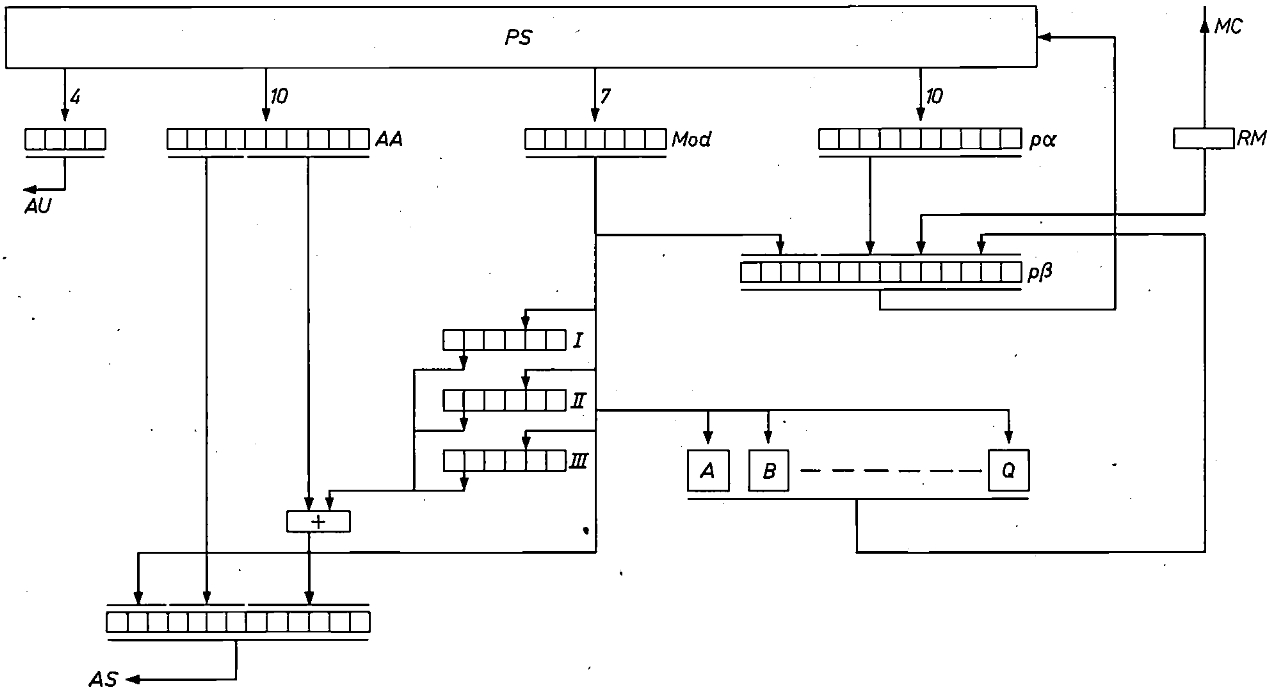


Fig. 8. Simplified diagram of the computer control circuits. An explanation is given in the text. *PS* programme store; *AU* arithmetic unit; *AA* arithmetic store address; *Mod* modification bits; *pα* store address register; *pβ* programme address register; *MC* main channel; *RM* recognition matrix; *I*, *II*, *III* modification registers; *AS* arithmetic store; *A* to *Q* bistable decision circuits.

pβ register to have a capacity four bits larger than that of the *pα* register. Upon examination of table I the reader will find that the number of bits in the programme address is indicated as 10 (+4). This notation results from the fact that the programme store is divided into sections of $2^{10} = 1024$ lines. The programmes are so arranged that successive instructions go on being written into the same section of the store as long as possible. While this situation continues 10 bits are sufficient to characterize a programme line and there is no need to modify the first four bits in the *pβ* register. This need arises only when a jump to another 1024-line section of the memory becomes necessary. The modification is then effected by directing four of the seven modification bits to the *pβ* register.

This device, which has also been applied for the arithmetic store address and the indication of the operation required, has made it possible to restrict the number of bits per line of the programme memory to 32. The possibility arose from the fact that the modification bits are used in only a limited number of instructions.

Modification bits are so named because they are used in the modification of address in the programme and arithmetic stores. Modification of an arithmetic store address may be desirable when a certain programme cycle is to be applied to a series of numbers in the arithmetic store in succession. For the purpose of creating a new arithmetic store address the contents of three index registers *I*, *II* and *III*, which have been provided for this purpose, can be combined with the last six bits of the arithmetic store address under the control of the data contained in the modification register.

Jumps in the programme address can be made to depend directly on a certain result obtained by the arithmetic unit (direct jump), or alternatively, this result can be used to determine whether jumping will be necessary or not after another part of the programme has been completed (delayed jump). In the latter case the result from the arithmetic unit is stored temporarily in one of the bistable circuits *A* to *Q*.

These bistable circuits can also be set directly on the basis of the data in the modification register. This facility is used, for instance, when two programmes *X* and *Y* that are otherwise different, have a certain programme cycle in common. One of the bistable circuits *A* to *Q* is then set in such a way at the beginning of the common cycle that at the end of it the programme is continued in the proper manner.

Each programme cycle completed by the computer is started off by some input message. The heading of each input message includes a significant character defining the nature of the message and the nature of the operation to be performed. From the main channel *MC* this significant character is directed towards a recognition matrix *RM* which analyses it and then puts it directly into the $p\beta$ register. It then becomes the address of the first instruction of the programme cycle to be completed.

The arithmetic unit

We shall not go into the details of the arithmetic unit since the principles followed in its design are no different from those found in other existing computers. All that may need elucidation is the information given in the heading of table I, to the effect that the computer operates on the parallel-series principle. This means that in adding — to which all other operations can be reduced — the various decimals are handled one after the other, i.e. serially. The addition of the binary units in which the separate decimals are expressed, however, is carried out as a parallel operation.

Input and output of data

The computers can take in or put out data via the main channel as well as via the drum channel. In this connection we must mention a problem which arises from the fact that a number of different clock frequencies are used in various parts of the equipment. Table I lists 660 000 c/s as the clock frequency of the computer. This means that the shortest interval between two successive changes of condition in the computer, while it is carrying out an arithmetic operation, is 1.5 μs.

The use of such high switching speeds naturally demands a number of special precautions, e.g. the reduction of the wiring capacitances to a minimum. In the case of the main channel, via which data are conveyed to and from the flight progress boards and the teleprinters, this requirement cannot be met to the same extent as inside the computer. Practical considerations require the computer and the various input and output devices to be located in different rooms and the distances to be covered will depend on the layout of the available building. The maximum distance for which the Satco equipment has been designed is 300 m, and for such a distance the delay caused by the coaxial cables and by the circuits connected to it requires a clock frequency lower than that of the computer.

However, the clock frequency on the main channel still greatly exceeds the frequencies needed for setting the teleprinters or the electromagnetic display units on the flight progress boards. This explains why buffer circuits have been interposed between the main channel and the progress boards or teleprinters. If, for example, the computer has a message to send to the flight progress board, it will transmit it to the buffer circuit for the progress board at the speed determined by the main channel clock frequency. Upon receipt of the complete message the buffer circuit switches to a clock frequency determined by an internal generator and matched to the speed at which the units on the progress board can be set.

Changes in clock frequency also occur in the exchange of messages over the drum channel. We will not enumerate all such changes but merely point out a problem common to all cases mentioned. It is that certain circuits, e.g. pulse counters, are successively controlled by pulses of widely different repetition frequencies. The problem is further complicated by the fact that the generators producing these pulses are not synchronized. Special care must therefore be taken to ensure that no pulses are cut in two when the clock frequency is changed.

A simplified diagram of the circuit used for this purpose is given in *fig. 9*. The pulse patterns applied to conductors A_1, A_2, B_1 and B_2 are shown below the diagram. No phase correlation exists between the A and the B patterns and it must be possible to put either of them through to conductors D_1 and D_2 at will. The signal to switch over from one pattern to the other is given via conductors C_1 and C_2. If the potential on C_1 is high, the A pattern passes, and if the potential on C_2 is high, the B pattern passes. NP_1 to NP_6 inclusive are pairs of "and "gates, whose outputs will be high if all their inputs are high simultaneously. *OP* is a pair of "or" gates, whose outputs will be high if either of their inputs becomes high. H_1 to H_4 inclusive are bistable circuits that are assumed to be "up" if their left-hand outputs are made "high" by applying a pulse to their left-hand input.

The various impulse patterns in fig. 9 have been drawn on the assumption that the C_1 potential is high to start with and that both H_1 and H_2 are "up". Under these conditions the right-hand inputs of NP_5 are high, so that the A pulses pass via *OP* to D_1 and D_2. If now, at an instant chosen at random, C_2 is given a high potential instead of C_1, the next A_1 pulse to follow will pass via $NP_1{}^2$ and bring H_1 "down". This has no effect on $NP_5{}^1$ so that pulse A_1 still passes complete, but it does make the right-hand input to $NP_5{}^2$ low, as a result of which the path of the next A_2 pulse is blocked. This latter pulse is nevertheless effective in bringing down the bistable circuit H_2 via $NP_2{}^2$. $NP_5{}^1$ is blocked in consequence.

H_2 now being down and C_2 high, the first B_1 pulse to follow (this may be an entire pulse or only part of one) is able to bring H_3 up via $NP_3{}^1$. As H_4 still remains down, however, the B_1 pulse will not pass beyond $NP_6{}^1$. but the B_2 pulse which now follows will go straight through because the left-hand input to $NP_6{}^2$ has become high. This B_2 pulse will also bring up H_4, which opens $NP_6{}^1$ for the next B_1 pulse. The series of

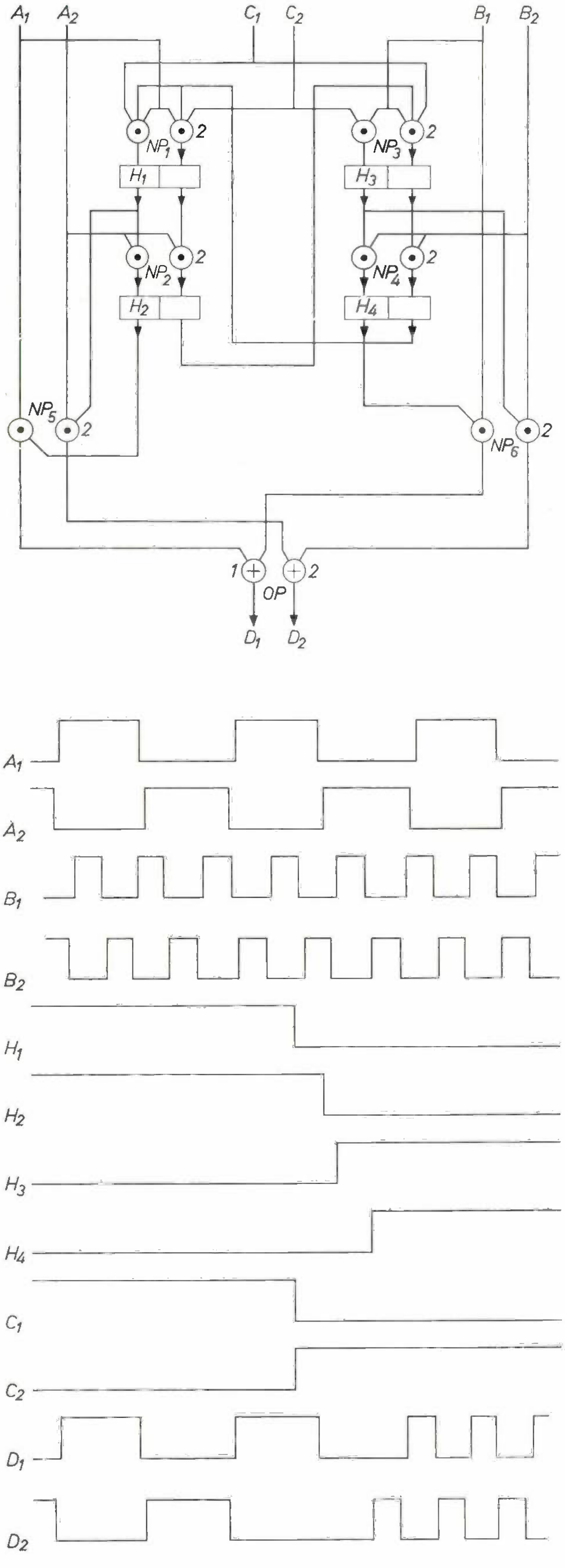


Fig. 9. Simplified diagram of the circuit used to connect either pulse source *A* or pulse source *B* to output terminals *D* without risk of mutilating either an *A* or a *B* pulse on switching over. The operation is explained in the text.

A pulses therefore always terminates in a complete A_1 pulse, while the *B* series always starts with a complete B_2 pulse.

The display board

One of the sections of the flight progress board in fig. 2 is shown in detail in *fig. 10*. For each of the characters making up the lines on the progress boards there is a display unit like the one shown in *fig. 11*. All

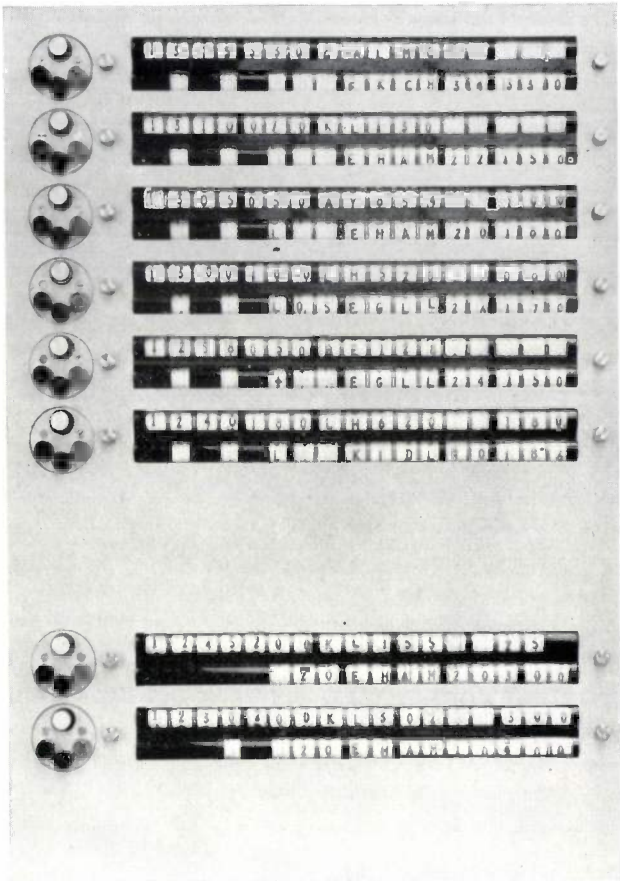


Fig. 10. Detail of a panel on the automatic flight progress boards, showing the individual display units in the different lines.

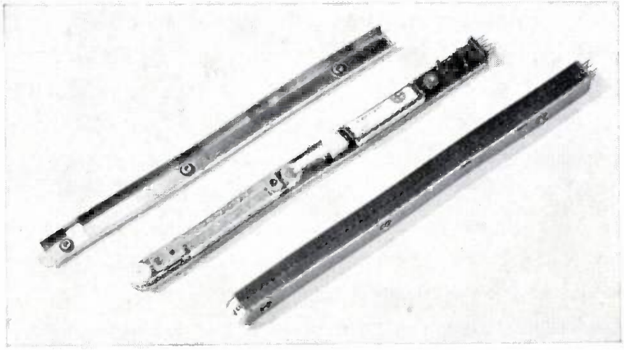


Fig. 11. Display units, in assembled and dismantled form. Letters and numerals are printed on an endless tape which is advanced by an electromagnetic drive. The edge of the tape is cut away according to a coded pattern. Seven brushes can make contact where the edge has been cut away and the combination of contacts made permits verification of the position of the tape.

the necessary characters — up to a maximum of 39 — are printed on an endless tape running over two rollers. A ratchet and pawl mechanism attached to one of the rollers is driven by an electromagnet which, on being energized, advances the tape by one character.

To provide a means of verifying the position of the tape, its edge is cut away in a coded pattern. Seven contact brushes rest on the edge of the tape and each makes contact when a cut-away section of the tape edge passes under it. At any given instant the combination of contacts made by the brushes provides an indication of the position of the tape.

Pulses obtained by half-wave rectification of 50 c/s alternating current are used to energize the magnet that advances the tape step by step. The tape is thus advanced one character every 20 ms, and completes a full revolution in less than 0.8 seconds. The shape of the pulses passing through the magnet coil is shown in *fig. 12.* At points A, A_1, A_2, ... the energizing pulses begin, and at points B, B_1, B_2, ... the armature has again come to a standstill. After each step the tape will therefore be at rest from B to A_1, from B_1 to A_2, etc. During this period of immobility the position of the seven contact brushes is compared with the code combination corresponding to the required character stored in the control circuit. If the two combinations are found to match, the stepping magnet is disconnected.

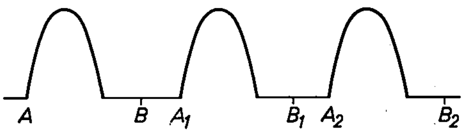


Fig. 12. The magnet coil of the automatic display unit of fig. 11 is energized by pulses obtained by single-phase rectification of alternating current. At points A, A_1, A_2, ... the current is starting and at points B, B_1, B_2, ... the mechanism has come to rest again. During periods BA_1, B_1A_2, etc. the seven code contacts are used to check whether the tape has reached the desired position or not.

Only one line can be set at a time on each of the progress boards, but the 31 display units in a line are energized simultaneously. To keep the wiring as simple as possible, the energizing coils are connected in a switching matrix of the form shown in *fig. 13.* All the 31 coils of a line are switched on by closing the line switch on the corresponding horizontal, and by connecting each of the verticals to earth via a bistable switch circuit. As each of the display units reaches the desired position, the comparison circuit we have mentioned opens the corresponding bistable switch circuits ES. When all units have been positioned, the line switch is also opened.

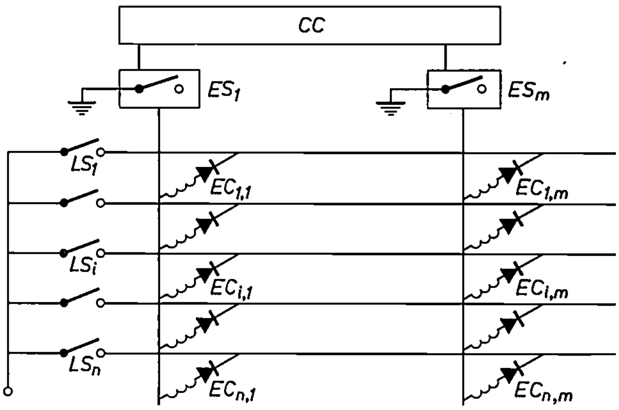


Fig. 13. The energizing coils EC of all display units in a flight progress board are connected in a switching matrix. When the units of line i are to be started, switch RS_i is closed and at the same time all the electronic switches ES_1 to ES_m of the m units of a line. Switches ES are energized from a comparison circuit CC which opens these switches as and when the various units reach their required positions. When all units have been positioned, the line switch opens again.

The requirement of simplicity applies even more to the wiring for the seven contact brushes in each display unit. These are connected in the switching matrix shown in *fig. 14.* On energizing the display units of a line, a second contact on the line switch connects all brush contacts to earth on one side. On the other side these contacts are connected to the verticals. All seven verticals of a display unit are connected to the comparison circuit via a seven-section electronic switch. This switch has 31 positions and scans all 31 display units of a line during the time these units are at a standstill. After each step made by the display units a check is thus made to verify whether they have reached the desired position or not.

Steps taken to ensure reliability

When automatic equipment like that of the Satco system is to assist in the control of air traffic, the safety of this traffic demands that the equipment should meet the most stringent requirements of reliability. Two different aspects of reliability must be considered in this connection.

In the first place, the risk of a complete service breakdown as the result of the failure of a single component or of a group of components must be kept as small as possible. In the second place, no internal or external cause should lead the equipment to produce erroneous results.

As far as the first aspect is concerned it will be obvious that the utmost care has been exercised in selecting the components used, and in determining the conditions under which they are operated. Not only have the various supply voltages been stabilized, but separate voltage control and alarm circuits have been built into the various sub-sections of the equipment. In addition, each equipment section includes an alarm

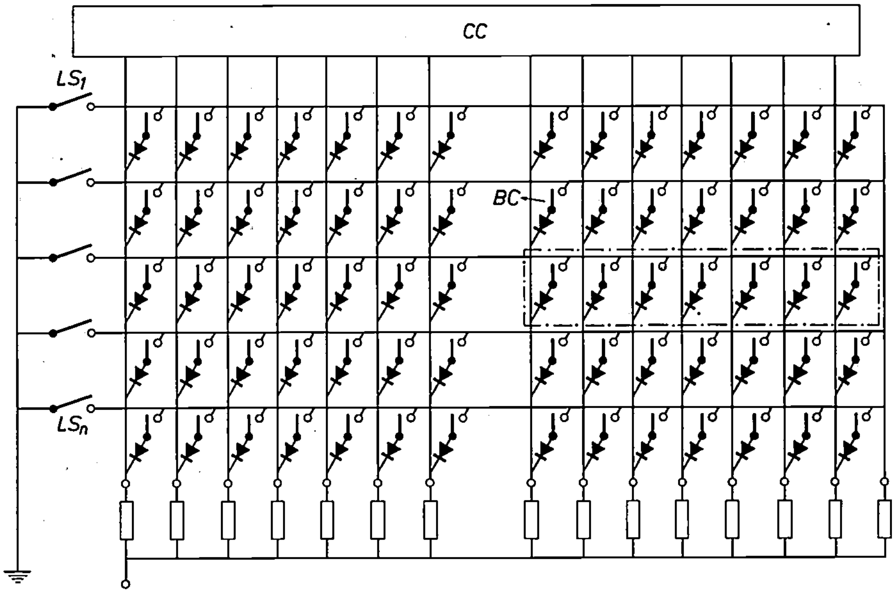


Fig. 14. Like the energizing coils, the brush contacts *BC* of all display units are connected in a switching matrix. When one of the line switches *LS* is closed, the brush contacts of the display units in the corresponding line earth the verticals which are connected to the comparison circuit *CC*. The latter contains a seven-section switch that has as many positions as there are display units in a line. In each position the switch connects a group of seven brushes (one group is shown in a broken-line frame) to the comparison circuit. When the combination of earthed verticals shows that the unit has reached the desired position, the comparison circuit opens switch *ES* in fig. 13.

circuit that comes into operation when the temperature in that section rises above a pre-determined value.

As mentioned before, both the computers and the magnetic drums have been provided in duplicate. To avoid the risk of spare equipment failing to function at the very moment a call is made upon it, all of it is kept in continuous operation.

The principles enumerated above were put to the test in a Satco system of simpler design which was taken into service as early as January, 1961, and has remained in continuous operation ever since. The *mean time between failures* (*MTBF*) of this equipment was found to exceed 1000 hours.

As far as the second aspect of reliability is concerned, the precautions to be taken against the production of erroneous results are of a very different nature. We have to guard, not only against non-systematic errors in the circuits of computers, magnetic drums and input and output devices, but also against the possibility of mistakes being made by the traffic control officers in feeding data into the equipment.

Basically, the provision of two identical computers creates the possibility of having every instruction carried out by both machines and comparing the two results afterwards. However, such a procedure has the serious disadvantage that the amount of equipment involved is more than doubled — two computers, plus the equipment needed for the comparison of results — so that errors also are more than twice as likely to occur. In addition, neither computer would be a full spare for the other, since the failure of one of them eliminates the possibility of making a comparison. Finally, the comparison method also impairs the speed of operation. We may exemplify this by pointing out that many operations require the reading of data from the magnetic drums. As the read-out time depends on the position the drum happens to occupy at the instant the read-out instruction is given, it is not constant and the necessity to make a comparison would always oblige the equipment to wait until the computer with the longest read-out time had produced its result.

Furthermore, results can be compared only between equipment that has been provided in duplicate, and this is not always desirable or feasible. It would not, for example, be feasible for the complicated electronic equipment associated with the automatic flight progress boards.

On the basis of these considerations the checks on the proper operation of the equipment have to a considerable extent been founded on the use of parity bits. Further, a number of special circuits have been introduced for the purpose, and, finally, part of the programme is devoted to it.

In discussing the code used for the transmission of characters via the main channel, we have seen that six of the seven bits used for each character are sufficient to define its meaning, while the seventh bit serves the sole purpose of making the total of 1's in the code odd under all circumstances. An error in the transmission of a bit will always cause the number of 1's in a character to become even. Such an error is easily recognized.

This parity check is repeated at each step in the carrying-out of an instruction. When a number is put into storage, the store circuit checks the parity, character by character. Later, when the number is transferred from the store to a register in the control section, the parity check is repeated. In this way checking is much more intensive than if it were limited to the end of the complete operation. It is an additional advantage that the operation can now be interrupted at each step when an error is detected, for the defective part of the equipment can now be found much more quickly.

As an example of a case where special measures are necessary we may mention code conversion. We have seen that numerals are transmitted via the main channel in a seven-unit code that is based on the international five-unit telex code, while the computer uses the excess-three code to write in each decimal in four binary bits. Each numeral that the computer receives via the main channel is therefore first written into a seven-unit register. It then passes through a code converter and is subsequently put into an arithmetic register with a parity bit of its own.

The parities of the code groups before and after conversion are entirely unrelated and the value of the parity bit has therefore to be determined anew after the conversion. As a special safety precaution, entirely separate circuits have been provided for the determination of the new code group and of the new parity bit directly from the original code group. Should either of the two circuits fail, this will show up at the next parity check.

A second example of special checking arrangements is found in the buffer circuits of the teleprinters for introducing data into the computers. A buffer circuit receives the characters sent by the teleprinter at its characteristic speed and stores them in a number of registers. When the message has been fully received, the buffer circuit tries to establish connection with the computer via the main channel and, when it has done so, it transmits the message at the speed which is characteristic of the main channel.

Each character the teleprinter transmitter sends into the buffer circuit is returned from its buffer register towards the teleprinter receiver. The machine therefore prints its own message and this provides the control officer with a check not only on his own actions, but on the correct operation of the teleprinter and on the receiving register in the buffer circuit.

The last type of precaution to be mentioned is the provision of cycles in the programme for testing the accuracy of input data. When discussing the example of a basic input, we saw that this follows a fixed pattern. When the computer receives a message of this type, it does not start processing it at once, but begins with a programme cycle for the purpose of checking whether the message follows the prescribed pattern and whether it contains unacceptable data, e.g. an erroneous indication for a reporting position. If the computer can accept the message, it confirms the fact to the originating teleprinter. The control officer has to await this confirmation before he may put in a new message. If the message is found to contain an error, this is signalled to the control officer. The message is rejected by the computer and the officer has to repeat it in corrected form.

The various safety measures described have resulted in the elimination of any risk of equipment errors impairing the safety of air traffic.

Summary. The object of the equipment described is to carry out automatically a number of operations in civil air traffic control. These include the making of flight path calculations and searching for possible conflict situations. Calculations are made by an electronic computer provided in duplicate. Results are displayed on automatic flight progress boards. Basic flight data are fed into the computer by teleprinter; for modifications the keyboards on the progress boards are used. Output data appear on the progress boards and on strip printers.

As there is a great volume of alpha-numeric input and output data, the computer stores numbers in binary-decimal notation in order to save time. Addition is done in parallel for the bits of each decimal and in series for the various decimals. Programme instructions are of the 1-plus-1 address type. Successive programme instructions therefore need not be stored in consecutive memory lines, but places can be kept open so that jumps can be effected by very simple changes, e.g. of a single address bit.

Each computer has two high-speed ferrite core stores; two magnetic drums are used by the two computers in common. One of the core stores is the arithmetic store, the other the programme store. As a result, an instruction and the number to which it refers can be read in very quick succession from the two stores, thus saving time. Apart from occasional modifications for operational reasons, the programme is invariable. As it has to be read out and written back very many times, the memory has been so designed that reading out is not completely destructive. The information is retained in latent form and can be written back without any risk of error. A description of the design is given.

Messages are exchanged between the computers and the input and output devices via a single channel, the main channel, which only one computer may use at a time. As a result, the volume of input and output equipment can be varied without modification to the wiring. One computer will carry the entire traffic load if necessary.

Flight data are made to appear on the progress boards by means of display units in which an endless tape, carrying letters and numerals is driven by an electromagnetic ratchet-and-pawl mechanism; in the event of a mains failure no flight data are lost. The edge of the tape is cut away in a coded pattern and seven contact brushes permit remote checking of the position of the tape.

A number of safety measures are described that are taken to keep the reliability as high as possible. These measures guard both against equipment failures, and also against the production of erroneous results.

Automatic telegraph exchanges with electronic stores

P. Harkema

621.394.614.4

Although the telegraph as a medium for the electrical transmission of messages is of considerably earlier origin than the telephone — see the last article in this issue — it was not developed on an extensive scale until a much later date than the telephone. There are two main reasons for this.

The first is that initially there was no telegraph instrument which could be easily operated by members of the public. This situation remained unchanged until the advent of the modern start-stop teleprinter.

In the second place, the telegraph has been much more a long-distance and international traffic medium than the telephone, at least in the early stages of telephone development. The great development of transmission technology did not come until the telephone, with its far greater market for traffic, had created the economic basis on which telephone carrier systems could be founded. By superposing carrier telegraph channels on carrier telephone systems, inexpensive transmission paths became available to the telegraph as well.

Once these two handicaps were removed, the telex service was quickly developed alongside the public telegraph system. Subscription to this service enables private persons or organizations to have installed on their own premises individual teleprinters on which they can exchange messages directly with other subscribers.

The ease and cheapness of private message traffic by telex service caused such a rise in the traffic that it proved possible to reduce the telephone services' lead in another respect, that of automation. Automatic switching of telegraph connections is basically no more difficult to achieve than that of telephone circuits, but because there was little traffic in the public telegraph system the groups of lines had remained small, while traffic theory proved conclusively that the efficiency of a system in which small groups are switched by automatic devices must inevitably be low.

Automation therefore first appeared in the telex service. In common with telephone systems, in automatic telex networks a direct connection over which traffic in both directions is possible is always built up between calling and called party. At first glance this may seem surprising because telex traffic is typically unidirectional, the aim being to convey a message from caller to called party: a reply is not usually expected until later, when a connection has to be made in the opposite direction. The channel in the return direction is therefore used very uneconomically but there are two reasons why this is unavoidable. The first is that it must be possible to transmit over the return channel a signal confirming that the connection has been established as far as the called party, and the second that the caller has to assure himself of the identity of the other party before he starts transmitting his message. Allowance must be made for the possibility that either through the mutilation of a character or through an error of the caller himself, connection is made to someone other than the desired subscriber. This uneconomic use of the return channel is justified if there is sufficient traffic to make the efficiency of the telegraph channels as a whole satisfactorily high. There is, however, an important category of telegraph networks in which that condition is by no means satisfied and that is the category of closed networks belonging to official, semi-official and private organizations, in which the only traffic possible is between the offices of the organization concerned. In general such networks are characterized by relatively large numbers of line groups, each of which is very limited in size, often consisting of not more than one channel.

It will be clear from the above that for the last type of network automatic switching from the calling to the called party is uneconomic and that even the provision of a very moderately used return channel for every circuit cannot be justified. Such networks are in the same position now as public telegraph neworks were in the early days when connections were built up by circuits extending from one office to another, with a transmitter at the sending end and a receiver at the other. Transmission was on an office-to-office basis, the message being noted at the incoming line termination of each office, then taken to the transmitter of the next outgoing line and retransmitted from there, and so on, until the destination was reached. Retransmission was subject to delay if there were other messages awaiting transmission.

Awkward as this method of operation, so described, may sound, it nevertheless contains a number of essentially valuable elements. For instance, it cancels the need to hold return channels available for signalling and identification and these can therefore be used for normal traffic in the return direction. Identification is less necessary because even if a message arrives at the wrong office, the latter, being part of the same organization, is obliged to ensure that the message is forwarded to the correct destination.

Ir. P. Harkema is on the staff of N.V. Philips' Telecommunicatie Industrie, Hilversum.

A very important element in this mode of operation is that good line efficiency is obtained at the price of some delay in forwarding. The introduction of delays in retransmission from the intermediate offices is possible because the message is taken down at the "incoming" instrument and thus preserved. The situation can be summarized by saying that reasonable efficiency is attainable if delays are accepted and the storage function is introduced at junctions.

By adhering to these principles it has even proved possible to automate closed networks with their small groups of lines. In automated networks of this kind it is not, as in telephony, the sections of line which are switched through. Instead, the messages themselves are, so to speak, switched through from office to office. These two methods of co-operation have respectively come to be known as "line switching" and "message switching". The section of this article which now follows will be devoted to a description of systems developed by Philips for message switching.

Message-switching exchanges

Philips were called upon to develop, for one particular organization, the "Société Internationale des Télécommunications Aéronautiques," (SITA), a series of exchanges with automatic message switching, in which each project embodied a greater degree of automation than the project which had preceded it.

These projects will now be discussed in turn, to bring out in a logical sequence the various problems that had to be solved.

SITA is an organization which operates a telegraph network for the common benefit of a large number of airline companies. The majority of messages transmitted over this network are concerned with the reservation of aircraft seats. A peculiarity of this traffic — and one frequently found in other closed telegraph networks — is that a not inconsiderable proportion of the messages are intended for more than one addressee. A message of this sort is necessary if, for instance, several booking offices have to be informed that no more seats are available on a particular flight. It was therefore necessary to devise a satisfactory method of automatic dissemination of multiple-address messages.

Original manual operation

In its original form the SITA network consisted solely of sections of line from office to office without any means of switching through. It was operated in the way outlined above for the older telegraph networks, although the task of SITA operafors was lightened by the use of reperforators, i.e. receiving perforators, at the incoming circuit terminations, and of automatic punched-tape transmitters at the outgoing ends. Transcription errors were thereby eliminated and the work of the operators was limited to tearing off and reading the received messages and transferring the tape to the transmitter associated with the outgoing line of the station of destination. The tapes were placed on racks (colloquially called "washboards") in order of arrival and, when their turn came, inserted in the automatic transmitter (see block diagram in *fig. 1*). For multiple-address messages a copy for each of the addresses was made on a teleprinter in a local circuit and these copies were then taken to the appropriate transmitting positions.

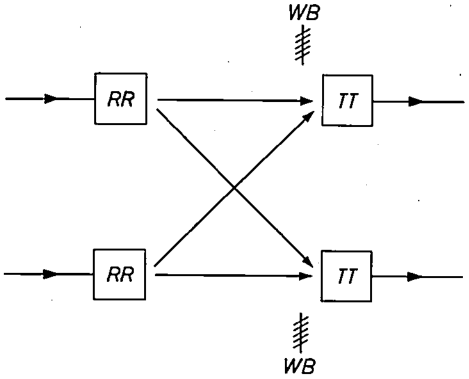


Fig. 1. Diagram of a manually operated torn-tape telegraph exchange in which incoming lines terminate on reperforators *RR* and outgoing lines are connected to automatic transmitters *TT*. The tapes have to be taken from the receivers to the transmitters. If a tape cannot be put into a transmitter immediately it is placed in a waiting bay *WB*.

In the large exchanges of the SITA network such as London one side of the telegraph room was lined with reperforators and facing these was a row of automatic transmitters. The operators walked continuously to and fro between the two rows, transferring punched tapes from one to the other (see *fig. 2*).

The first step towards automation

By 1954 the traffic, the number of incoming and outgoing circuits and the number of operators needed in the London exchange of the SITA had increased to such an extent that the method of operation began to give rise to difficulties in the form of excessive message delay, while there was also an increasing danger of message tapes getting lost.

When consultations concerning a possible solution of this problem started between British European Airways, who operate the London SITA exchange, and Philips, the introduction of complete automation would have called for changes in operational procedures. This required protracted international discussion, however, and since there was an urgent need to surmount existing difficulties, it was decided to adopt a temporary solution.

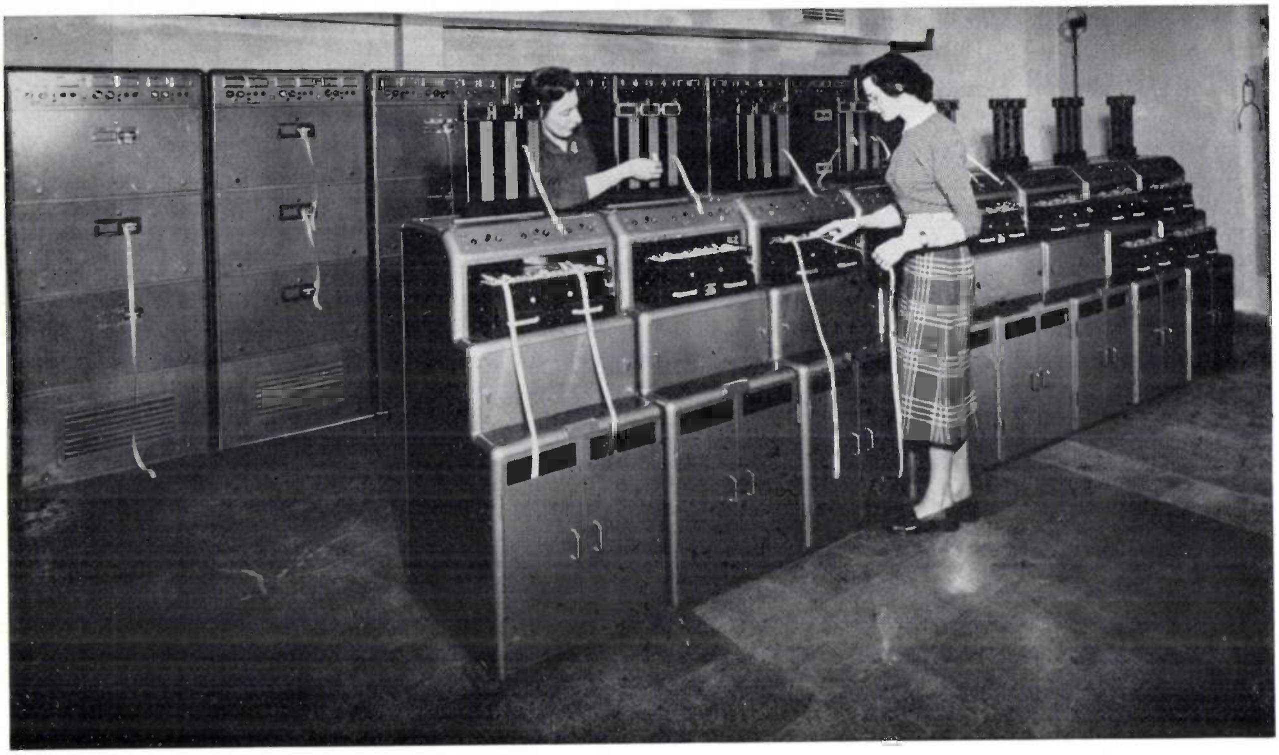


Fig. 2. Photograph of a torn-tape exchange based on the principle of fig. 1. The reperforators are in the background, with the transmitters and waiting bays in front.

The need for international co-operation before taking such a step will be appreciated if we remember that under the method in use until then, the message heading, i.e. the section in which the address appears, was always read by an operator, so that a certain amount of freedom in its make-up could be allowed without necessarily leading to handling errors. The international code designations of the various offices of destination were not suitable for automation. The same was true, *mutatis mutandis*, of the sequence in which the information had to appear in the heading. And, finally, to be suitable for automatic transmission the messages had to contain characters indicating not only the end of the address section but also the end of the message as a whole. Once agreement had been reached on these points, the staffs of all the offices would have to be trained in the new procedure, which would have to be followed strictly.

The provisional solution chosen for the London exchange was basically very simple, consisting merely, as shown diagrammatically in *fig. 3*, in disconnecting the transmitters from the outgoing channels and inserting an automatic selector. This selector was operated by means of a pushbutton panel at each transmitting position. The operator had only to load the tape into the transmitter and press the button corresponding to the appropriate outgoing route. The selector then located itself on the contact to which the outgoing line was connected and the message went out. If the line was not free the selector waited until it was, then positioned itself without further delay.

Despite its simplicity, this solution brought considerable improvement. As all outgoing circuits could be reached from any transmitting position, it was possible, as *fig. 4* shows, to alternate the transmitting positions and the cabinets containing the reperforators. This arrangement enabled the operators to attend to the reperforators on both sides of their positions without getting up from their chairs. Multi-address messages could be inserted as often as necessary in the same transmitting head, so that the need to prepare copies of tapes disappeared. All the former comings and goings and the inherent confusion and delays were eliminated, operator efficiency increased by 25 %, and delays fell to a very low average value.

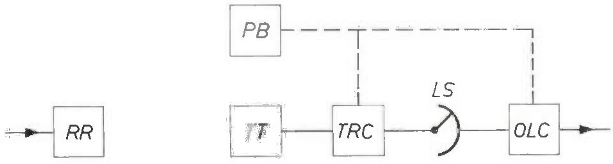


Fig. 3. Diagram of a torn-tape telegraph exchange in which automatic transmitters *TT* are not connected to outgoing line circuits *OLC* directly, but via line selectors *LS*. For each transmitter there is a transmitter circuit *TRC* and a push-button panel *PB*. The desired route is selected on the latter and selector *LS* then automatically finds a line for that route.

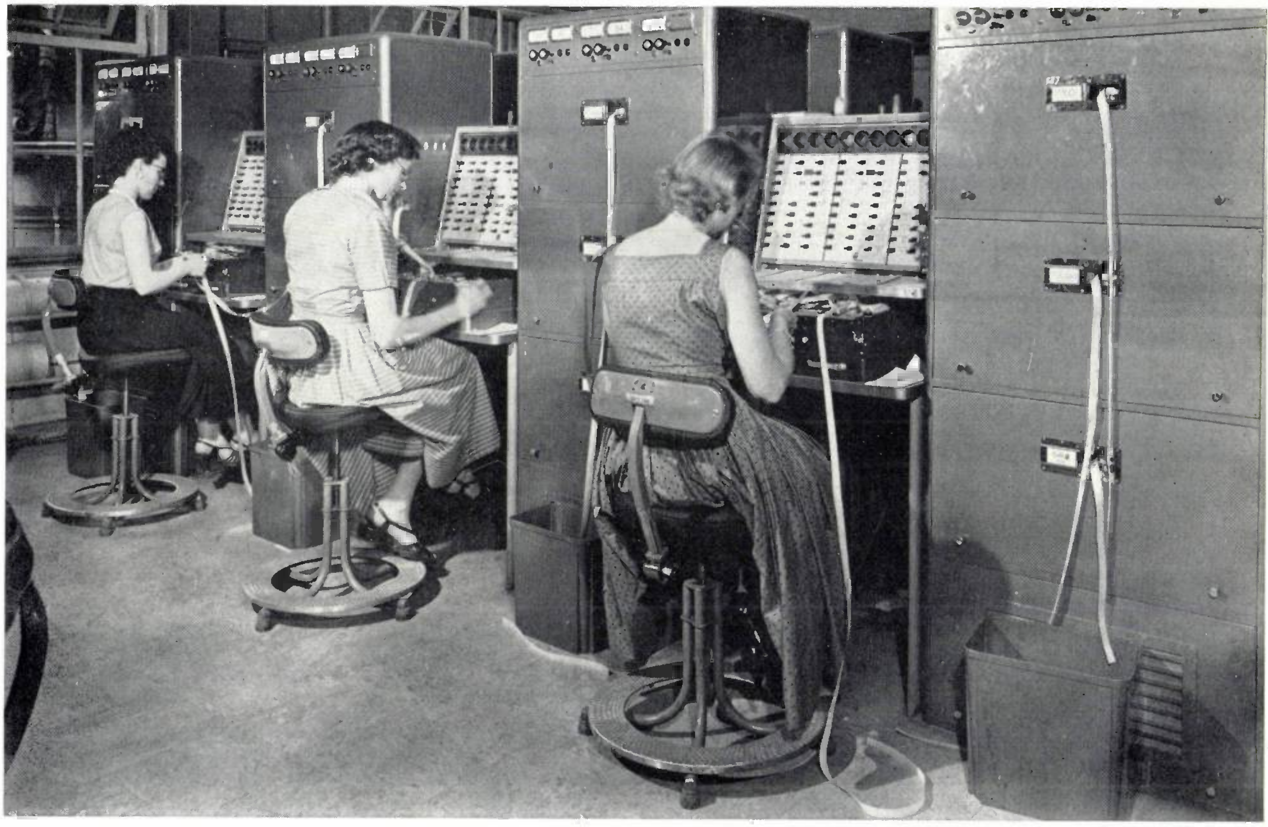


Photo BEA, London

Fig. 4. Photograph of an exchange based on the diagram in fig. 2. The receiver racks are now in between the operators' positions with their automatic transmitters. The operators no longer need to leave their positions in order to transfer tapes from the reperforators to the transmitters.

Introduction of electronic storage

The SITA exchange in Paris, which was operated by Air France and originally also had its transmitters and receivers connected permanently to lines, reached the same state of overloading shortly after the London exchange, so that the consultations concerning renovation of the two almost overlapped.

In Paris as in London the time was not considered ripe for full automation, but the simple solution adopted for London was not suitable in Paris. This was because the Paris exchange had approximately twice as much traffic to handle as the London one, so that certain limitations inherent in the London solution might have led to difficulties. In addition, a more permanent solution was wanted than in London, where it was proposed to replace the temporary method of operation by a more automatic one five or six years later.

The main limitation in the London exchange consisted in the fact that if a desired outgoing line or group of lines was busy, messages had to wait in the transmitting heads until a line became free. This meant that in the event of a sudden surge of traffic for a particular outgoing route, there might be so many messages waiting in transmitting heads that traffic in other directions would be brought to a standstill. Although certain organizational measures could have been taken to cope with this problem, it was feared that it would give rise to difficulties at the much larger Paris exchange. The number of automatic transmitters per operating position could not be increased without the operators having to walk to and fro, a situation which had to be avoided at all costs. A solution was therefore sought in the introduction of intermediate stores to which messages could be transferred from the transmitting head if the outgoing circuits were engaged, and where they could be held until the lines were free again.

A conceivable form of intermediate store is a device using perforated paper tape. Apart from the not inconsiderable consumption of paper, such a device has the disadvantage that repeated transmission of, for example, the same multi-address message may give rise to difficulties. Ferrite core memories were therefore chosen instead. As the shortest element in a teleprinter character is 20 ms long, it was possible to use fairly large cores with an outer diameter of 3.5 mm, thus simplifying not only the wiring but also the writing and reading amplifiers.

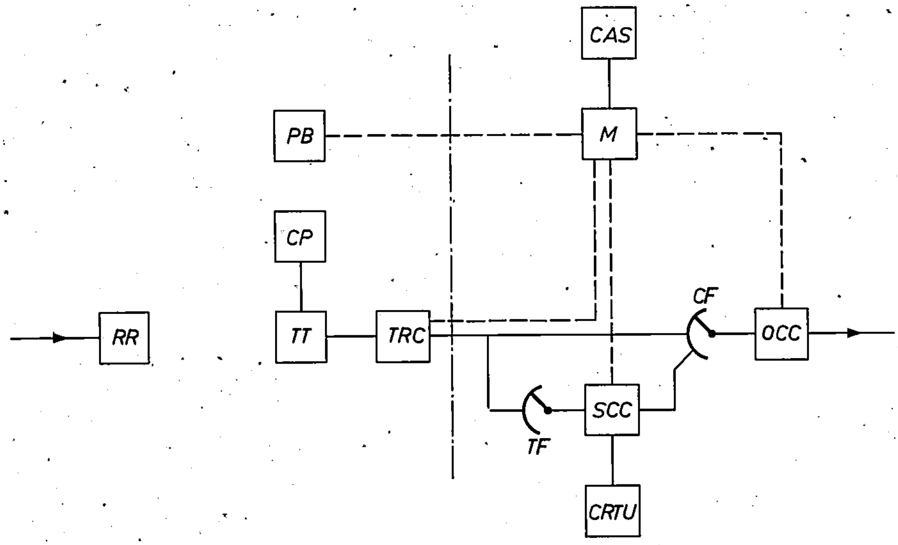


Fig. 5. Diagram of an ES-1 automatic torn-tape system as used in the exchange for the SITA organization in Paris. Messages can be held temporarily in electromagnetic stores if transmission to a free line is not immediately possible. For a description of how the system operates see the text. *RR* reperforator; *TT* automatic transmitter; *CP* control panel; *PB* pushbutton panel; *TRC* transmitter circuit; *CF* connecting circuit finder; *OCC* outgoing line circuit; *TF* transmitter finder; *SCC* secondary connecting circuit; *CRTU* (combined receiving and transmitting unit): electronic store with writing and reading amplifiers; *M* marker; *CAS* central address store.

Fig. 5 shows a simplified trunking diagram of the SITA exchange in Paris, and *fig. 6* a photograph of an operator's position. It can be seen from the diagram that the incoming circuits still terminate on reperforators. The operator tears the length of tape carrying a complete received telegram from the receiver *RR* and inserts it in an automatic transmitter, *TT*. She reads the address(es) and presses all the appropriate routing buttons on the panel *PB* visible in fig. 6, finally pushing the start button on control panel *CP*. On receiving this signal the relay store associated with the panel *PB*, in which the selection information has been temporarily stored, establishes contact with marker *M*. This is a central device which operates so fast that the waiting time for *PB* is only a fraction of a second. Once contact is made, the selection information and the identity of the automatic transmitter are conveyed to the marker. The push-button panel is then freed for another connection.

The description which now follows refers to a multi-address message, but the simplification for a message with a single destination will at once be clear. The marker has connections to all the line circuits *OCC*, of which there is one for each outgoing channel. The marker can therefore test whether there is a free channel on the routes selected. Any line circuits which are free are immediately seized by the marker, but if for one or more of the routes, there is no free line available it also seizes a free connecting circuit *SCC* and the associated store *CRTU*. Selectors *CF* and *TF*, which are associated with the line circuits and the intermediate store respectively are positioned in parallel on the calling transmitter by the marker, which then withdraws from the connection. Once the transmitter starts, the message is forwarded simultaneously to the channels which have been found free and to the intermediate store. The effect so far as the operator is concerned, is that each message leaves her operating position immediately, and is held, if necessary, in an intermediate store which can be located clear of her operating position so as to allow her complete freedom of movement.

If it proves necessary to call in the assistance of an intermediate store, the marker, before releasing, finds a connection to another central device, the central address store *CAS*. The routes on which the message will still have to be transmitted, and the identity of the intermediate store in which the message is held, are recorded in the central address store by the marker.

Every line circuit of a route which becomes available again after the transmission of a message sends the marker a signal which causes the latter to enquire of the central address store whether there is a message for that route waiting in an intermediate store. If there is, the marker holds the line and positions the relevant selector on the intermediate store concerned. If by chance further lines become simultaneously free for the message then further selectors are positioned on the intermediate store. Having set up these connections the marker withdraws from the connection again. This cycle is repeated until the message has been transmitted on all the desired routes.

The number of intermediate stores to be installed obviously has to be calculated on the principle that even during the heaviest traffic there must still be capacity available for the storage of messages. On the other hand, economy demands that the traffic channels, which are often very expensive, should be loaded to the maximum possible extent, and loads of 80% can in fact occur. Such high loading of a traffic route, which may only have one channel at its disposal, inevitably results in considerable delay in forwarding messages. This delay might cause messages to pile up in interme-

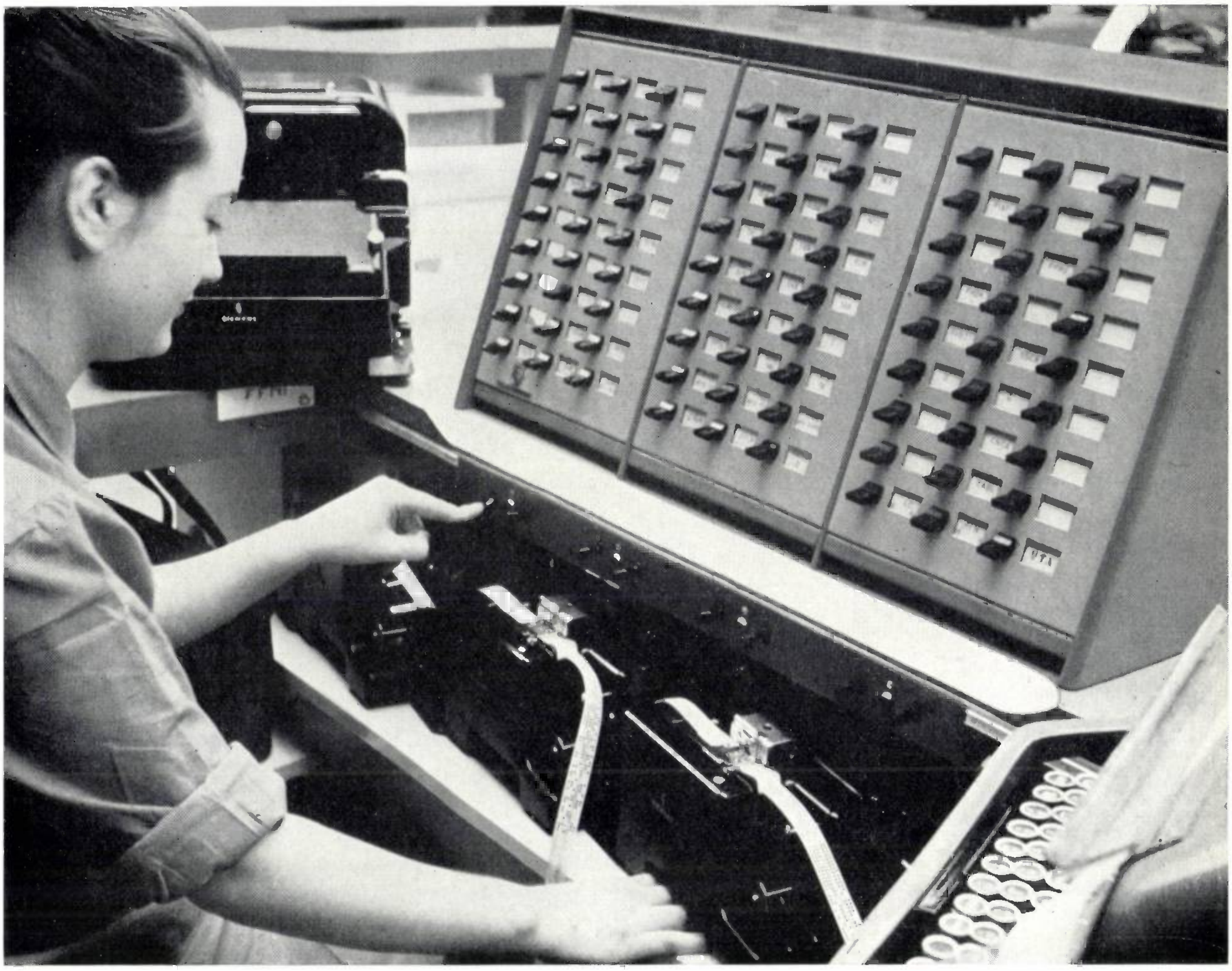


Fig. 6. Operator's position for an exchange based on the arrangement in fig. 5, as adopted in the SITA switching centre in Paris. In front of the operator are four tape transmitters, and above them, the keys that operate them. At the top is a pushbutton panel which is common to all transmitters and has a button for each outgoing route. The reperforators are arranged in two tiers on each side of the operator.

diate stores, so that a sudden traffic peak on a busy outgoing connection could produce a shortage of intermediate stores.

To be able to cope with this situation in all circumstances it would be necessary to install an uneconomically large number of stores. This problem has been solved by providing the line circuits of very heavily loaded circuits with their own stores, at the rate, as *fig. 7* shows, of two stores per line circuit. Each of these stores has its own selector *CF* enabling it to connect with a transmitter circuit *TRC* or a connecting circuit *SCC* with an intermediate store. This means that, although messages cannot always be forwarded at once, the route, as seen from the switching centre, has doubled its acceptance capacity, so that overloading of the central group of stores is avoided. When messages are retransmitted from the stores of the double line circuits the line is switched from one store to the other by contact *I* after the departure of each message, so that no message has to queue for a disproportionately long time. If both line stores are empty, the line is connected direct to a finder *CF* by contact *II*.

As the stores in the line circuits do not have to satisfy the requirements for repeated transmission of the

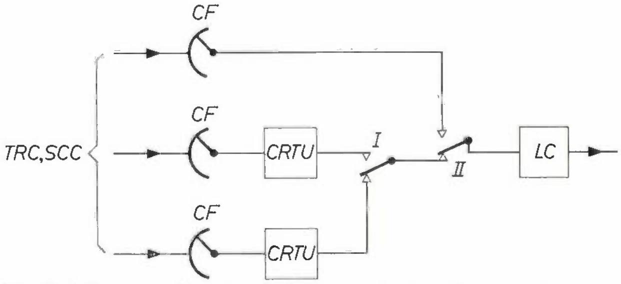


Fig. 7. Diagram of a circuit for an outgoing line carrying very heavy traffic. Circuit finders *CF* have the same function as in fig. 5. When traffic is not heavy, contact *II* is in the upper position and messages can be put straight on to the line. As soon as the line is busy, the circuit finders associated with the *CRTU* electronic stores come into operation. Outgoing messages are not then sent direct to the line but are stored temporarily. Contact *II* now occupies the position shown while contact *I* is switched alternately to the top and bottom store. This ensures that no message will accidentally have to wait an excessively long time in either store.

same message, ferrite cores are not necessarily called for. In the Paris exchange, however, they proved an economic solution to the problem of providing the desired capacity of 2000 characters. In the semi-automatic London SITA exchange, to be described below, a larger capacity was required and it was more practical to have mechanical devices employing perforated tape. A device with a capacity of 40 000 characters using magnetic tape has since been developed by Philips for the same purpose: this is described in a separate article on p. 250 of this issue.

The second London SITA storage

The use of electronic stores in the Paris SITA exchange is limited to the outgoing side of the exchange, the storage function on the incoming side being still performed by perforated tapes. The time taken to receive the message in its entirety on tape, to tear it off by hand, read the address and insert the tape in the transmitter, and to press routing and start buttons constitutes the minimum possible delay in forwarding messages and cannot be avoided. This form of human intervention was, however, still considered desirable because it provided the opportunity to rectify imperfections in the message format.

This principle had also been valid when the first, provisional, London SITA exchange was built. By the time the second exchange was ordered, international consultation had made such considerable progress that agreement existed on a standardized layout of the address section of messages. Nevertheless, it was decided, after due deliberation to carry automation much farther than in Paris while retaining a certain degree of human intervention because at that time the layout of the addresses was still not completely decided.

Full automation has been adopted only in the recent expansion of the exchange. The new part operates com pletely in parallel with the old, so that offices linked to the exchange can be connected to either. The simplified trunking diagram of the section first installed is shown in *fig. 8*, while a double operating position can be seen in *fig. 9*. Examination of the latter shows that a large pushbutton panel is fitted in the table top and that the vertical part of the position houses two teleprinters placed behind glass to deaden the sound. All the reperforators and transmitters have disappeared.

This can also be seen in fig. 8, which shows that the incoming channels terminate on line circuits *ICC*, connected on the switching centre side to a contact in the arc of a linefinder *LF*.

A message may come in on any incoming line without prior warning. Its text, however, is always preceded by the letters ZCZC. These letters, which have no inherent significance, are used to start the automatic equipment, so that linefinder *LF*, to which an incoming connecting circuit *RCC* with a ferrite-core store is connected, has time to position itself on the contact of the calling line. *LF* is a high-speed selector, the positioning of which is fully completed before the significant part of the message comes in. This part is then conveyed to *RCC* and recorded in the store.

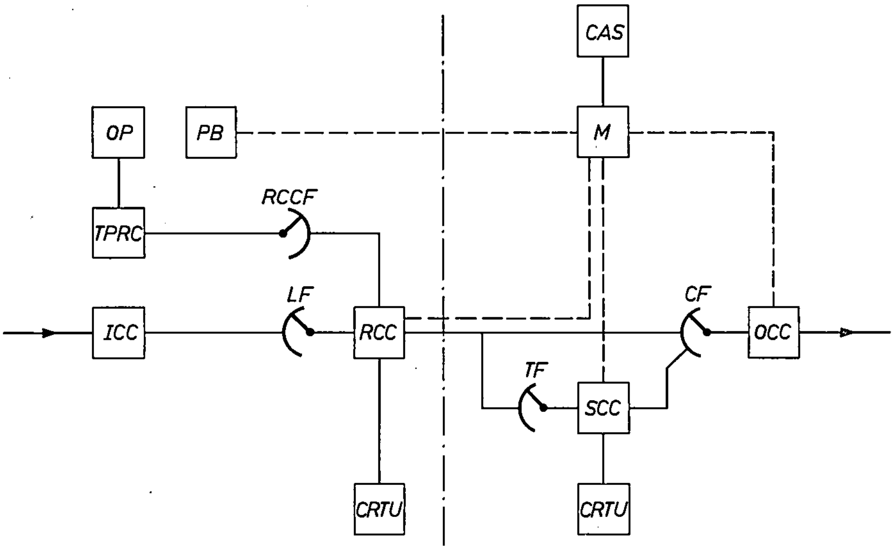


Fig. 8. Diagram of a tapeless exchange in which the address of destination has still to be read off by an operator, who must then press a button on board *PB* for each appropriate route. For a more detailed explanation see the text. *ICC* incoming line circuit; *LF* linefinder; *RCC* incoming connecting circuit; *RCCF* incoming connecting circuit finder; *TPRC* teleprinter circuit; *OP* operators teleprinter; *PB* pushbutton panel; *CF* connecting circuit finder; *OCC* outgoing line circuit; *SCC* secondary connecting circuit; *CRTU* (combined receiving and transmitting unit): electronic store with writing and reading amplifiers; *TF* transmitter finder; *M* marker; *CAS* central address store.

After the linefinder has been positioned, the incoming connecting-circuit *RCCF* starts hunting. *RCCF* belongs to a free teleprinter circuit *TPRC*, with which a teleprinter *OP* is connected at an operator's position. Once this connection is established, read-out of the message from the store begins and the first lines appear on the operator's teleprinter. As soon as the whole of the address part has been printed, the operator presses one or more routing buttons on her panel *PB*, finally pressing the start button of the appropriate teleprinter. No more of the actual text of the message is passed on

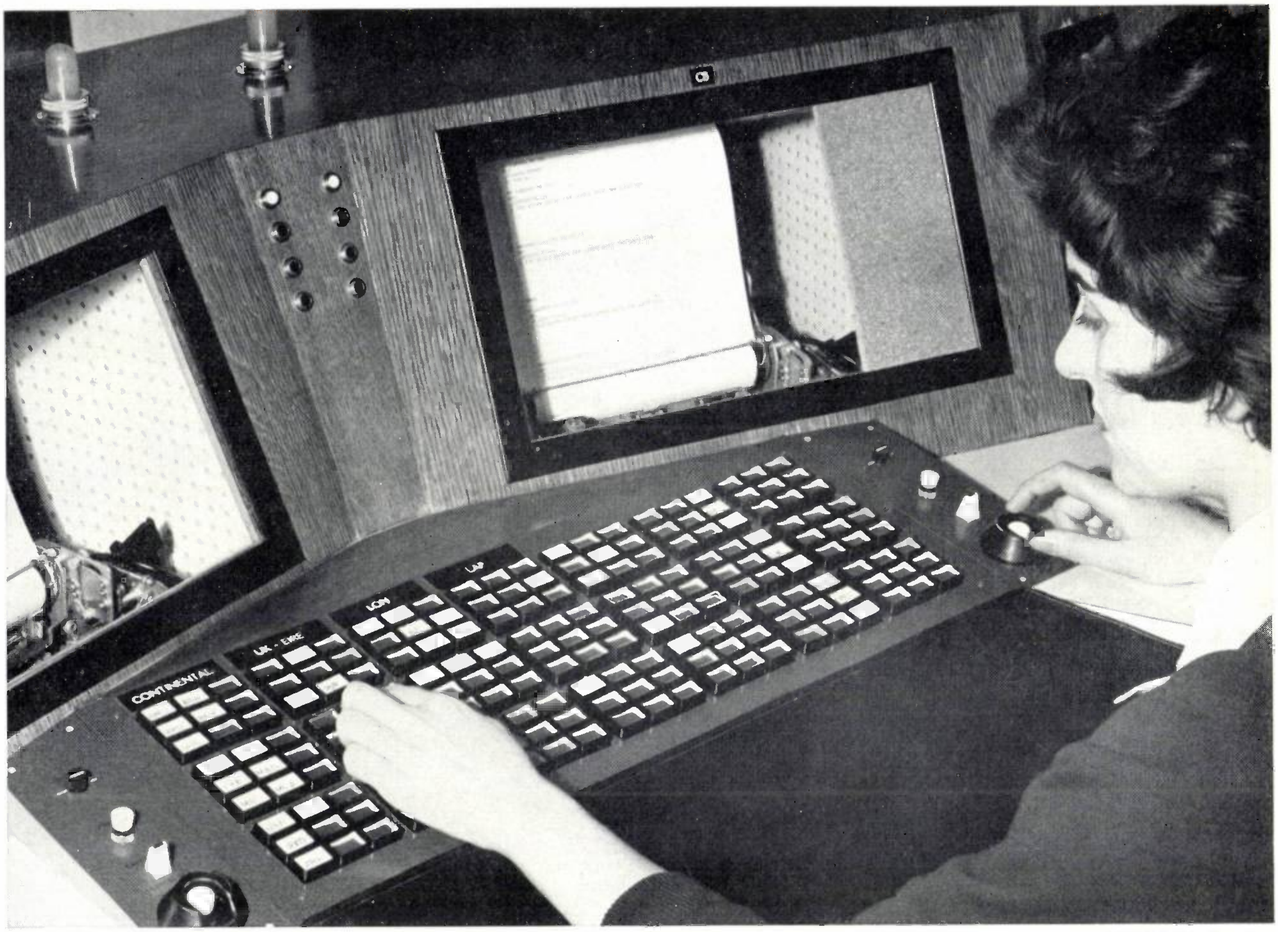


Foto ART-WOOD, London

Fig. 9. Half of a double operating position for an exchange arranged following the diagram in fig. 8. Each operator controls two teleprinters which are installed behind glass to reduce the noise. After reading the address from the teleprinter, the operator presses one or more of the routing buttons on the panel in front of her. The operating time is sufficiently short that each operator can easily attend to two teleprinters.

to her teleprinter and the message is forwarded completely automatically.

If we substitute the incoming connecting circuit *RCC* for the automatic transmitter of fig. 5, we see that the two diagrams are otherwise entirely identical, so that further explanation of how this automatic switching takes place is superfluous. It is, however, interesting to note that once the necessary connections have been set up by the marker, retransmission of the message from the store of *RCC* begins, either directly to a free line of indirectly to the intermediate store of a secondary connecting circuit *SCC*.

There are thus two groups of stores in the exchange and one may wonder why both functions cannot be performed by the same store. It will, however, be appreciated that the chance of not a single incoming connecting circuit being free when a message comes in must be extremely small, and if we do not wish to be forced, for this reason, to install an uneconomically large number of connecting circuits, the holding time of these circuits has to be kept low, which would be out of the question if they were also used as queueing stores. Moreover, separation of the functions makes it possible to give the stores in the incoming connecting circuits half the capacity of those in the secondary connecting circuits. The greatest permissible length of a message is 2000 characters and this is the capacity of the stores in the secondary connecting circuits. That of the stores at the incoming side, however, is only 1000 characters, which is possible because retransmission from these stores always takes place so quickly that when character 1001 comes in, position 1 in the store has already been cleared. The use of this cyclical mode of operation means that the incoming stores need be less expensive and there is less objection to making a large number of them available.

From an operational point of view it is a great advantage that the incoming lines no longer terminate at the operators' positions but at equipment in the automatic part of the centre. The number of operators can therefore be much more readily adapted to the traffic situation as it is at any particular moment.

As the traffic increases or decreases, more or fewer of the positions can be occupied, whereas in the Paris exchange the positions must be attended until the decrease in traffic allows an operator to attend two positions at once, which involves walking to and fro.

Fully automatic switching

It is not hard to appreciate that in the diagram of fig. 8 the operator's position with its associated teleprinter circuit and relay store can be replaced by an automatic circuit which extracts the address section from the message in the incoming store, decodes the addresses and then passes them to the marker which selects the connection. This arrangement has been adopted for the extension of the London SITA exchange and, as already stated it requires the composition of the message to follow a very strict pattern. Not only can the two sections of the centre work perfectly together, but the traffic capacity can be expanded without the addition of new operating positions.

While the transition to fully automatic operation certainly means more complicated control equipment, it is not necessary to introduce new storage functions. As in this article we are mainly concerned with the question of how the automatic switching of messages can be rendered possible by the introduction of storage functions, we will not go into the details of this more complicated control.

A simplified system

The method we have followed in discussing the problem of automatic switching, — the examination in chronological order of a number of relevant projects — has the advantage of enabling the logical development of the systems to be convincingly demonstrated. In it, however, lurks the danger that readers mayconsider the logic of this development to be to some extent inescapable. It should, however, be remembered that this development was adapted to the clearly defined requirements of a definite class of user. As already stated, the messages which are transmitted over the airlines' network are mainly concerned with reservations for passenger flights and air freight. Although great importance is naturally attached to rapid forwarding of such messages, the economic factor continues to be very significant. That is why the solutions described have not only been aimed at minimizing the delays due to switching in the exchange but are also designed to load the circuits, which are often very expensive, to their maximum capacity.

The latter is the purpose behind, for instance, the double type of line circuit shown diagrammatically in fig. 7. This technique prevents the accumulation of messages in the common equipment and removes it to the line equipment of the heavily loaded routes. It is important, nevertheless, to realize that delays of the order of ten minutes in the forwarding of messages from the stores of these double line circuits to the line itself are far from hypothetical.

Such delays are often considered permissible in airline reservation traffic. Alongside the commercial network operated by the airlines, however, there is the telegraph network of the official aeronautical services, which are responsible for the technical control of flights and for the safety of air traffic. Delays of the order of ten minutes in the forwarding of messages can be completely unacceptable when air traffic safety is at stake.

We have seen that delays in retransmission are always due to heavy loading of lines. That is why the aeronautical services have, for example, laid down that a traffic route which has no more than one channel at its disposal must not be loaded more than 40%. For larger groups of channels the permissible specific load increases somewhat but it still remains on the low side of what the airline companies would like to permit.

It is fairly obvious that if, for the reasons stated above, the number of circuits available is ample, some of the requirements which are progressively satisfied by the solutions described lose a great deal of their force. There was considered therefore to be good reason for designing a simplified system based jointly on the experience acquired from the projects described and on the earlier principles of the diagram in fig. 3.

The resultant circuit diagram, which is intended for medium-sized centres, is given in *fig. 10*. Incoming cir-

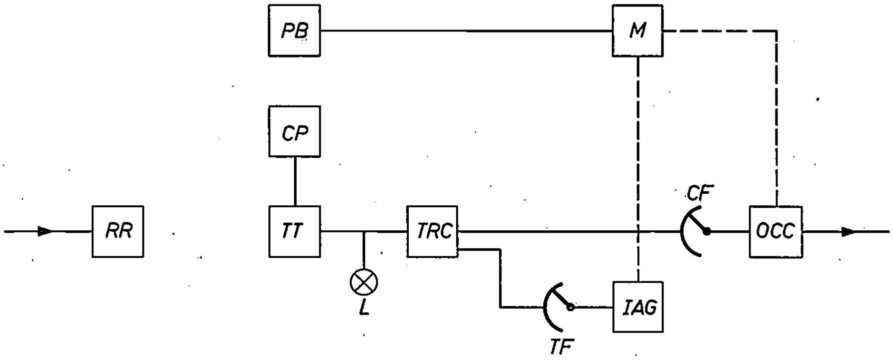


Fig. 10. Diagram of an ES-0 exchange developed from the diagram in fig. 5. This arrangement, described in further detail in the text, can be used to advantage in a centre which is not too large and has moderately loaded outgoing lines.

cuits again terminate on reperforators *RR*, a limited number of which are installed at each operator's position. Each operator also controls a certain number of perforated-tape transmitters *TT* in which she inserts the tapes torn from the reperforators. After reading the destination address(es) of a message she presses first the appropriate routing button(s) on the push button panel *PB* and then, finally, the start button on her control panel *CP* for the transmitter concerned.The data relating to the routes selected are then transferred to the central marker *M* which then tests whether any line circuits *OCC* in these routes are free.

If one or more lines in the desired direction are free, the connecting circuit finders *CF* of the line circuits concerned are directed towards the transmitter circuit *TRC* associated with the transmitter in which the message is waiting. If there is no free line in one or more of the selected routes, the marker marks an individual address store *IAG* which is free and also directs transmitter finder *TF* to the transmitter circuit concerned.

The individual address store comprises a number of relays in which the marker records the routes for which no free line is available. This store, however, is unable to take the text of the actual message. The marker then releases the connection while transmitter *TT* — provided at least one of the selected routes was found free — is started from the line circuit.

If not all the selected routes were free and an individual address store has therefore been switched in, this store causes a signal lamp *L* beside the transmitter to light up, thereby indicating that a single transmission will not be sufficient to forward the message to all its destinations. When the first transmission is ended, the tape leaves the transmitter, closing a contact and causing lamp *L* to start flashing. This is a signal to the operator to reinsert the tape and to push the start button again. If a connection to a destination which was originally busy is or becomes free, the marker *M*, and the individual address store *IAG* in which all the destinations were noted, together ensure that the connecting circuit finder of the line concerned is positioned on the transmitter circuit *TRC*, and the transmitter is started again. This process is repeated until the message has been transmitted to all its destinations.

Continuity

Examination of the diagrams in figs. 10, 5 and 8 will reveal a considerable degree of relationship between them. This relationship is such that it is possible to progress from the circuit of fig. 10 to that of fig. 5 and hence to fig. 8 and from the latter to the circuit — not shown — of a fully automatic exchange, without rendering any considerable proportion of existing equipment superfluous. The requirement that the equipment should be suitable for gradual adaptation as the degree of automation increases is therefore satisfied.

Summary. Four types of automatic telegraph exchange are described which are suitable for use in networks handling small groups of lines. The introduction of electronic stores enables messages to be held temporarily at junctions if no line to the next junction is immediately available. This gives increased line efficiency. These stores also make it possible to satisfy a requirement which is often specified, the automatic transmission of individual messages to a number of different destinations.

The greater the traffic offered and the number of routes to be served, the more numerous the functions for which electronic stores are economically justified and which can be performed automatically in the four types of exchange. The series is designed so that gradual conversion to successive levels of automation is possible. The original development work was done in close cooperation with the "Société Internationale de Télécommunicaton Aéronautiques" (SITA) and in connection with the Société's exchanges in London and Paris.

Magnetic tape store for telegraph characters

H. van Kampen

621.318.24:621.394

Teleprinters operating on the stop-start principle are used on a large scale in modern telegraph traffic. Switching functions in networks based on such teleprinters are now generally being automated. Either of two different principles may be adopted in this process, the choice depending on the nature of the network to be automated. In public telex networks, the connection between the originator and the receiver of the message is established before the exchange of information begins. In many non-public networks, however, a method has been chosen whereby messages are forwarded from centre to centre and, where necessary, recorded temporarily in a store until the circuit to the next centre is free.

An article by P. Harkema in the present issue [1] contains a description of automatic switching centres designed upon this principle and incorporating groups of storage units in which ferrite cores are used as store elements. One advantage of such stores which interests us here lies in the fact that the messages they contain can be read out two or more times for transmission to a number of different destinations. The article also points out that it is occasionally necessary to provide certain outgoing lines with individual stores. This may be necessary when a line has so much traffic to handle that messages intended for it accumulate and threaten to overload the group of central stores. Provision of an individual store may also be desirable if the outgoing circuit employs a radio channel. On teleprinter circuits with radio channels, it is a common practice to use "TOR" (Telex Over Radio) equipment. This equipment is designed to detect mutilations which occur in the transmission of telegraph characters, and upon detection of a mutilation, ensures the retransmission of as many characters as are necessary to put the mutilation right. The number of such repetitions, especially when propagation conditions on the radio path are poor, may be high, so that the mean signalling speed is considerably lower than the nominal one. The number of characters arriving at the channel input will then obviously be greater than the number leaving from the channel output. It is therefore necessary to insert at the input a recording device that can always accept messages from the stores in the central groups at nominal speed, and so ease the load on this group, while it can retransmit to the outgoing channel at a much lower average speed.

In the cases mentioned, where individual stores are necessary, the messages need not be extracted from the store more than once. Also the storage capacity sometimes has to be so much greater than that of a central group store that ferrite core stores would be uneconomic and other solutions to the problem must be sought.

Perforated paper tape has long been in use for storage purposes in telegraph centres. It satisfies the requirement as to storage capacity but has the disadvantage that it can only be used once. Where traffic is heavy, expenditure on tape is far from negligible. There is the further point that perforator and tape transmitter mechanisms are rather complicated and demand regular maintenance. Moreover the channel may be put out of operation if a reel of perforated tape runs out at an awkward moment.

These considerations led to the development of the store described below, in which the storage function is performed by an endless magnetic tape of sufficient length to record 40 000 telegraph characters. This capacity is sufficient for all practical purposes. The disadvantages of paper tape have been wholly obviated by this solution. *Fig. 1* shows how the 20-metre magnetic tape, on which 40 000 characters can be recorded, is housed in a cassette. Separate writing and reading heads, each with its own tape-feed mechanism, are fitted above the cassette.

The design

Investigation into the requirements that a magnetic tape store for telegraph characters must satisfy soon revealed that the design currently in use in popular instruments for recording music and speech would not do. In the first place, as is clear from what has been said above, the writing and reading functions must operate independently of each other and simultaneously. A magnetic head and a tape-feed mechanism must be available for each function. At the same time, separate tape-feed mechanisms would also make it possible to use different speeds for writing and reading. Teleprinters at present in use throughout the world do not all have the same speed of operation. In Europe and adjacent regions the CCITT has standardized a speed of 50 bauds. In the United States, however, speeds of 45.45 and 56.88 bauds have long been in

Ir. H. van Kampen is on the staff of N.V. Philips' Telecommunicatie Industrie, Hilversum.

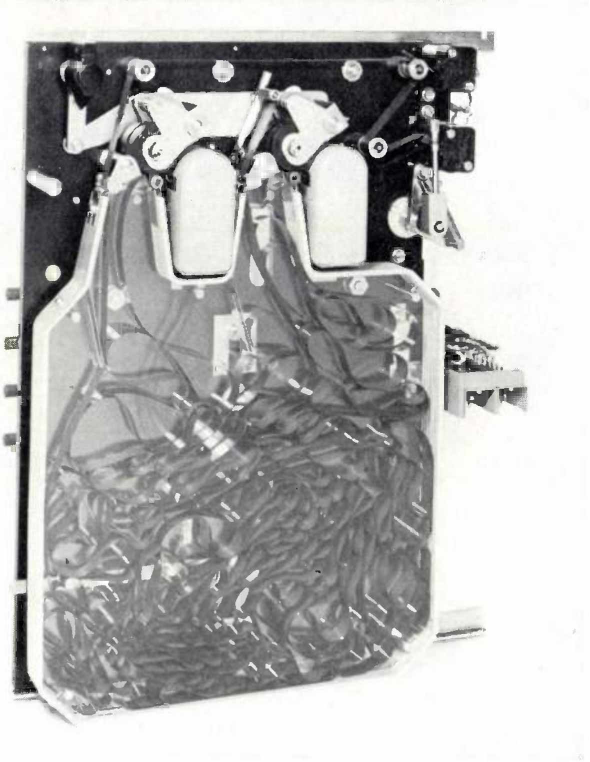


Fig. 1. The 20 metres of magnetic tape needed to record 40 000 teleprinter characters are housed in a cassette. Above the cassette are the writing and reading heads with their tape feed mechanism.

use, while the considerably higher speed of 74.2 bauds is also at present in use. It is therefore possible, by choosing different writing and reading speeds, to use the storage device to couple networks in which different signalling speeds are used.

The desirability of being able to work the store with TOR equipment makes it essential to be able to record characters on the magnetic tape one at a time. It must also be possible to control the starting and stopping of the tape from outside.

In tape recorders for music and speech the tape is moved at uniform speed over the recording and playback head. Changes in the magnetization of the tape induce voltages in the coil fitted round the magnetic circuit of the head. These voltages are proportional to the frequency of the signal to be read out. At signalling speeds of 45.45 to 74.2 bauds the signal frequencies are so low that no useable output signal can be obtained.

For these reasons a read-out method is used in which the output voltage depends solely on the value of the magnetic field at the point of read-out and not on the changes in the field along the tape. The output voltage therefore does not depend on the speed at which the tape moves over the head and is not affected by the fact that the tape does so step by step.

A teleprinter character, as the reader probably knows, consists of seven elements transmitted serially: a start element, five code elements and a stop element. Only the five code elements need be recorded on the tape. A separate track for each of them is available on the $^1/_2$ in. wide tape. A series of alternating positively and negatively magnetized elements are written on a sixth track for control purposes.

As already observed, the magnetic tape is moved forward one step for every telegraph character. The time taken to do so is considerably shorter than the time needed to transmit a complete character at the highest speed which occurs, namely 74.2 bauds. The tape-feed mechanism need therefore not be modified if the signalling speed is changed. Adaptation is entirely confined to the electrical control and in the present case can be done remotely. The distance the tape is moved per character is 0.5 mm, so that 20 metres of tape are needed for 40 000 characters. This length of tape can be accommodated in a cassette of reasonable dimensions, as shown in fig. 1. From what has already been said, it will be obvious that the amount of tape needed is not affected by the signalling speed chosen.

Writing and reading

As we have seen, the characters in the store dealt with here are not read by determining the variations of the magnetic field along the tape, but by measuring the absolute amount of magnetization at individual points. The method adopted is based on the principle put forward by Kilburn et al [2]. We have departed from their procedure by impressing the magnetization perpendicular to, instead of along, the surface of the tape, so that distinct external fields and therefore individually measurable signals are obtained, even when there is a succession of elements with the same polarity

A writing head and a reading head are shown diagrammatically in *fig. 2*. It should be realized that six such combinations are used. *Fig. 3* gives an enlarged picture showing how the five tracks carrying code elements and the single track carrying control elements are arranged on the magnetic tape. In analogy with the procedure usually adopted for perforated paper tape, it is the third track which is reserved for control purposes.

No further explanation of how the writing head works will be necessary. The magnetic circuit of the reading head is divided at one point in such a way as to form a closed magnetic sub-circuit. Round the arms of this sub-circuit are two windings w_1 and w_2. Winding

[1] P. Harkema, Automatic telegraph exchanges with electronic stores, Philips tech. Rev. **26**, 240-249, 1965 (No. 8/9).

[2] T. Kilburn, G. R. Hoffman and P. Wolstenholme, Reading of magnetic records by reluctance variation, Proc. IEE **103**, Suppl. No. 2, 333-336, 1956.

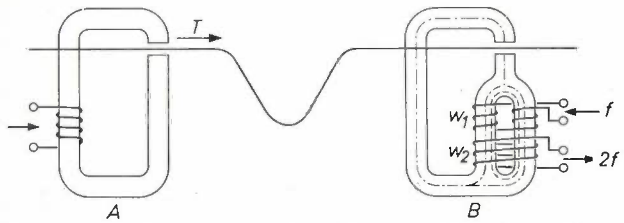


Fig. 2. Schematic representation of the reading and writing heads. Writing head *A*, on the left, impresses magnetization perpendicularly on the surface of the tape *T*. The magnetic circuit of reading head *B* is divided at one point into two arms round which windings w_1 and w_2 are fitted. The function of this sub-circuit is described in the text.

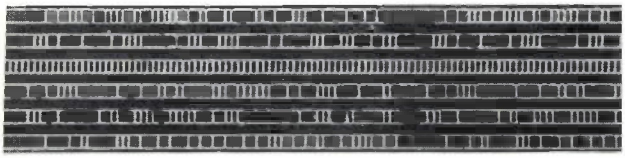


Fig. 3. Drawing showing how the five code elements of a teleprinter character are recorded on five tracks of the magnetic tape. A sixth track (track 3 on the tape) has a series of alternating positively and negatively magnetized elements impressed on it and is used for control purposes.

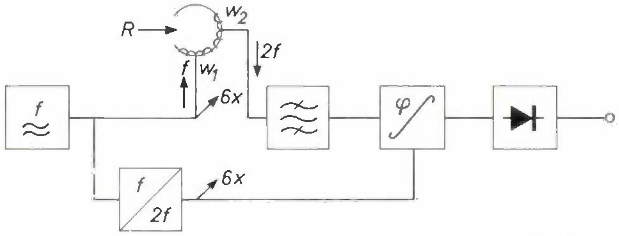


Fig. 4. Circuit in which windings w_1 and w_2 of the reading head *R* are included. *φ* is the phase discriminator.

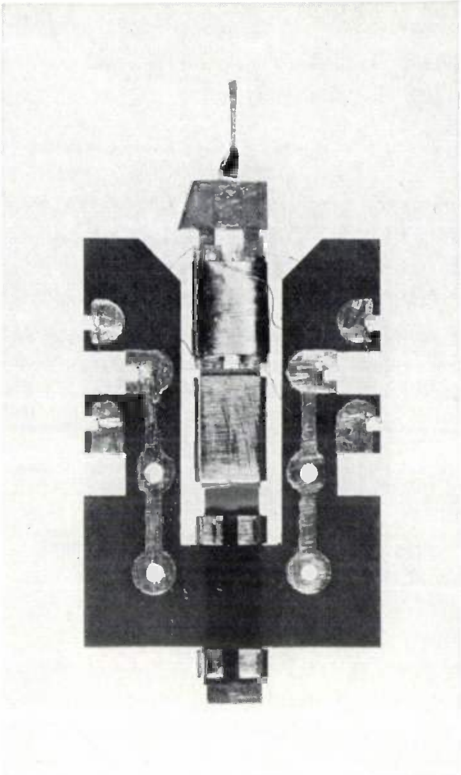


Fig. 5. A part of reading head *B* shown in Fig. 2. The leads of windings w_1 and w_2 are terminated on a synthetic resin bonded paper panel, with the magnetic material at the centre. The head ends in a narrow point over which the magnetic tape runs. Below this point can be seen the division of the magnetic circuit into two arms. The reading heads are extremely flat because six of them have to be fitted together within the $^1/_2$ inch width of the magnetic tape.

w_1 is arranged so that any current passing through it will set up circulating magnetization in the sub-circuit but not in the main circuit. Winding w_2 is a test winding and insensitive to variations of circulating magnetization in the sub-circuit, but is sensitive to variations in the main circuit.

An auxiliary a.c. current of frequency *f* permanently flows through winding w_1 with an amplitude such that magnetic saturation periodically occurs in the sub-circuit. As a result of this saturation, the magnetic resistance in the main circuit is always high except at the moments of polarity reversal of the current through w_1. Only at those moments is it possible for the field of the tape to cause any perceptible magnetization in the main circuit. As there are two polarity reversals per cycle of the a.c. current through w_1, an alternating voltage with a frequency of 2*f* is generated in w_2.

Careful consideration will show that if the magnetic field reverses its direction when the tape moves, the phase of the a.c. voltage in w_2 will change by 180°. Use is made of this fact in the reading circuit shown in *fig. 4*. A generator sets up an a.c. voltage of frequency *f* which is applied to winding w_1 of the reading head *R*. An a.c. voltage with a frequency 2*f* is taken from w_2 and applied to phase discriminator *φ*, via a band-pass filter for the suppression of unwanted frequency components. The voltage used for comparison is a voltage of invariable phase and frequency 2*f* obtained from the generator frequency by doubling. The phase discriminator is adjusted so as to deliver a positive output voltage for one direction of the field on the magnetic tape and a negative voltage for the other.

Fig. 5 shows the construction of the part of a reading head on which windings w_1 and w_2 have been fitted. The magnetic circuit is closed by a yoke common to all six reading heads.

Tape feed

In each drive mechanism the tape is moved over a small steel roller. This roller is driven via gear-wheels by a motor, to be described below, and has a diameter such that the tape moves a distance of 3 mm for every complete revolution of the motor. As each character is allocated 0.5 mm of tape, the motor makes one sixth of a revolution for every character to be recorded. The roller is smooth and cannot be used to drive the magnetic tape directly. As the drawing on the left of *fig. 6* shows, the tape is held against the metal drive

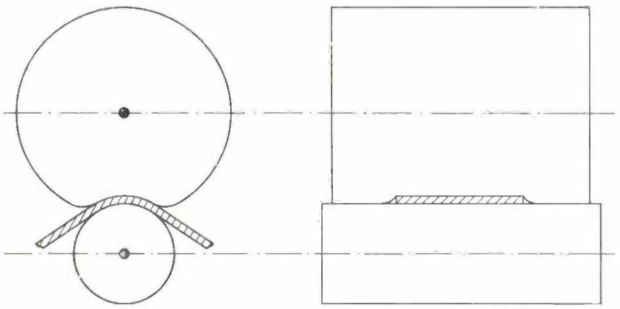


Fig. 6. Enlarged drawings of the magnetic-tape feed mechanism. The left-hand one shows how the tape runs between a small steel roller and one made of resilient rubber. As the steel roller is too smooth to drive the tape directly, the rubber roller is made wider than the tape. The drawing on the right shows how the steel roller drives the rubber roller on either side of the tape, which is thus driven by the centre section of the rubber roller.

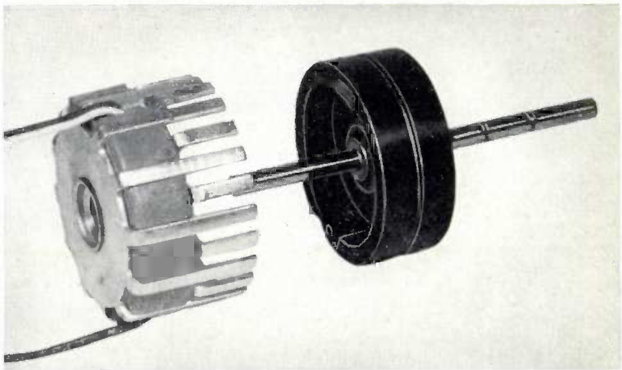


Fig. 7. The ferroxdure rotor (right) and one of the two stators of the drive motor. The stator consists of a disc-shaped coil between two plates, each of which has twelve teeth bent towards the stator.

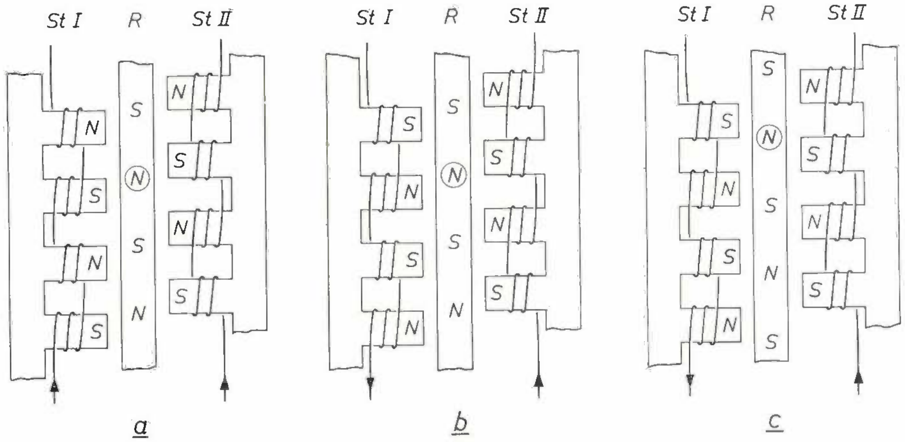


Fig. 8. Schematic diagram of rotor R and two stators $St_{\rm I}$ and $St_{\rm II}$. In the unoperated condition a, a north pole of the rotor is positioned on the line connecting two south poles. In b the rotor is still in the same position, but the direction of the current in $St_{\rm I}$ has been reversed. The rotor is subjected to an upward force and therefore moves to the position it occupies in c.

roller by a roller of fairly resilient rubber. The rubber roller, being broader than the tape, projects beyond it on both sides, as illustrated in the right-hand drawing, and so presses against the steel roller as well. The friction between the rubber and steel rollers is sufficient for the latter to drive the former without slipping. The rubber roller in turn transports the magnetic tape, also without slip.

The motor which provides the driving power has two identical stators and a ferroxdure rotor. The rotor and one of the stators are illustrated in *fig. 7*. Each stator consists of a disc-shaped coil between two soft-iron discs, each of which has twelve upturned teeth. When the coil is energized, twelve pairs of poles are set up at the stator teeth. The rotor is permanently magnetized in such a way that twelve pairs of poles also exist along its perimeter. As *fig. 8* shows, the two stators $st_{\rm I}$ and $st_{\rm II}$ are offset relative to each other by half a pole pitch.

When both stator coils are energized in a given direction, as for example in fig. 8a, the rotor will always position itself so that a north pole of the rotor lies on a line joining two south poles of the stators. If, as in fig. 8b, where the rotor still occupies the same position as in fig. 8a, the current through one of the stator coils is reversed, a north pole of the rotor will be repelled by the two stator north poles below it, and attracted by the two south poles above it. The direction in which the rotor will turn is therefore decided, and it will rotate half a pole pitch upwards to the position it occupies in fig. 8c. It can easily be shown that, if the current direction is switched in one stator coil and the other alternately, the rotor will advance regularly. The rotor has a total of 48 discrete positions and the motor has to make 8 steps to transport the tape over the length of one character.

Operation as a whole

Fig. 9 is a schematic representation of the main parts of the store. On the left can be seen the receiving distributor A. This receives the telegraph characters whose elements arrive one after the other. The distributor is triggered by the arrival of the start element of a character. The five code elements are recorded in the same number of bistable (flip-flop) circuits. When the stop element is received, the contents of these five bistable circuits are transferred to five more in writing amplifiers $C_{1,2,4,5,6}$. Simultaneously, the correspon-

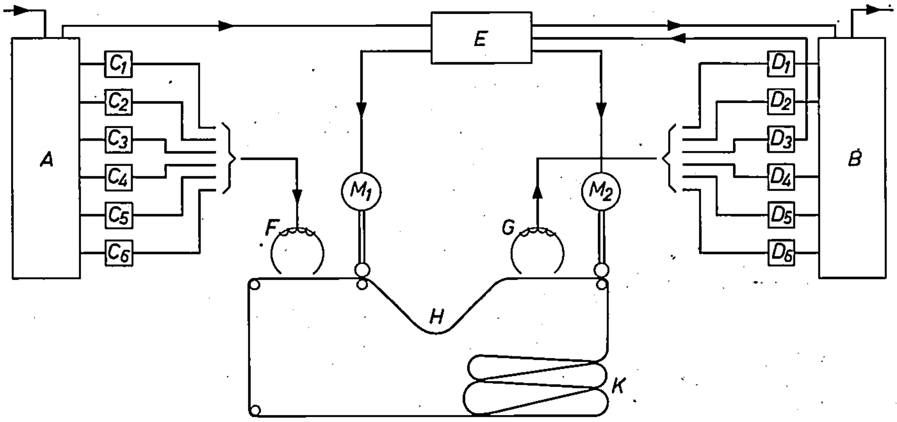


Fig. 9. Diagram illustrating the working relationship between the various parts of the store. A receiving distributor; B transmitting distributor; C_1 to C_6 writing amplifiers; D_1 to D_6 reading amplifiers; E control circuit; F writing head; G reading head; H storage loop; K tape in cassette; M_1 and M_2 drive motors.

ding bistable circuit in amplifier C_3 is changed over, so that the character element recorded by this amplifier is given opposite polarity to that of the element which preceded it. The receiving distributor is is immediately ready for receipt of the next character.

As the character elements are transferred to the writing amplifiers, a signal is sent to control circuit E, which switches on the drive motor M_1 associated with heads F and makes it rotate 8 steps. While the tape is advancing 0.5 mm during this process, the six elements are recorded on it at the same time. As we have already seen, the movement of the tape is so rapid that it is always completed before the receiving distributor has received all of the next character even if this character comes immediately after the preceding one.

As long as the store has no more text for retransmission, the magnetic tape between the writing and reading head is taut. On becoming taut it operates a lever which opens a contact, causing drive motor M_2 for reading head G to switch off. When the next message comes in, the motor of the writing head starts and part H of the magnetic tape will hang slack. The reading-head motor is therefore switched on again and the tape starts moving under the reading head.

As we shall explain presently, there is no text on the length of tape between the writing and reading heads in the non-operated condition. Further, in this condition the elements of the control track written on this part of the tape all have the same polarity. No further change of polarity on this track occurs until the first character in the next incoming message is written. As soon as reading amplifier D_3 detects a change of polarity on the control track, the reading-head motor is once more stopped. Retransmission of the character which is under the reading heads then depends on whether the telegraph channel connected to the output of transmitting distributor B can accept the character or not. If the outgoing channel can take the character, it sends an appropriate signal to transmitting distributor B. Amplifiers $D_{1,2,4,5,6}$ will consequently read the elements of the character simultaneously from the tape and transfer them at once to five bistable circuits in the transmitting distributor. The latter transmits the five code elements of the character sequentially, prefacing them with a start element and appending a stop element. At the moment transmission begins, however, the motor at the reading head is started and the tape is moved forward until the next change of polarity on the control track is detected. Pending the next instruction from the circuit associated with the outgoing channel, the next character waits in readiness under the reading heads.

If the message has been completely received, the receiving distributor will detect no more start elements and the write motor will cease stepping. The motor at the reading end will continue moving the tape forward until it is taut. The lever referred to above now operates and switches off the read motor. As the length of tape between the writing and reading heads contains a part of the message which has not been transmitted, the write motor is restarted by the control circuit, which allows it to run long enough to feed in all the tape necessary for complete retransmission of the "tail" of the message. The fact that the write motor is now running while no characters are coming in means that the polarity of the writing current through the third head does not change. Since a new message may start arriving while extra tape is being fed through, the circuit is arranged so that upon receipt of a start element the write motor immediately stops advancing extra tape and starts stepping again under control of the incoming characters.

A final point which might be added is that when

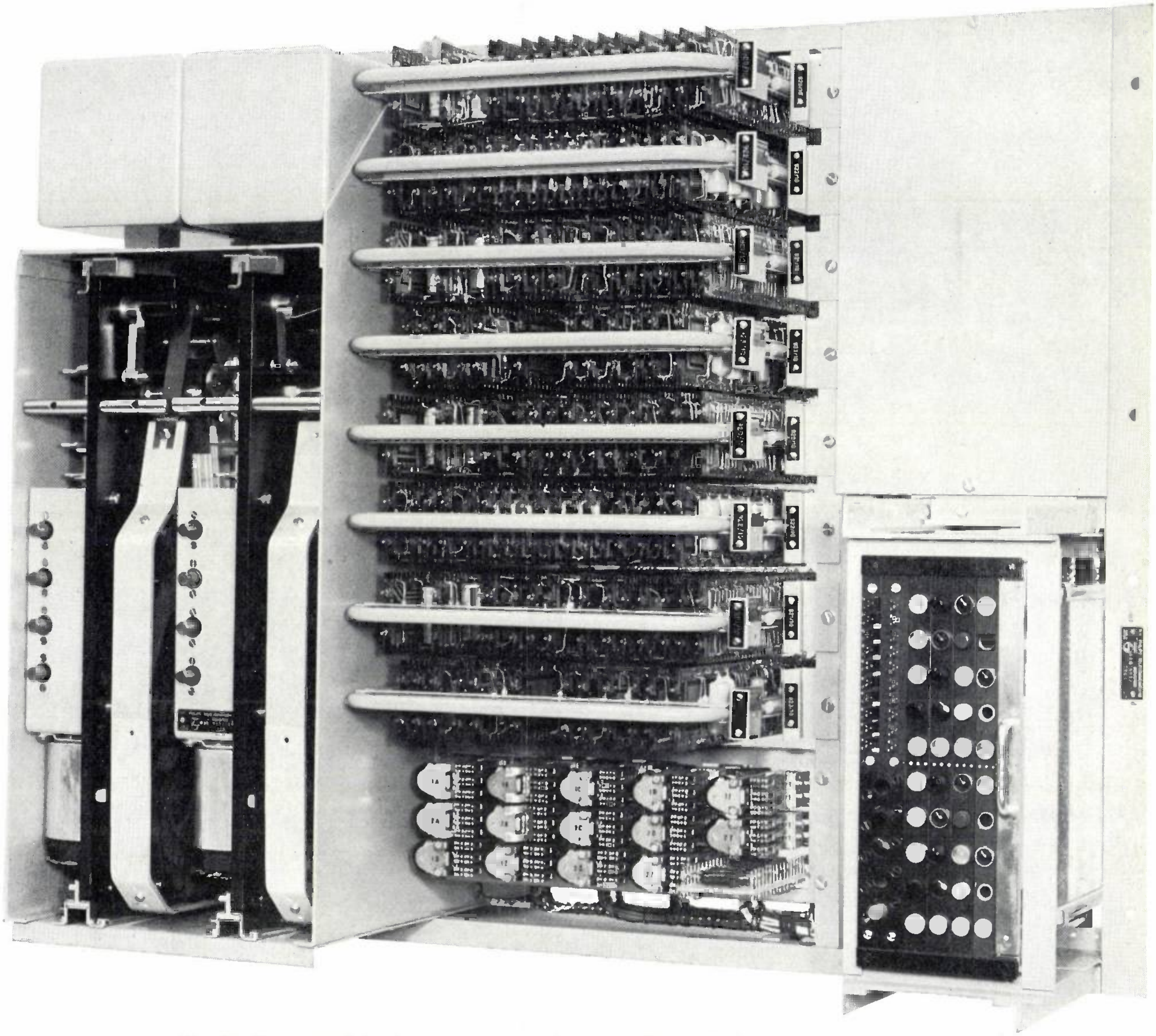


Fig. 10. Two complete stores are mounted on one frame. Left: two cassettes containing magnetic tape and, above them, the reading amplifiers in hermetically sealed containers. The other electronic equipment is mounted on panels which can be removed in groups for maintenance. Right: a supervision panel with lamps and jacks.

the various character elements are recorded the magnetic layer is always saturated. It is therefore unnecessary to erase the recorded information after read-out.

Arrangement of the equipment

As *fig. 10* shows, the cassettes illustrated in fig. 1 are mounted in pairs on a frame with the associated equipment. The two sets of reading amplifiers are mounted immediately above the cassettes in hermetically sealed containers. These are necessary because the amplifiers have to operate within very close tolerances. The writing and reading heads in the cassettes are magnetically screened to protect them from disturbing influences such as the earth's magnetic field.

The remaining electronic equipment, is less sensitive, and is mounted on narrow strips of SRBP arranged vertically in the middle of the rack but combined in horizontal groups to form units. For maintenance each such unit can be withdrawn from the rack by means of the bar running along the front of it.

A number of pilot lamps, test jacks, etc., are assembled on a supervision panel installed to the right of the rack.

Summary. A teleprinter character store for use in automatic telegraph centres is described. A capacity of 40 000 characters has been attained by the use of magnetic tape as the storage medium. In view of the low frequencies used to transmit telegraph characters, a static method is employed in reading characters from the tape. The latter does not run at a uniform speed but is moved forward a step at a time. This is done by means of a stepping motor of very simple design. The store is suitable for the various speeds at which teleprinters are normally operated, the necessary matching being effected completely electrically, without any adjustment of the mechanism.

Television transmitters for the ultra-high frequency band

J. A. van der Vorm Lucardie

621.397.61.029.6

A number of frequency bands in the very high and ultra-high frequency regions of the radio spectrum have been set aside by international agreement for the provision of television and sound broadcasting programmes. In these regions, which are generally referred to by their initials VHF and UHF, a total of five bands are available. Bands I, II and III are in the VHF spectrum and bands IV and V in the UHF. Band II is used for sound broadcasting, while the others have been assigned to television broadcasting.

For frequency-band allocation the world is divided into three regions, and the location of the frequency bands varies slightly for each of these regions. That for Europe, the Near East and North Africa was last defined in 1961 during the Stockholm conference [1]. As the use of frequencies in UHF bands IV and V was still in a very early stage, it proved possible to define the bandwidth available per television channel — 8 Mc/s — and the positions of the vision carrier frequencies within bands IV and V in a uniform manner.

For the sake of completeness the positions of the five bands are here quoted:

Band I :	41 — 68 Mc/s	(7.3 — 4 m),
Band II :	87.5 — 100 Mc/s	(3.4 — 3 m),
Band III:	162 — 230 Mc/s	(1.85 — 1.3 m),
Band IV:	470 — 582 Mc/s	(64 — 51 cm),
Band V :	582 — 960 Mc/s	(51 — 31 cm).

It will be seen from this table that bands I and III are relatively narrow and they can therefore accommodate only a limited number of TV channels.

The service area of a television transmitter — i.e. the area in which the field strength is sufficient to ensure good picture quality — depends on the height of the aerial and is confined within a radius of approximately 35 miles (60 km) for powerful stations. The interference area, however, extends much further and therefore the distance between two transmitters which it is intended to operate on the same frequency has to be several hundred kilometres. It is consequently impossible, even in countries where there is only one television programme, to obtain a completely closed pattern of service areas, as is desirable in Europe, using only channels which are available in bands I and III. If this is to be achieved, a number of additional channels in the UHF spectrum are necessary.

In addition to this need, however, a far greater one has been created by the desire to transmit a second or even a third television programme. Fortunately, the much greater width of bands IV and V amply allows the demand to be met in these bands.

In designing television transmitters for the UHF range a number of problems are encountered which are more or less peculiar to these high frequencies. We shall devote particular attention to such points in the present article.

The power required for transmitters in bands IV and V

A certain minimum field strength is necessary to ensure a good-quality television picture. An increase in the field strength need not involve increasing the transmitting power but can also be achieved by arranging that the RF power is not radiated uniformly in all directions but strongly concentrated in the horizontal plane. A number of aerial designs are available for this. It is consequently customary, in describing television transmitters, to speak of their "effective radiated power", which is defined as the product of the power applied to the aerial and a factor dependent on the type or design of aerial employed. This factor is expressed in decibels and is called the aerial power gain.

When allowance has been made for the effective aerial height, the gain of the receiving aerial and the noise contribution of the receiver, it is found that the effective radiated power of a UHF station has to be approximately 10 dB higher than that of a VHF station for the same quality of reception. The effective radiated power of large stations in bands I and III being 30-100 kW, an ERP of 300-1000 kW is necessary in bands IV and V.

At the frequencies with which we are concerned here the limit of the service area more or less coincides with the optical horizon as seen from the aerial, which is therefore erected in as elevated a position as possible. The short wavelengths in bands IV and V make it possible to obtain much greater aerial gain than in bands I, II and III, while keeping the size of the transmitting aerial within economic bounds. The limit is determined by mechanical considerations. The maximum gain factor that can be attained is approximately 50. If, finally, allowance is made for

Ir. J. A. van der Vorm Lucardie is on the staff of N.V. Philips' Telecommunicatie Industrie, Hilversum.

[1] Final acts of the European VHF/UHF Broadcasting Conference, Stockholm, 1961, published by the International Telecommunication Union, Geneva.

the fact that very considerable losses occur in the coaxial cable connecting the transmitter and the aerial in its high position at band IV and band V frequencies, it is finally found that transmitter powers of 10-40 kW are necessary.

On the basis of this result Philips' Telecommunicatie Industrie first developed a 10 kW transmitter. This was followed by a 20 kW transmitter, while for smaller service areas a 2 kW transmitter is now also available. All of these are suitable for colour television. When necessary, double power can be obtained by connecting two transmitters in parallel. An advantage of this arrangement is that if one transmitter develops a fault, the transmission can continue without interruption, though at reduced power. We shall return to this point later.

General arrangement of a television transmitter

What is generally called a television transmitter is in fact a combination of two transmitters, one of which transmits the vision signal and the other the accompanying sound signal. As the aerial and its feeder cable form a very costly element, both transmitters are always connected to the same aerial by means of a coupling network, for which the term combining unit has been generally adopted in television practice. The combining unit prevents the sound-transmitter and vision-transmitter from interacting on each other.

The essential parts of a vision-transmitter are shown in *fig. 1.* A crystal oscillator *A* generates a signal whose frequency is a sub-multiple of the transmitting frequency. The transmitting frequency is attained by multiplication in stage *B*. In *C* the radio-frequency signal is modulated by the video signal, which has previously been amplified in the video modulator *E*. As a rule, modulation takes place as close as possible to the output stage, and sometimes in the output stage itself. The effect of this is to simplify transmitter tuning and improve the stability. In output stage *D* the modulated signal is amplified to output level.

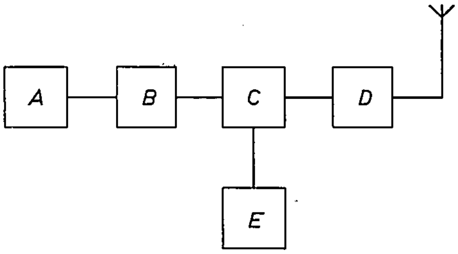


Fig. 1. Diagram showing the main parts of a vision-transmitter. *A* crystal oscillator; *B* frequency multipliers; *C* modulated amplifier; *D* final amplifier; *E* video modulator.

Amplitude modulation is used for picture transmission. The bandwidth occupied is limited by suppressing a large part of the lower sideband of the modulated signal. In television transmitters in which modulation is effected in the output stage, the general practice is to have the output stage followed by a filter combined with the combining unit. This combination is called a filterplexer. In the transmitter to be described below, a combining unit is adequate since a klystron is used as the final power amplifier. Suppression of the lower sideband is effected with a simple coaxial filter inserted before the final amplifier input.

In keeping with normal practice, the sound-transmitter employs frequency modulation. Once again, the process starts with the generation of a signal with a frequency much lower than the transmitting frequency. This is done in the oscillator A_1 (see *fig. 2*), whose frequency is directly modulated with the sound signal, the latter being amplified in *F*. Oscillator A_1 cannot therefore be crystal-controlled and as a result its stability is limited. If the frequency of A_1 were to be raised to its final value solely by multiplication, as is done in the vision-transmitter, the large multiplication factor required would result in equally large magnification of the fluctuations of the mean frequency of oscillator A_1 in relation to its nominal value. The output frequency from A_1 is first doubled in V_1 and then compared with the frequency of a crystal oscillator A_2 in circuit *C*. This comparison circuit delivers a control voltage which corrects the mean frequency of A_1 when deviations are observed. Stage V_2 triples the output frequency of V_1 and the output signal from V_2 is then mixed in stage *M* with that from a crystal oscillator A_3 whose frequency is equal to half the mean transmitting frequency, minus the output frequency of V_3. The frequency of the output signal from *M* is therefore — after the suppression of unwanted products of mixing — half the mean trans-

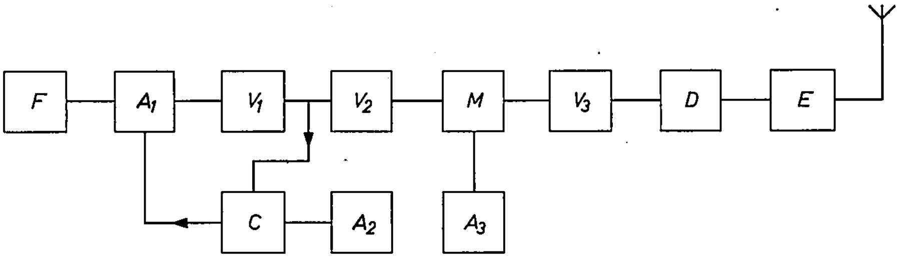


Fig. 2. Diagram showing the main parts of an FM sound-transmitter. A_1 frequency-modulated oscillator; *F* sound-amplifier/modulator; V_1 doubler; *C* comparison circuit; A_2 crystal oscillator; V_2 tripler; *M* mixer circuit; A_3 crystal oscillator; V_3 doubler; *D* power amplifier; *E* output amplifier.

mitting frequency. The transmitting frequency is finally attained in a doubler V_3, which is followed by the power amplifier *D* and the output stage *E*.

As we have already observed, the sound-transmitter operates in the same frequency band as the vision-transmitter, and therefore the design problems raised by the two are largely the same. There are, however, two problems which are peculiar to the vision-transmitter and owe their origin to the fact that its power is approximately five times that of the sound-transmitter and that the bandwidth it requires is several Mc/s, compared with the few kc/s needed for the sound-transmitter. In the description which follows, we will therefore confine our observations to the vision-transmitter.

At the high frequencies of bands IV and V the choice of transmitting valves for the output amplifier constitutes a problem, at least for output powers of 10 to 20 kW. We will begin by examining this point.

Transmitting valves for the output stage of the vision transmitter

In the final amplification stage of their band I and band III transmitters, Philips generally use tetrodes in push-pull connection. At frequencies in the UHF range, two difficulties arise. In the first place the wavelength at these frequencies is comparable with the dimensions of the components used. To avoid unwanted radiation and coupling effects which can result in various types of instability, one must use coaxial techniques, in which the advantages of the simple design of the push-pull circuits are lost. This also means, however, that the full transmitting power can be attained much more simply with one valve than two.

A second feature is that at frequencies in the UHF range the electron transit time leads to incorrect valve operation. We shall not analyse this phenomenon here but merely point out that it greatly reduces the efficiency and power output of tetrodes. The obvious remedy for it is to make the electrode spacing as small as possible. There is, however, a limit to this reduction because, particularly in the case of valves for power of 10 kW and more the thermal demands on the material become greater and greater. This drawback can be partly overcome by replacing the glass by a ceramic insulator but even then the limit of what is possible is still reached sooner or later in large valves.

In the transmitter under discussion use has accordingly been made of klystrons or velocity-modulation tubes. The operating principle of the klystron has been described in numerous publications, including previous articles in this journal [2], and will therefore not be discussed here. We will merely remind readers that in this type of valve a beam of electrons is passed through a cylindrical chamber which is interrupted by gaps at, basically, two places. If an a.c. voltage is applied across the first gap that the electron beam traverses, a velocity modulation of the electrons occurs which leads to a density modulation of the electron beam after the electrons have travelled a certain distance in the valve. If the dimensions of the valve are chosen so that the area where this density modulation is greatest coincides with the location of the second gap, energy can be drawn from the electron beam by loading this gap with a tuned circuit. The r.f. power thus obtained is many times greater than that needed to maintain the modulation voltage across the first gap, so that the valve can operate as an amplifier.

The amplifying effect can be increased further by arranging a number of gaps one after the other and bridging each gap with a tuned circuit. The Philips 11 kW klyston type YK 1001, used in the 10 kW transmitter, and the 22 kW klystron type YK 1061, used in the 20 kW transmitter, have four gaps and four circuits in the form of resonant cavities, and are sometimes called four-cavity klystrons. For the frequencies for which these klystrons are used the gaps are so spaced that both valves are about 5 ft. 6 in. long.

Before deciding whether klystrons or normal triodes or tetrodes are to be preferred for the frequency range in question, a number of considerations relating to design, operation and economy must be taken into account. From a design view point the very considerable power gain of the klystron has the advantage that the number of power amplifiers which precede it can be greatly reduced. This means that, despite the large dimensions of the klystron, the overall dimensions of a klystron transmitter need not be greater than those of a transmitter equipped with triodes or tetrodes.

Operationally, it is of great importance to be able to predict impending klystron failure, as will be seen below. This enables the klystron to be replaced before it fails, thus eliminating the risk of failure during a broadcast and consequent interruption of the programme. The fact that it takes much more time to replace a klystron than the much smaller triodes and tetrodes is, by comparison, of secondary importance. There is often no means of predicting the end of the life of the latter type of valves with any certainty and the possibility of their failure during a broadcast cannot be disregarded. In view of the pre-

[2] See, for example, B. B. van Iperen, Velocity-modulation valves for 100 to 1000 watts continuous output, Philips tech. Rev. 13, 209-222, 1951/52.

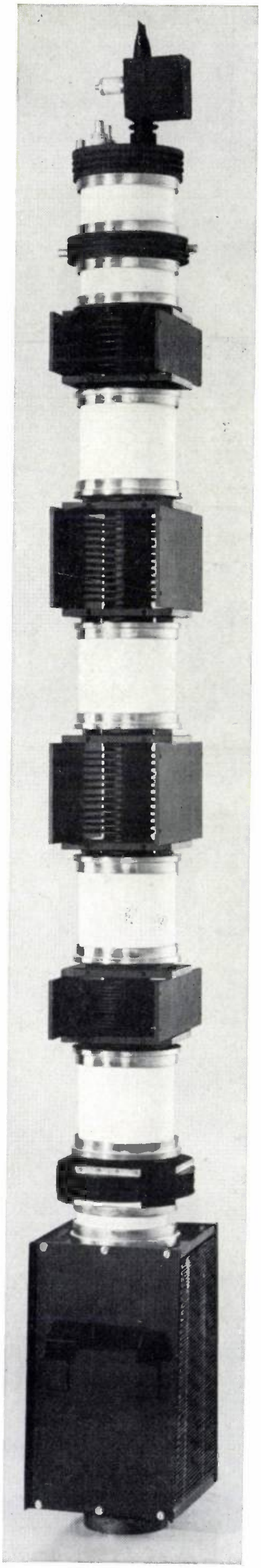


Fig. 3. Photograph of the type YK 1001, four-cavity klystron without focusing magnets or resonant cavities. The getter ion-pump is visible at the top of the valve.

←

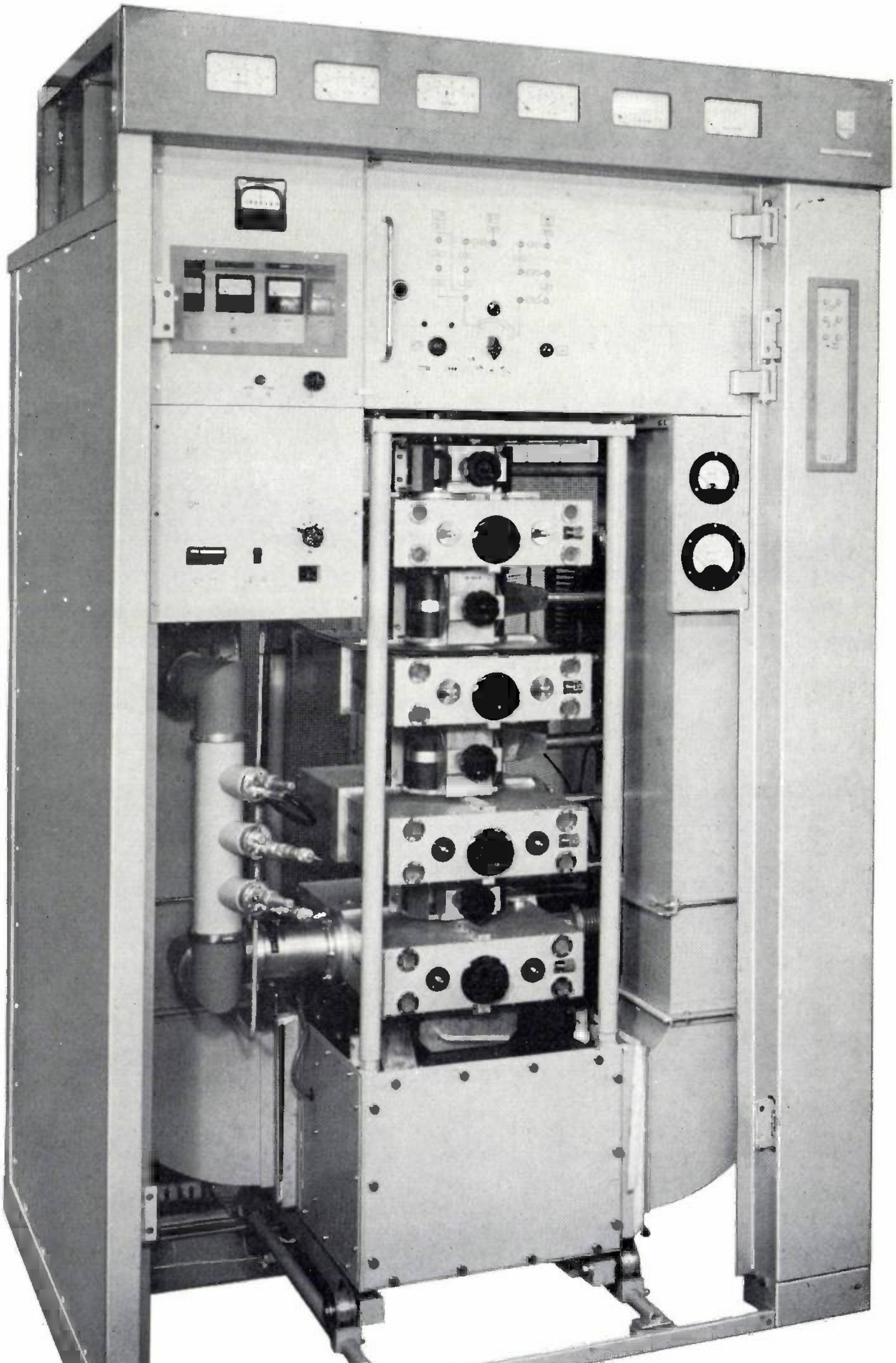


Fig. 4. Arrangement of a type YK 1001 klystron in the output stage of the 10 kW vision-transmitter.

sent trend towards television stations which are not permanently attended, the time involved in travelling to and from the transmitting station is of far more consequence than the time required to change the valves. It is therefore preferable that such changes should be made in the course of routine maintenance visits.

The fact that klystrons have a much longer life than tetrodes economically outweighs the lower purchase price of tetrodes. With regard to the power consumption of a klystron transmitter, the efficiency of a klystron operated as a class A amplifier can be considerably improved by appropriate measures, which will be discussed in greater detail below. This and the fact that the total number of amplification stages can be smaller in a klystron transmitter mean that the overall efficiency of a klystron transmitter is not inferior to that of a transmitter equipped with tetrodes, although the tetrodes themselves, being connected as class B or class C amplifiers, are themselves more efficient than the klystron.

As already observed, a large part of the lower sideband of the amplitude-modulated vision-transmitter carrier has to be suppressed. It is an advantage of the four-cavity klystron that the gain of the valve can be made markedly selective — a point to which we will return — so that the filter inserted in the circuit prior to the modulated stage can be simplified still further.

Thus, although from the point of view of design and economy the advantages of klystrons approximately balance those of triodes and tetrodes, the reliability of the klystron transmitter seems to tip the balance in its favour, and it was this consideration which determined the principle adopted in the Philips 10 kW and 20 kW transmitters.

The type YK 1001 klystron

As the design principles embodied in the construction of the type YK 1001 11 kW and type YK 1061 22 kW klystrons are very similar and the klystrons have practically identical dimensions, it will suffice here to describe a number of details of the YK 1001.

Fig. 3 shows this valve without the accessories that are visible in *fig. 4*, which is a photograph of the same valve arranged in position in the output stage of the 10 kW transmitter. The valve is best described with reference to *fig. 5*, which combines a diagram of a section through the klystron with the circuit in which the various components of the valve are included.

The electron beam traversing the entire valve in the longitudinal direction is provided by an impregnated cathode [3]. The life of this very ruggedly constructed type of cathode is favourably influenced here by the fact that the cathode is clear of the regions in which high-frequency alternating fields occur, and this enables the cathode to be given ample dimensions.

The cathode has a concave emitting surface, so that the electrons emerging from it are already beamed to some extent; the focusing electrode which the electrons now pass and which has a negative potential of about 300 V relative to the cathode, ensures that the electrons are sharply beamed before they enter the valve proper. Before this happens they also pass an accelerating anode which, like the walls of the valve, is at earth potential. As the cathode is maintained at approximately –18 kV, the electrons enter the drift space at full velocity.

On leaving the transit region, the electrons encounter the collector. As will be seen from the diagram, the latter has a voltage of about –5 kV relative to the wall of the drift space region. The result is that the electrons, on leaving the last section of the drift space, are decelerated and arrive at the collector with reduced velocity. In this way a portion of the energy imparted to the electrons during their acceleration is recovered and the efficiency is improved.

When the depth of modulation is large, some electrons will already have a low velocity on leaving the drift space. To prevent these electrons from going only part of the way to the collector and returning to the last section of the drift space, the collector is shaped so as to impede these returns as much as possible.

Although the electrons, on entering the drift space, are concentrated by the operation of the focusing electrode, the beam will tend to disperse again owing to the repelling effect of the electrons on each other. The length of the drift space in relation to its width is so large that if no countermeasures were taken some of the electrons would strike the wall of the drift space, eventually causing excessive heating of the klystron wall. It is therefore necessary to maintain a focusing effect along the entire length of the beam. The most economical means of doing this is by using an axial magnetic field. As the beaming effect is independent of the direction of the field, so long as the latter is axial, it is permissible to use either a field which has a fixed direction over the length of the klystron or a field that reverses its direction several times along the axis of the valve.

In the earliest klystrons to appear on the market a field of constant direction was employed which was generated by a number of circular coils surrounding the klystron. As will be seen in fig. 4, such coils do not figure in the YK 1001 klystron. They have been replaced by a number of permanent magnets made of ferroxdure. It can be shown [4] that by employing an alternating field (i.e. alternating in space but not in time) which is generated by a number of small, per-

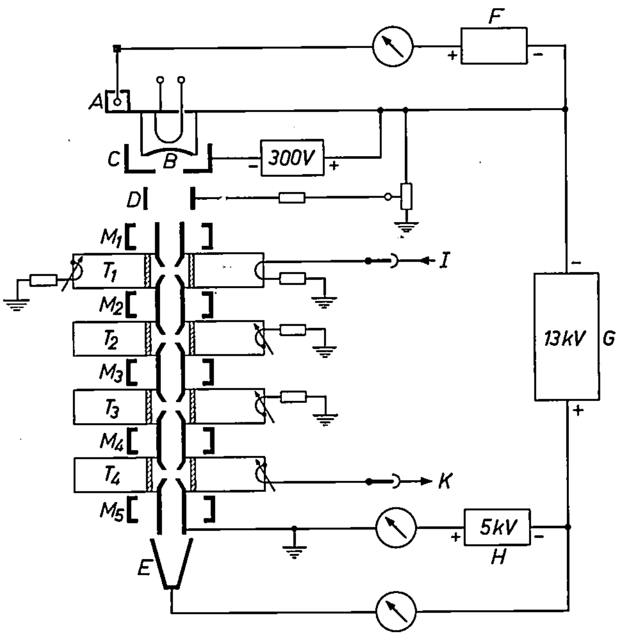


Fig. 5. Combined cross-sectional drawing of the type YK 1001 klystron and diagram of the circuit in which it is used. *A* ion pump; *B* cathode; *C* focusing electrode; *D* acceleration anode; M_1 to M_5 focusing magnets; T_1 to T_4 resonant cavities; *F* power supply rectifier for the ion pump; *G* 13 kV HT rectifier; *H* 5 kV HT rectifier; *I* RF signal from preceding stage; *K* output signal to aerial.

manent magnets, it is possible to save considerably on magnetic material compared with an arrangement whereby a field of fixed orientation is set up with a single large magnet. The use of permanent magnets also means a saving in weight by comparison with the use of magnetic coils. Of even more importance, however, is the fact that the klystron, being no longer surrounded by coils is freely accessible.

The beaming effect obtainable with a magnetic field of alternating direction does not manifest itself until a certain minimum value of the beam voltage [4] is reached. The cathode voltage is therefore brought to its full value at once and then the beam current is increased by means of the accelerating anode. When the HT rectifier is switched on, the voltage control of the accelerating anode is set so that this anode is at cathode potential. Then the accelerating anode is brought to earth potential in steps.

Despite the very careful degassing to which transmitting valves are always subjected, it is necessary to allow for the possibility that when the klystron is being put into service, and to a lesser degree when it is in actual operation, residual gas may be released from various components. The presence of small amounts of gas has a harmful effect on the life of the cathode and on the operation of the valve as a whole. The klystron therefore has a permanently fitted getter ion pump. As fig. 3 shows, this pump is a small cube-shaped box mounted at the top of the tube. It also serves as a vacuum gauge, operating on the principle indicated by Penning [5], with which the vacuum can be measured not only before the valve is put into service but also while it is in operation. The former possibility is important because klystrons have a life of many thousands of hours and consequently the spare valves which have to be kept in reserve at every transmitter are sometimes held in store for a long time before they go into use. A very tiny leak which cannot be detected during manufacture may have reduced the vacuum considerably during that period, without by any means having rendered the valve unsuitable for use, because the ion pump can easily extract the small quantity of gas that has entered. If any deterioration is measured, the pump is switched on for a time before the valve is put into operation.

Regular measurement of the vacuum during operation gives a check on the condition of the valve, thus making it possible to ensure that the valve does not fail at an undesirable moment, e.g. while the transmitter is on the air. This, and the regular check on the condition of the cathode provided by measurement of the beam current, make it practically certain that a broadcast need never be interrupted by the failure of a klystron.

Tuning the output stage

It will be seen from fig. 3 that to maintain the vacuum, the modulation gaps in the valve are bridged by ceramic sleeves. As these sleeves are in the r.f. alternating field, the material of which they are made must satisfy very stringent requirements. The material chosen is an aluminium oxide ceramic.

These sleeves are not visible in fig. 4, being enclosed in the resonant cavities bridging the modulation gaps. Each of these cavities takes the form of a rectangular box fitted round the tube in two halves. One such half is shown in *fig. 6*. As can be seen from the photograph, one of the walls of the box is adjustable and thus can be used to tune the cavity to the desired frequency. The photograph also shows an adjustable coupling loop. A loop of this kind is used in the first resonant cavity (T_1 in fig. 5) to inject the RF power excitation needed. In cavities T_2 and T_3 the loop is used to couple an external load resistance to the RF field. Adjustment of the coupling loops and tuning of the resonant cavities together give the desired form of tuning charac-

[3] R. Levi, Dispenser cathodes, III. The impregnated cathode, Philips tech. Rev. **19**, 186-190, 1957/58.

[4] J. T. Mendel, C. F. Quate and W. H. Yocom, Electron beam focusing with periodic permanent magnet fields, Proc. IRE **42**, 800-810, 1954.

[5] F. M. Penning and K. Nienhuis, Construction and applications of a new design of the Philips vacuum gauge, Philips tech. Rev. **11**, 116-122, 1949/50, and A. Klopfer and W. Ermrich, A small getter ion-pump, Philips tech. Rev. **22**, 260-265, 1960/61.

teristic. The coupling loop in the last cavity, T_4, is used to take the RF power from the klystron.

It can be seen in *fig. 7* that the resonance frequencies of the various cavities do not coincide with the carrier frequency. The flatness of the resonance curves of cavities T_1, T_2 and T_3 can be modified by increasing or decreasing the resistive load on the cavities by means of the coupling loops. The flatness of the curve of cavity T_4 is of course determined by the aerial load. The result of connecting in tandem the four circuits tuned in this way is shown at the bottom of fig. 7.

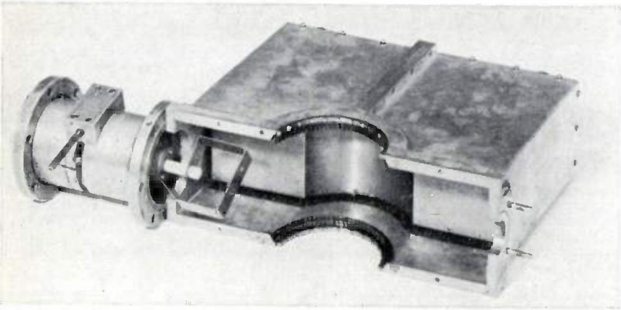


Fig. 6. One half of a box-shaped resonant cavity. One wall can be adjusted to tune the cavity. An adjustable coupling loop can be used to inject or extract RF power.

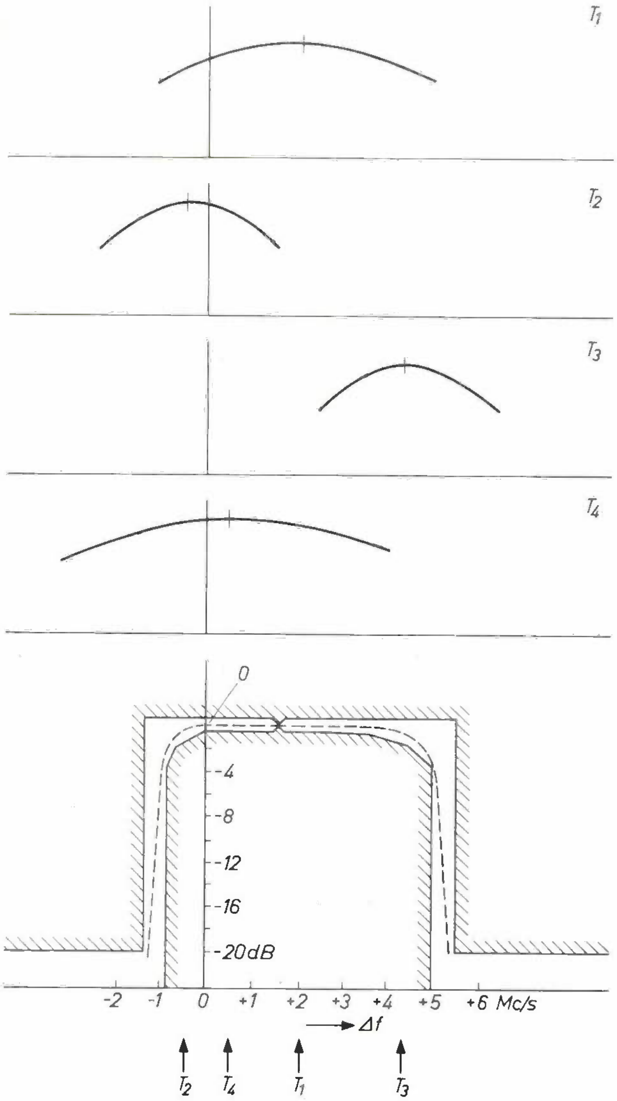


Fig. 7. The resonant cavities are tuned so that the maxima of the amplitude-frequency response curves for the various resonators are slightly to the side of the nominal carrier frequency. The overall gain characteristic of the klystron is therefore as shown at the bottom of the figure.

The range of frequencies over which the resonant cavities can be tuned extends from 470 to 790 Mc/s. The same klystron can also be used for the band V top frequencies, up to 960 Mc/s, but a second set of resonant cavities is then necessary.

The driver stages

Our discussion of the driver stages will be restricted to the stage immediately preceding the klystron — i.e. where modulation takes place — and to the modulation amplifier. The penultimate stage of the transmitter is driven by a signal at carrier frequency. The amplifier valve used in it is the type YL 1100 tetrode. The circuit of which this valve forms part is reproduced in *fig. 8*, which also shows part of the physical construction. The modulator circuit can also be seen in the diagram.

To increase the stability of the amplifier the grounded-grid circuit has been chosen, with a tuned circuit inserted between grid and cathode and another between the grid and anode. The grid has a d.c. connection to earth. Both tuned circuits, as fig. 8 shows consist entirely of concentric conductors.

Conductors *a* and *b* form a coaxial system through which the filament voltage is connected to the valve. The space inside conductor *b* is therefore free of RF fields. Conductors *b* and *c* together constitute the grid-cathode circuit. At the top, the grid-cathode capacity terminates the coaxial line formed by *b* and *c*. A variable capacitor at the other end of this coaxial line is used to adjust the grid-cathode voltage to a maximum.

Cylindrical can *d* forms a cup over the top of conductor *c*: the space between the two forms the grid-anode circuit, which is tuned with a piston fitted in the lower part of the box *d*. The cover top of can *d* is removable, so that the valve is accessible for replacement. All connections to the electrodes of the valve are made by means of circular arrangements of spring contacts and replacement is therefore a simple operation.

The banks of spring contacts for both the screen-grid, anode, and control-grid connections are mounted on flat metal rings. The ring for the control grid contacts is fixed to the top of conductor *c*. This ring and the one for the screen-grid contacts, which is fitted on top of it, are separated by a mica ring. The capacitance

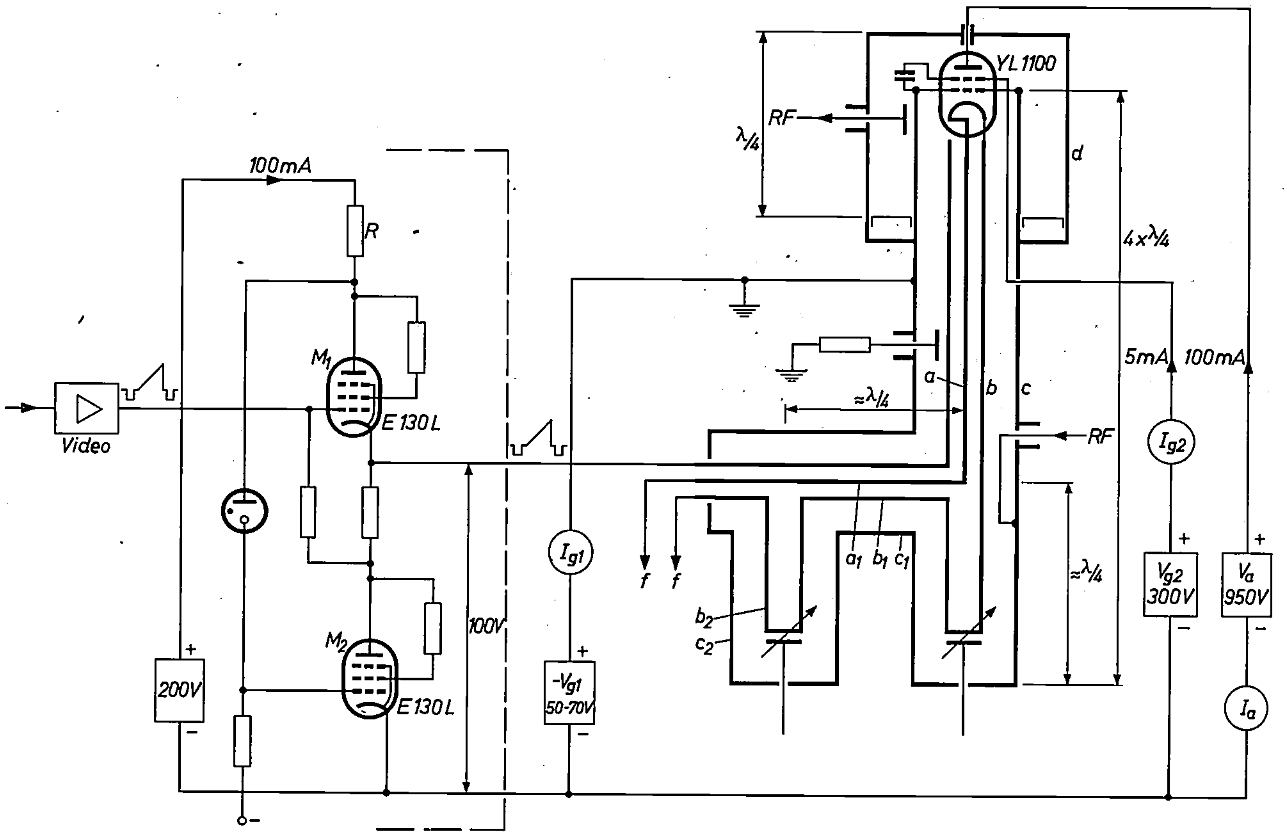


Fig. 8. Circuit diagram of the penultimate amplification stage of the video transmitter and the modulator. M_1 modulator valve; M_2 cathode impedance valve for M_1; *a-b* coaxial pair via which the heater voltage is taken to the YL 1100 valve; *b-c* coaxial cavity between the grid and cathode of the YL 1100 valve; b_1-c_1 quarter-wave line to coaxial cavity b_2-c_2, forming a short-circuit for the transmitting frequency and preventing the r.f. voltage from getting into the modulator.

hus created keeps the r.f. potential of the screen grid the same as that of the control grid, while the two d.c. potentials differ. A similar arrangement ensures that the anode d.c. voltage can be connected to the valve without d.c. connection with the box.

As we have already said, the transmitter has to be tunable over the frequency range 470 to 790 Mc/s. The length of conductors *b* and *c* is chosen in such a way that the system is electrically approximately a whole wavelength long for the frequency in use and consequently there will be voltage antinodes at the extremities of *b* when the transmitter is correctly tuned. About a quarter wavelength from the lower end there will be a voltage node — and hence also a current antinode — and it is here, at one side, that the coupling loop is fitted via which the grid-cathode circuit is excited by the preceding amplifier stage. At the top end ihe r.f. power is drawn from the anode circuit with the aid of a capacitive coupling.

Opposite the coupling loop, there is a branch system comprising conductors a_1, b_1 and c_1. The video modulation voltage is applied to the amplifier tube via conductor *b*1. To ensure that no r.f. power finds its way into the modulator via b_1, another branch b_2-c_2 is fitted about a quarter wavelength from the centre of *b-c*. At the end of this branch there is a variable capacitor which can be adjusted so that circuit b_2-c_2 assumes a condition of series resonance. The branch then in effect sets up a short-circuit at the point of connection, thus preventing r.f. power from entering the modulator.

The modulator amplifier

The extremely high power of the klystron means that an r.f. driving power of only 15 W is needed to ensure the full output power of 10 kW. The klystron is connected as a linear amplifier and modulation takes place in the stage which precedes it.

As the r.f. amplifier in fig. 8 operates in a grounded-grid connection, it is not modulated at the grid, but at the cathode. This means that the cathode current of the YL 1100 valve has to be furnished by the modulator but, as this current is only 100 mA, no difficulty arises.

The cathode-grid impedance of the YL 1100 tube forming the load on the modulator output varies considerably with the amplitude of the modulation voltage. To reduce this effect the grid-cathode circuit is loaded by a external fixed resistor which is capacitively coupled to conductor *b*. Nevertheless, the output impedance of the modulator still varies too much to make special measures unnecessary. These measures are designed to reduce the internal resistance of the modulator as far as possible, for only then can a sufficiently linear amplitude response be obtained with variable load.

As can be seen from fig. 8, modulator tube M_1, for which a type E 130 L pentode has been chosen, is connected as a cathode follower. The anode impedance of valve M_2 — also an E 130 L — whose grid is coupled to the anode of M_1, acts as a cathode resistance. Careful design of this negative feedback circuit enabled the internal resistance of the modulator to be reduced to a very low value, measurement showing it to be less than 4 ohms.

Sideband suppression

The video signal has a bandwidth of approximately 5 Mc/s. The amplitude modulation of the vision-transmitter produces two sidebands, one on either side of the carrier frequency, so that a total bandwidth of 10 Mc/s is occupied. It has been agreed internationally to suppress a large part of the lower sideband to save space in the frequency spectrum. In bands I and II, where 7 Mc/s is available per channel, this sideband is cut off in accordance with the curve shown schematically in *fig. 9.* The curve is flat to 0.75 Mc/s below the carrier frequency and reaches zero at a point 1.25 Mc/s below it.

The curve shown in *fig. 10* has been adopted as a standard for the r.f. amplitude response of television receivers. If an r.f. signal is detected whose lower sideband has been suppressed in accordance with fig. 9, the amplitude response of the video signal obtained is flat, except for a 0.5 dB deviation at a frequency of 0.75 Mc/s. This distortion, which is visible at 0.75 Mc/s in a normal picture, has been accepted as unavoidable in bands I and III. Nevertheless, it can be avoided if, as in *fig. 11,* the transmitter response is not allowed to drop until the frequency is 1 Mc/s below the carrier. This solution is a possibility in bands IV and V, because of the extra space available.

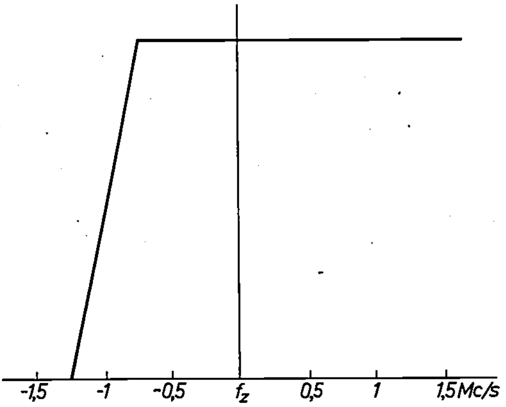


Fig. 9. The lower-frequency end of the amplitude-frequency response curve of the vision-transmitter, shown diagrammatically. The response is flat to 0.75 Mc/s below the transmitting frequency f_z, falling to zero at 1.25 Mc/s below f_z.

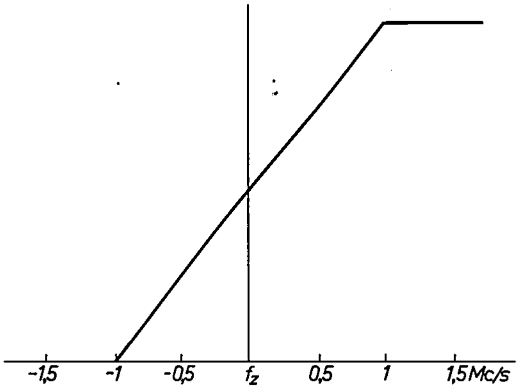


Fig. 10. The amplitude response curve of the average television receiver. It begins to drop 1 Mc/s above the transmitting frequency f_z and reaches the zero line 1 Mc/s below it.

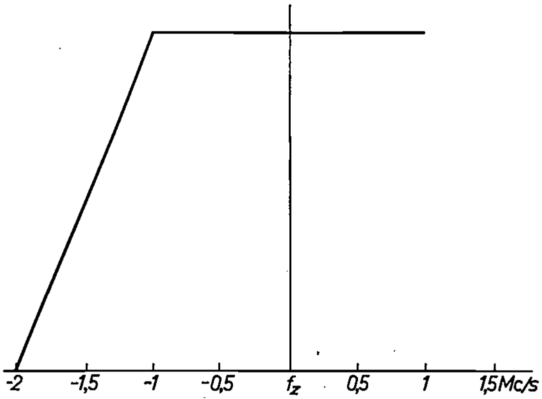


Fig. 11. The Philips vision-transmitter can be tuned so that the residual sideband characteristic of the transmitter assumes the form seen here. In combination with the receiver characteristic shown in fig. 10 it gives an overall response which is entirely flat.

Partial suppression of the lower sideband introduces phase errors into the video signal, and the steeper the slope of the transmitter response curve, the greater these errors will be. Phase errors are also present in the receiver. For economic reasons these are corrected in the transmitter. The phase correction usually is essentially a pre-distortion of the video signal. When sudden level variations take place in the video signal, pre-distortion causes this signal to "overshoot", i.e. there is for a moment a larger variation in level than that warranted by the actual signal. This means that a sudden change from white to black causes temporary overmodulation of the transmitter. As a bandwidth of 8 Mc/s is available in bands IV and V, the response curve can be given a gentler slope in these bands, so

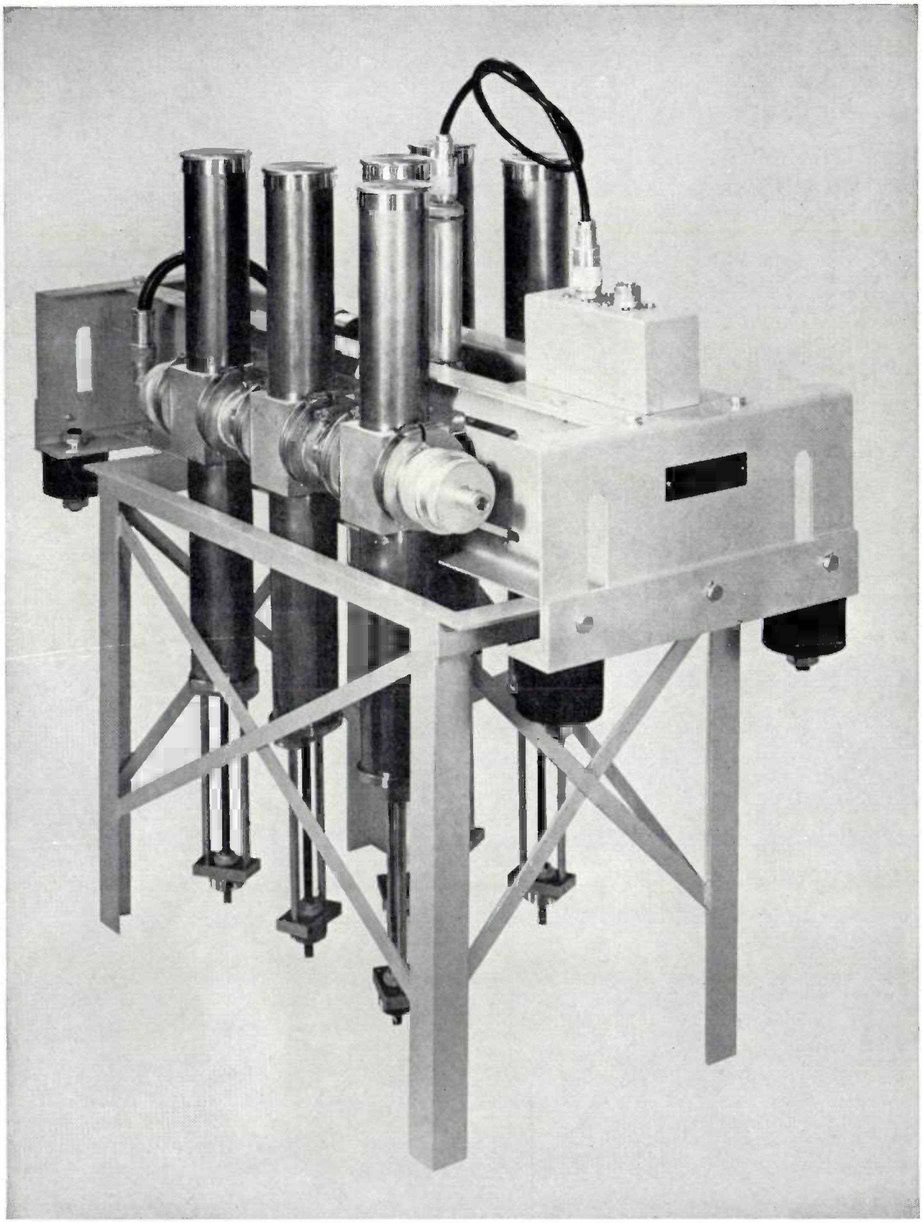


Fig. 12. The component parts of the Nyquist receiver are mounted on a frame. In front, the Nyquist filter; behind it, the notch filter; in between, the detector, and, on the right, the phase correction filter.

that the phase distortion is decreased and less phase correction is needed. The bands IV and V transmitter described here can be adjusted to give a residual-sideband response like that shown in fig. 11.

Monitoring the transmitted picture signal

For adjustment of the transmitter and regular supervision of the transmitted vision signal it is necessary to have a monitoring receiver with an amplitude response which is exactly as shown in fig. 10. As this response curve was first indicated by Nyquist, this receiver is generally referred to as a Nyquist demodulator. Because of the important task this receiver performs, its response curve must not vary. As the input signal to the receiver can be taken straight from the transmitter output, sufficient power is available for the signal to be applied without pre-amplification to a crystal detector via a pair of filters which will be described below. This arrangement not only enables the number of electrical components to be kept down, but it also obviates the use of amplifier valves which, as they age, become an important contributory cause of response curve drift in receivers.

The signal to be monitored passes through a high-pass filter, a band-rejection filter, a detector and a phase-correction filter, in that order. The first two filters determine the amplitude-frequency response of the receiver and must therefore satisfy high standards of stability. They are composed of coaxial elements which, to eliminate the effect of temperature variations on the characteristics, are made partly of Invar. Both filters are accurately adjusted for the transmitter frequency chosen and must therefore be replaced if this frequency is changed.

The first filter, which is used to give the amplitude-frequency response curve the lower-end slope shown in fig. 10, is a high-pass filter generally referred to as the Nyquist filter. The second is a band-rejection filter for preventing the sound transmitter signal, which, of course, is also present in the transmitter output lead, from entering the receiver. This type of filter is known in television engineering as a notch filter. The two filters are mounted on a chassis together with the crystal detector and phase correction filter, as shown in *fig. 12.*

In both the Nyquist filter and the notch filter the r.f. signal is fed in via a coaxial line which has shunt coaxial stubs at quarter-wavelength intervals. These stubs are clearly visible in fig. 12. A cross-section of a stub is given in *fig. 13*; lengths l_1 and l_2 and capacitor C are variable for adjustment purposes.

When the design of the stubs was being determined, it was found that the exact formulae involved calculations which were practically impossible to perform. Use was made therefore of the fact that the behaviour of the admittance of the stubs in the significant range of frequencies is very closely approximated by that of a circuit consisting of a capacitor and an inductor in series, with an inductor or capacitor connected in parallel. The approximate values thus found for the practical dimensions could then be checked with the exact formulae.

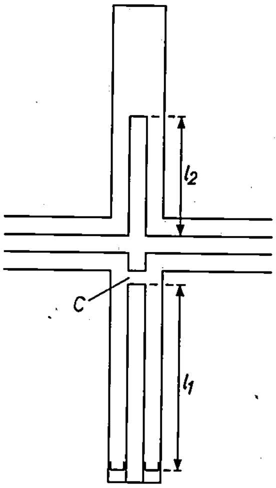


Fig. 13. Schematic cross-section through a coaxial stub as used in the two r.f. filters of the Nyquist receiver.

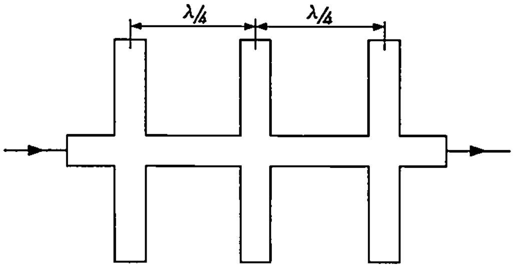


Fig. 14. In the Nyquist filter three coaxial stubs are fitted at intervals of a quarter wavelength along a coaxial line.

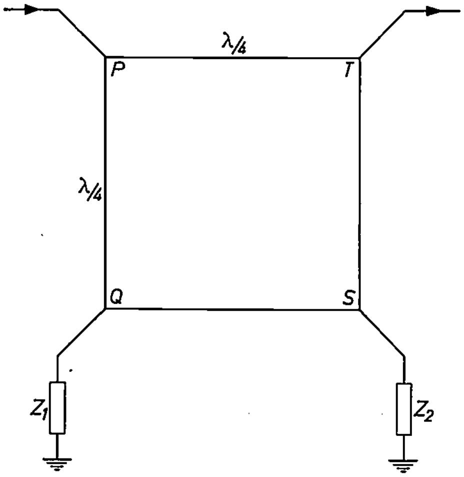


Fig. 15. Schematic diagram of the bridge circuit used in the notch filter.

The design of the Nyquist filter is shown schematically in *fig. 14*. The notch filter, whose function is to keep the vision signal free of interference from the sound signal, must present an attenuation of 40 dB at the sound carrier frequency. If this filter were designed on the same lines as the Nyquist filter, it would not be possible to ensure the desired attenuation with the circuit Q attainable at the frequencies concerned. This is, however, possible with the bridge circuit shown in *fig. 15*. The bridge consists of a closed system of four coaxial lines which are in principle each a quarter wavelength long. The input and output are at two adjacent corners of the bridge, while impedances Z_1 and Z_2, which take the form of coaxial stubs of the same type as used in the Nyquist filters, are connected to the other two corners. Although the elements used have a finite Q, the bridge circuit enables the same effect to be obtained as with elements of infinite Q.

It will be seen from fig. 12 that a third stub is fitted in the notch filter. This is connected to line PT (see fig. 15) and its sole purpose is to simplify adjustment. Finally lines PQ and ST have been made three-quarters of a wavelength long instead of a quarter wavelength for constructional reasons. The complete circuit is shown diagrammatically in *fig. 16*.

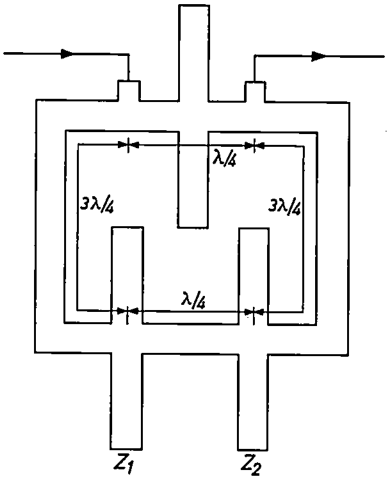


Fig. 16. The notch filter, built up from coaxial lines and coaxial stubs. The stub marked Z_3 is intended solely to facilitate adjustment.

The notch filter, as already mentioned, is followed by the detector, fitted with silicon diodes. The circuit is designed so as to satisfy the very high requirements laid down for linearity of the detection characteristic.

The phase correction filter of the monitoring receiver is composed of a number of bridged-T sections. The properties of this filter are best represented by the "group delay" characteristic, as this is most suitable for practical measurement. The filter is normally given the characteristic shown in *fig. 17* but if desired a filter with a different characteristic can be supplied.

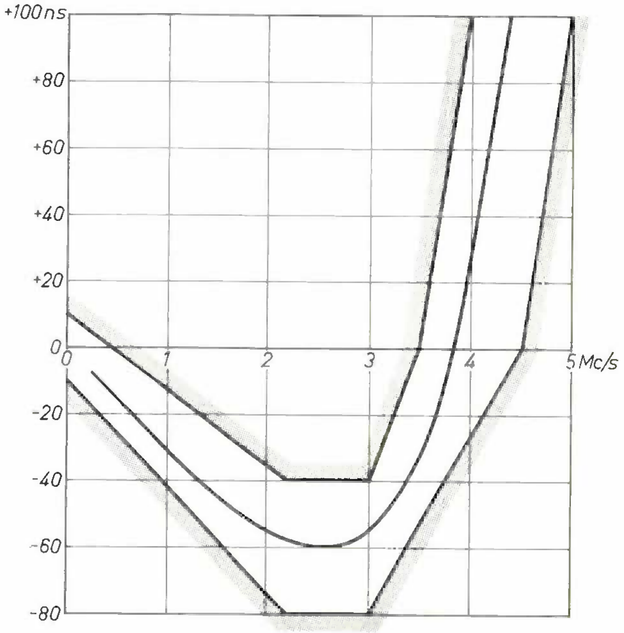


Fig. 17. Group delay characteristic for normal setting of the phase-correction filter in the Nyquist demodulator.

Paralleling transmitters

In the interest of continuity of television broadcasts, special measures are taken to avoid interruptions. For example, equipment can be duplicated. For economic reasons and to be certain that stand-by equipment will not fail at the very moment its services are needed, it is desirable to keep this equipment in operation. A good solution to the problem is, having decided upon the operating power of the transmitter, to have this power provided by two transmitter units, each of half the power and operating in parallel. If one of the units fails the transmission will go on without interruption, although at reduced power.

When two transmitters are connected in parallel in this way, it can be assumed that their frequencies are determined by a common crystal oscillator, and that their phase relationship is therefore permanently fixed. Two courses are then available. The first possibility is to feed both output signals to a bridge circuit for combination. The composite signal then goes from the bridge circuit to the aerial via the common aerial feeder. The drawback of this arrangement is that a broadcast can still be interrupted by a fault in the bridge circuit, the feeder or the aerial. If it is desired to eliminate even that possibility, the aerial can be divided into two identical parts and the two transmitters connected to these half aerials by separate feeders. No bridge circuit is then needed and the signals combine after radiation from the aerials.

Although the second arrangement reduces the chance of interruptions during broadcasts to a minimum, it was found, when Philips Telecommunicatie Industrie set up a transmitter operating on this principle, that certain precautions have to be taken in building and tuning the two halves of the aerial if unwanted side effects are to be avoided.

These side effects are due to the fact that in addition to the main lobe, whose axis generally slopes slightly downwards towards the horizon, the radiation pattern also comprises a number of side lobes. The minima between the various lobes are due to the fact that the signals of the two aerial halves in the direction of these minima are practically cancelled out by interference. In the case of a clear-cut minimum, when cancellation is practically complete, the residual component of the signal will depend very closely on the phase relation between the two signals radiated by the two half-aerials. As this phase relation also depends on the momentary frequency determined by the modulation, the picture quality may be very unfavourably affected, and negative pictures may even occur, in areas situated in the direction of radiation pattern minima. It will be obvious from what has been said that this drawback can be overcome by designing and adjusting the aerial so that no very pronounced minima occur.

Summary. The article describes a number of problems confronting the designer of television transmitters for ultra-high frequencies (470-960 Mc/s). When allowance is made for propagation conditions at these frequencies and the gain that can be attained with the aid of the directional effect of the transmitting aerial, transmitter outputs of 10-40 kW are found to be necessary. These outputs and the frequencies used bring designers to the limit of what is possible with triodes and tetrodes. Two types of transmitter built by Philips for these bands therefore employ klystrons. A description of the 10 kW four-cavity klystron type YK 1001 is followed by a description of the output stage tuning procedure. The four resonant cavities can be tuned in such a way that the conventional sideband suppression filter is made largely superfluous. The penultimate stage of the vision-transmitter employs a tetrode in grounded-grid connection and uses coaxial techniques throughout. The cathode current for this stage is supplied by the modulator, which has a very low internal resistance. The receiver used to monitor the quality of the transmitted signal has to possess extremely constant characteristics and is therefore of special design. The filters incorporated in it are composed of coaxial elements made partly of Invar. Finally the article points out possible sources of degradation in picture quality when transmitters are connected in parallel.

The early history of telegraphy

G. R. M. Garratt

621.394(091)

It has been said that the history of the art of communication is the history of the human race. While such a statement is perhaps too sweeping, it is undoubtedly a fact that the development of communication is an integral part of the growth of civilization. Every improvement in the speed and facility with which thoughts and ideas can be exchanged has had its social or economic effect, and we can appreciate fully the significance and trend of historical, social and political developments only if we view them against the background of the contemporary state of the art of communication.

Momentous changes in the art of communication were initiated in the first half of the 19th century, and in the context of this special issue of Philips Technical Review it seems worth while to give some attention to this most interesting period. The changes in question were marked by the advent of the electric telegraph, an invention based on the great discoveries in the field of electricity which were made during that period, and we shall see how the evolutionary stages of the electric telegraph correspond with discoveries in static electricity, galvanism, electrolysis and electromagnetism.

The visual telegraph

The rapid transmission of intelligence has been an aim of mankind from the earliest times and many systems of *visual* communication have been employed throughout the ages. It was not until the last decade of the 18th century, however, that the visual telegraph reached the zenith of its development in a vast communication system covering more than half a continent. In retrospect it is interesting to note that this climax occurred just at the time when the early forms of the electric telegraph were about to make their appearance and it will be of value if we first briefly review this amazing development of the visual system.

Communications are vital to the conduct of all military operations and they can never have been more necessary than they were to the French forces in 1793 when, torn by the internal excesses of the Revolution and attacked by enemies on every side, defeat might well have seemed inevitable. When, in 1799, a politician in The Hague tried to explain the causes of the astonishing military successes of the French which had occurred in the meantime, he mentioned several factors: "*. . . courage, aided by inventiveness which yielded the useful telegraph, the application of the balloon, an ample production of saltpetre and ingenious strategic plans*" [1]. It is certainly not without significance that it was the telegraph, the device which had permitted coordination of efforts on different fronts, which was given priority of mention in this quotation. The telegraph referred to was the visual telegraph or Semaphore devised by Claude Chappe.

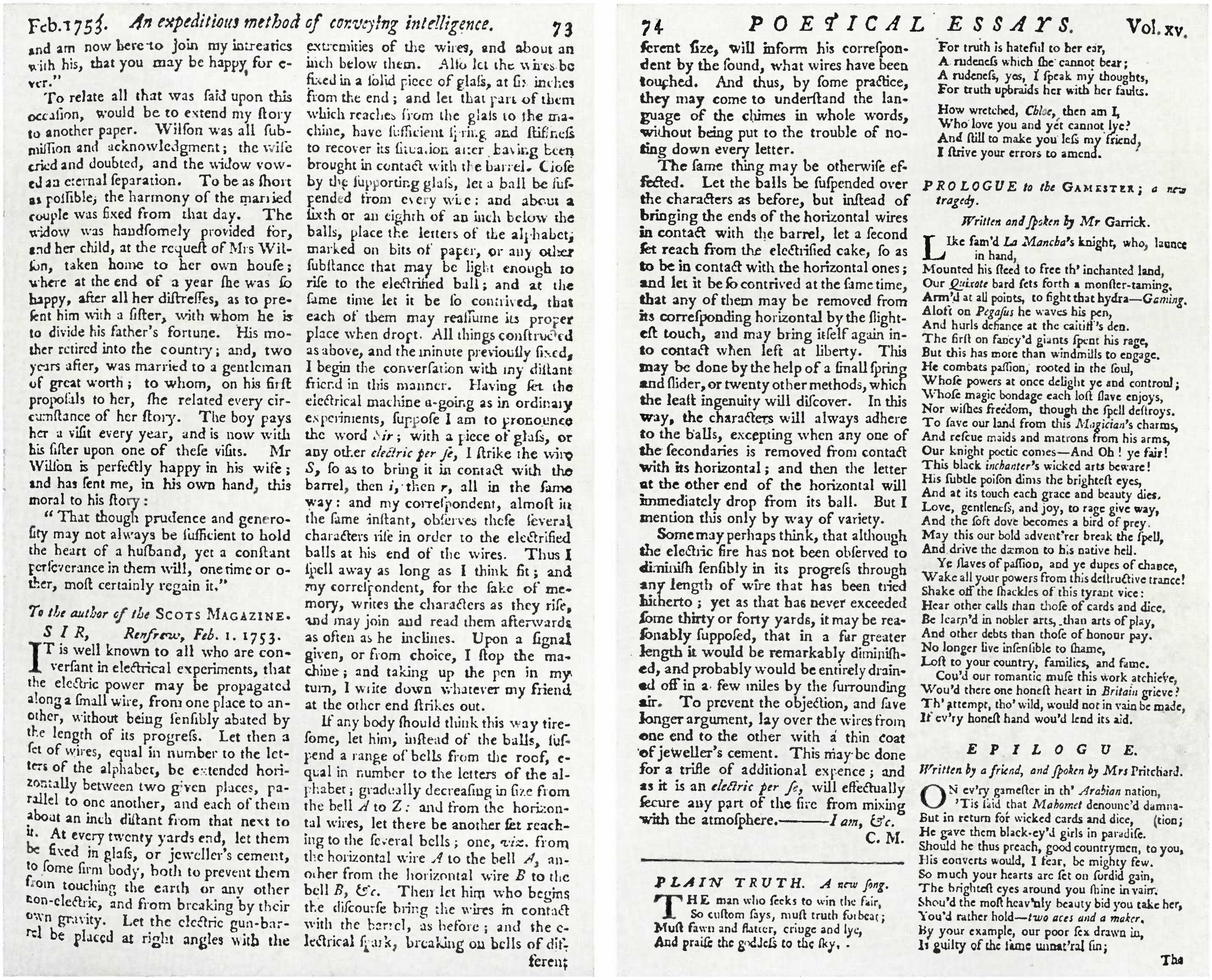
Claude Chappe — nephew of a well-known astronomer— had shown a keen interest in science when still a young man and had published several articles in the "Journal de physique" before the age of twenty. Early in 1790 he became interested in the problem of military communications and, together with his brothers, he started experimenting and devised a system based on the use of electricity. Unfortunately knowledge in this field was still insufficient (e.g. difficulties of insulation were insuperable at the time), and even with Chappe's energetic personality as a promoter, the experiments were unsuccessful. Chappe soon abandoned the idea and turned to the development of an optical system.

In the revolutionary atmosphere which prevailed in France at this period, it is not surprising that Chappe's conspicuous experiments earned him the suspicion of the fanatical crowds and on two occasions the apparatus which he had set up at the Etoile was torn down and destroyed. Chappe, however, persevered and by the early summer of 1793 he had completed a practical system which, when it was finally examined by a commission of the Convention, gave such excellent performance that Chappe was accorded the title of Ingénieur-Télégraphe and was immediately commissioned to erect a chain of telegraph stations between Paris and Lille, a distance of 145 miles.

Before continuing on the history of these and other chains, let us have a look at Chappe's apparatus in order to appreciate why his system was so much more successful than the numerous earlier attempts at visual telegraphy.

Chappe's apparatus is illustrated in *fig. 1.* At the top of a vertical pole a long wooden beam was pivoted at its centre so that it could be rotated in a vertical plane. Slender arms, rotatable in the same plane, were fitted at each end of the beam and signalling was a-

Mr. G. R. M. Garratt, M.A.(Cantab.), M.I.E.E., is head of the Communications Department of the Science Museum, London. — The present article is in part based on previous publications by the author on the same subject, especially chapter 22 in vol. 4 of the History of Technology, edited by Ch. Singer et al., Oxford University Press 1958.



Feb. 1753. *An expeditious method of conveying intelligence.* 73

and am now here to join my intreaties with his, that you may be happy for ever."

To relate all that was ſaid upon this occaſion, would be to extend my ſtory to another paper. Wilſon was all ſubmiſſion and acknowledgment; the wiſe cried and doubted, and the widow vowed an eternal ſeparation. To be as ſhort as poſſible, the harmony of the married couple was fixed from that day. The widow was handſomely provided for, and her child, at the requeſt of Mrs Wilſon, taken home to her own houſe; where at the end of a year ſhe was ſo happy, after all her diſtreſſes, as to preſent him with a ſiſter, with whom he is to divide his father's fortune. His mother retired into the country; and, two years after, was married to a gentleman of great worth; to whom, on his firſt propoſals to her, ſhe related every circumſtance of her ſtory. The boy pays her a viſit every year, and is now with his ſiſter upon one of theſe viſits. Mr Wilſon is perfectly happy in his wife; and has ſent me, in his own hand, this moral to his ſtory:

"That though prudence and generoſity may not always be ſufficient to hold the heart of a huſband, yet a conſtant perſeverance in them will, one time or other, moſt certainly regain it."

To the author of the SCOTS MAGAZINE.

SIR, *Renfrew, Feb.* 1. 1753.

IT is well known to all who are converſant in electrical experiments, that the electric power may be propagated along a ſmall wire, from one place to another, without being ſenſibly abated by the length of its progreſs. Let then a ſet of wires, equal in number to the letters of the alphabet, be extended horizontally between two given places, parallel to one another, and each of them about an inch diſtant from that next to it. At every twenty yards end, let them be fixed in glaſs, or jeweller's cement, to ſome firm body, both to prevent them from touching the earth or any other non-electric, and from breaking by their own gravity. Let the electric gun-barrel be placed at right angles with the extremities of the wires, and about an inch below them. Alſo let the wires be fixed in a ſolid piece of glaſs, at ſix inches from the end; and let that part of them which reaches from the glaſs to the machine, have ſufficient ſpring and ſtiffneſs to recover its ſituation after having been brought in contact with the barrel. Cloſe by the ſupporting glaſs, let a ball be ſuſpended from every wire: and about a ſixth or an eighth of an inch below the balls, place the letters of the alphabet, marked on bits of paper, or any other ſubſtance that may be light enough to riſe to the electrified ball; and at the ſame time let it be ſo contrived, that each of them may reaſſume its proper place when dropt. All things conſtructed as above, and the minute previouſly fixed, I begin the converſation with my diſtant friend in this manner. Having ſet the electrical machine a-going as in ordinary experiments, ſuppoſe I am to pronounce the word *Sir*; with a piece of glaſs, or any other *electric per ſe*, I ſtrike the wire *S*, ſo as to bring it in contact with the barrel, then *i*, then *r*, all in the ſame way: and my correſpondent, almoſt in the ſame inſtant, obſerves theſe ſeveral characters riſe in order to the electrified balls at his end of the wires. Thus I ſpell away as long as I think fit; and my correſpondent, for the ſake of memory, writes the characters as they riſe, and may join and read them afterwards as often as he inclines. Upon a ſignal given, or from choice, I ſtop the machine; and taking up the pen in my turn, I write down whatever my friend at the other end ſtrikes out.

If any body ſhould think this way tireſome, let him, inſtead of the balls, ſuſpend a range of bells from the roof, equal in number to the letters of the alphabet; gradually decreaſing in ſize from the bell *A* to *Z*: and from the horizontal wires, let there be another ſet reaching to the ſeveral bells; one, *viz.* from the horizontal wire *A* to the bell *A*, another from the horizontal wire *B* to the bell *B*, *&c.* Then let him who begins the diſcourſe bring the wires in contact with the barrel, as before; and the electrical ſpark, breaking on bells of different

74 *POETICAL ESSAYS.* Vol. xv.

ſize, will inform his correſpondent by the ſound, what wires have been touched. And thus, by ſome practice, they may come to underſtand the language of the chimes in whole words, without being put to the trouble of noting down every letter.

The ſame thing may be otherwiſe effected. Let the balls be ſuſpended over the characters as before, but inſtead of bringing the ends of the horizontal wires in contact with the barrel, let a ſecond ſet reach from the electriſied cake, ſo as to be in contact with the horizontal ones; and let it be ſo contrived at the ſame time, that any of them may be removed from its correſponding horizontal by the ſlighteſt touch, and may bring itſelf again into contact when left at liberty. This may be done by the help of a ſmall ſpring and ſlider, or twenty other methods, which the leaſt ingenuity will diſcover. In this way, the characters will always adhere to the balls, excepting when any one of the ſecondaries is removed from contact with its horizontal; and then the letter at the other end of the horizontal will immediately drop from its ball. But I mention this only by way of variety.

Some may perhaps think, that although the electric fire has not been obſerved to diminiſh ſenſibly in its progreſs through any length of wire that has been tried hitherto; yet as that has never exceeded ſome thirty or forty yards, it may be reaſonably ſuppoſed, that in a far greater length it would be remarkably diminiſhed, and probably would be entirely drained off in a few miles by the ſurrounding air. To prevent the objection, and ſave longer argument, lay over the wires from one end to the other with a thin coat of jeweller's cement. This may be done for a trifle of additional expence; and as it is an *electric per ſe*, will effectually ſecure any part of the fire from mixing with the atmoſphere.——*I am, &c.*

C. M.

PLAIN TRUTH. *A new ſong.*

THE man who ſeeks to win the fair,
So cuſtom ſays, muſt truth forbear;
Muſt fawn and flatter, cringe and lye,
And praiſe the goddeſs to the ſky.

For truth is hateful to her ear,
A rudeneſs which ſhe cannot bear;
A rudeneſs, yes, I ſpeak my thoughts,
For truth upbraids her with her faults.

How wretched, *Chloe*, then am I,
Who love you and yet cannot lye?
And ſtill to make you leſs my friend,
I ſtrive your errors to amend.

PROLOGUE *to the* GAMESTER; *a new tragedy.*

Written and ſpoken by Mr Garrick.

LIke fam'd *La Mancha*'s knight, who, launce in hand,
Mounted his ſteed to free th' inchanted land,
Our *Quixote* bard ſets forth a monſter-taming,
Arm'd at all points, to fight that hydra—*Gaming*.
Aloft on *Pegaſus* he waves his pen,
And hurls defiance at the caitiff's den.
The firſt on fancy'd giants ſpent his rage,
But this has more than windmills to engage.
He combats paſſion, rooted in the ſoul,
Whoſe powers at once delight ye and controul;
Whoſe magic bondage each loſt ſlave enjoys,
Nor wiſhes freedom, though the ſpell deſtroys.
To ſave our land from this *Magician*'s charms,
And reſcue maids and matrons from his arms,
Our knight poetic comes—And Oh! ye fair!
This black *inchanter*'s wicked arts beware!
His ſubtle poiſon dims the brighteſt eyes,
And at its touch each grace and beauty dies.
Love, gentleneſs, and joy, to rage give way,
And the ſoft dove becomes a bird of prey.
May this our bold advent'rer break the ſpell,
And drive the dæmon to his native hell.
Ye ſlaves of paſſion, and ye dupes of chance,
Wake all your powers from this deſtructive trance!
Shake off the ſhackles of this tyrant vice:
Hear other calls than thoſe of cards and dice.
Be learn'd in nobler arts, than arts of play,
And other debts than thoſe of honour pay.
No longer live inſenſible to ſhame,
Loſt to your country, families, and fame.
Cou'd our romantic muſe this work atchieve,
Wou'd there one honeſt heart in *Britain* grieve?
Th' attempt, tho' wild, would not in vain be made,
If ev'ry honeſt hand wou'd lend its aid.

EPILOGUE.

Written by a friend, and ſpoken by Mrs Pritchard.

ON ev'ry gameſter in th' *Arabian* nation,
'Tis ſaid that *Mahomet* denounc'd damna-
But in return for wicked cards and dice, (tion;
He gave them black-ey'd girls in paradiſe.
Should he thus preach, good countrymen, to you,
His converts would, I fear, be mighty few.
So much your hearts are ſet on ſordid gain,
The brighteſt eyes around you ſhine in vain.
Shou'd the moſt heav'nly beauty bid you take her,
You'd rather hold—*two aces and a maker.*
By your example, our poor ſex drawn in,
Is guilty of the ſame unnat'ral ſin;

The

Fig. 4. Facsimile of two pages of the Scots Magazine of 17th February 1753, showing the letter to the editor signed "C.M." which contained the earliest proposal to use electricity for the purpose of communication.

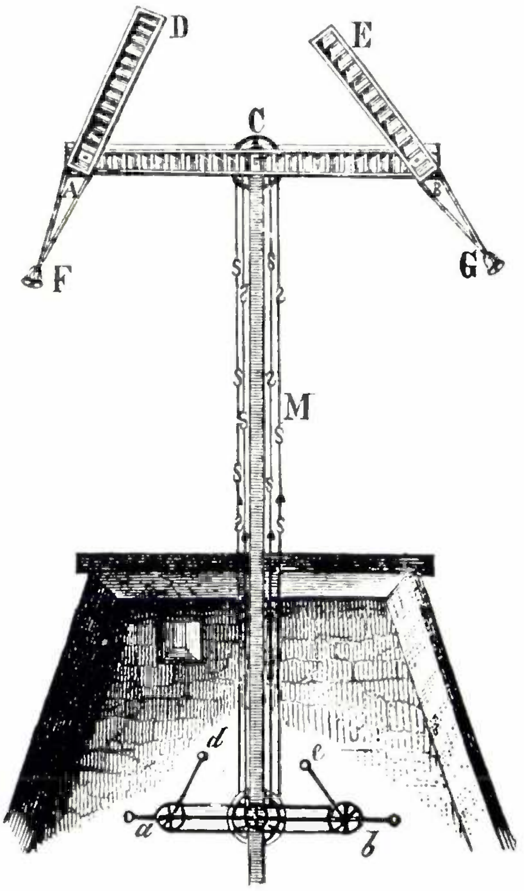


Fig. 1

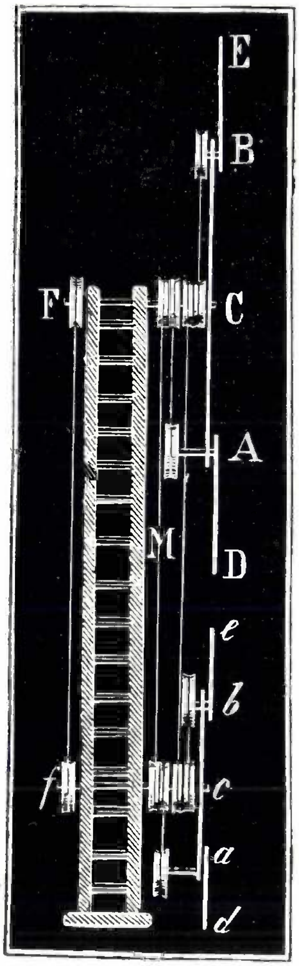


Fig. 2

Fig. 1. Chappe's optical telegraph. A pole on top of a tower carries a pivoted beam (*AB*) with two rotatable arms (*D*, *E*). The position of the beam and the arms is controlled from inside the tower by means of the "répétiteur" (*a*, *b*, *d*, *e*).

Fig. 2. Rope-and-pulley drive for the independent rotation of the three elements of the telegraph. (Figs. 1 and 2 from: H. Schellen, see note [3].)

chieved by altering the positions of the beam and the two arms. The beam and the arms were constructed from open rectangles filled with a series of wooden vanes, a construction resulting in low cost, light weight and little susceptibility to damage by storms, the latter advantages making it feasible to install the apparatus on high towers. Thus the apparatus, even with its modest dimensions could easily be viewed from a distance of 8-12 miles by means of a telescope (the use of which for visual telegraphy had already been proposed by Robert Hooke a century ealier) and even in an unfavourable light the silhouette of the apparatus could then be seen distinctly enough to allow of 196 different signals. Chappe made a judicious choice from those possibilities, establishing a code with 98 of the most distinct signs for the letters of the alphabet, nummerals, etc., and using the other 98 signs only for service instructions ("signaux réglementaires"). Means were also provided for signalling at night, with the aid of several lanterns fixed to the signalling arms. In addition to all the above-mentioned features, however, important though they were from a practical and an economic point of view, Chappe's apparatus possessed an ingenious mechanical system to which it probably owed much of its success. The setting of the beam and the two arms was controlled from the interior of the tower by means of ropes or chains and a system of levers whose configuration corresponded on a smaller scale with those of the main beam and signalling arms on the top of the mast. This mechanism, the so-called "répétiteur", was so designed as to enable either of the two arms or the main beam to be set quite independently of the others (*fig. 2*) and was instrumental in permitting a high rate of signalling. Moreover, when a message was being propagated via a chain of stations, personnel at one station could conveniently imitate a signal observed at the preceding station, without watching the signalling arms of their own pole and also without consulting a code for translatin gthe signals— a considerable asset in those times when personnel skilled in the procedure of coding and decoding were not readily available.

Chappe's first chain of telegraphs, between Paris and Lille, consisted of 15 stations. It was completed in July 1794 and one of the first messages to be transmitted via the system was a report sent from Lille to Paris to inform the government that their forces had recaptured the town of Quesnoy.

The successful working of the Chappe telegraph during the autumn of 1794 quickly led to the establish-

[1] Quoted after E. A. B. J. ten Brink, Publ. Genootschap Napoleontische Studiën, No. 10, 1957, page 347. Some other interesting and little known facts on the Chappe telegraph have also been borrowed from this publication for this article.

ment of other chains. One connected Paris with Strasbourg and another Paris with Brest. Over such long distances the number of stations in a chain was often considerable, especially if no high grounds were available for erecting the signalling towers. For example there were no fewer than 50 between Paris and Strasbourg. The Paris-Lille line was extended to Amsterdam in 1810, and in the archives at The Hague there are some interesting documents concerning this (*fig. 3*); these documents give evidence of the problems with which the work in organizing telegraph lines was beset.

One of these problems has already been mentioned: the enlisting and training of personnel. Large numbers of reliable persons were required for manning the numerous stations, and in view of the serious consequences which negligence by personnel at a single station could have, the maintenance of rigorous discipline was of prime importance. Claude Chappe, despairing of the worries of administration and of rivalry tragically ended hi slife by suicide in 1805, but his work was continued by his brothers who had aided him from the beginning and who had been appointed to high positions in the telegraph administration. In fact the Chappe system might well be called an achievement of the whole Chappe family [2].

During several decades the Chappe system of telegraphs continued to render invaluable service to the French government, but it was never used for private or commercial purposes. New lines were installed as late as 1823 (Paris-Bayonne), and when the system was finally closed down in 1852 it comprised a total of 556 stations with a total length of more than 4000 km.

Other European countries soon followed the French example in making use of visual telegraphy. Reports of the working of the original Chappe line from Paris to Lille began to reach England during the autumn of 1794, but when one recalls the traditional enmity which then existed between England and France, it could scarcely be expected that the British authorities could bring themselves to adopt an identical scheme. A number of quite similar proposals were in fact submitted to the Admiralty and the one adopted employed a large frame with six movable shutters. In 1796 a chain of 15 shutter-telegraphs, erected at a total cost of £ 3750, was opened between London and Dover. A chain to Portsmouth followed immediately, and later, spurred by the fresh outbreak of hostilities in 1803, lines to Plymouth and Yarmouth were brought into use. Between the years 1811 and 1816, however, the Admiralty's shutter-telegraphs were gradually replaced by the more efficient Chappe system, the greater speed of signalling having at last overcome national pride and prejudice. The last of these lines, that to Portsmouth was not finally closed until 1847 — a year after the Cooke-Wheatstone electric telegraph had been set up along the railway between London and Portsmouth.

For many years the Chappe telegraphs with their mysterious and incessantly moving arms had fascinated the public and, in some places, had even formed an attraction for tourists. The picture of these strange erections falling in disuse and being made obsolete by more modern devices, inspired sentimental reflections, like this one (Gustave Nadaud, 1849 [1]):

"Que fais-tu mon vieux télégraphe,
Au sommet de ton vieux clocher,
Sérieux comme une épitaphe,
Immobile comme un rocher?

. . . .

Tu fus l'énigme de notre âge:
Nous voulions, enfant curieux,
Deviner ce muet langage
Qui semblait le parler des dieux,
Lorsque tes bras cabalistiques
Lançaient à l'horizon blafard
Les mensonges diplomatiques
Interrompus par le brouillard."

Telegraphy based on static electricity

One of the most interesting documents in the whole history of telegraphy is a remarkable letter which was published in the Scots' Magazine on 17th February 1753. In this letter, which is reproduced in *fig. 4*, a writer whom we know only by his initials C.M. described in considerable detail a scheme by which he proposed to "hold a conversation with a distant friend by means of electricity". This was probably the earliest fully-fledged proposal to use electricity for the purpose of communication.

It is important to realise that when this proposal was made knowledge of electricity was in a very elementary state. The frictional electric machine had only recently been invented, as had also the Leyden jar; phenomena of "static" electricity such as the electric spark and the electric shock were well known, but the Galvanic cell, the behaviour of an electric current and the phenomena of electro-magnetism had yet to be discovered.

In his letter, C.M. proposed that a set of wires, one for each letter of the alphabet, should be provided

[2] Cf. the book published by Claude Chappe's elder brother Ignace-Urbain: "Histoire de la télégraphie", Paris 1824.

Fig. 3. Seal used by Chappe, "Directeur Télégraphique", on letters written in 1810 to governor Lebrun at The Hague concerning the extension of the Paris-Lille line to Amsterdam. (From: E. A. B. J. ten Brink and C. W. L. Schell, Geschiedenis van de Rijkstelegraaf 1852-1952, The Hague 1954.)

between the two friends who wished to communicate with each other. From every wire a ball should be suspended at each end and the letters of the alphabet, marked on bits of paper, placed below the balls. In order to transmit a series of letters, the sender was instructed to bring the respective wires in turn into contact with the "barrel of an electrical machine set a-going" and the receiver would then note the letters as they rose in turn towards the electrified balls.

So far as we know, no attempt was ever made to put C.M.'s ideas into practice. Since his method of "coding" required a separate wire for each letter, it would have proved very expensive but the main difficulty would have consisted in the insulation of the wires C.M. seems partially to have foreseen this difficulty since in his letter he proposes means to "secure any part of the electric fire from mixing with the atmosphere".

Many inventors coming after C.M. were defeated by this problem. We have already mentioned Claude Chappe, who in 1790 experimented with an electric telegraph, based on static electricity, but abandoned it because of the difficulties of insulation. Chappe's experiments are interesting nevertheless because he applied a completely different method of coding. His scheme made use of two clocks, one placed at the transmitter, the other at the receiver, the seconds dials of which were marked with the numerals 0-9. The clocks were carefully adjusted ("synchronized" we would call it) so that the seconds hands moved in unison, always pointing to the same numerals at the same instant. Only two wires connected the two stations. Both operators had to observe their clock and the sender would denote a particular numeral by discharging a Leyden jar through the wires at the moment when the clocks indicated this numeral. Incidentally, this principle of "time coding", so important in later times, was already known to the Greeks who had thought of employing it with a system of visual telegraphy.

One of the last of the telegraphs depending on static electricity was that proposed in 1816 by Sir Francis Ronalds. His coding scheme was similar to that proposed by Chappe though rather more elaborate (*fig. 5*). He carried out many practical experiments and we can guess how he, too, came to grips with the problem of insulation. In the garden of his London home he put up a pair of large wooden frames between which he suspended eight miles of wire (*fig. 6*) and he proved that the discharge of the wire at one end was indicated instantaneously at the other. Experience with this arrangement probably taught him that static electricity

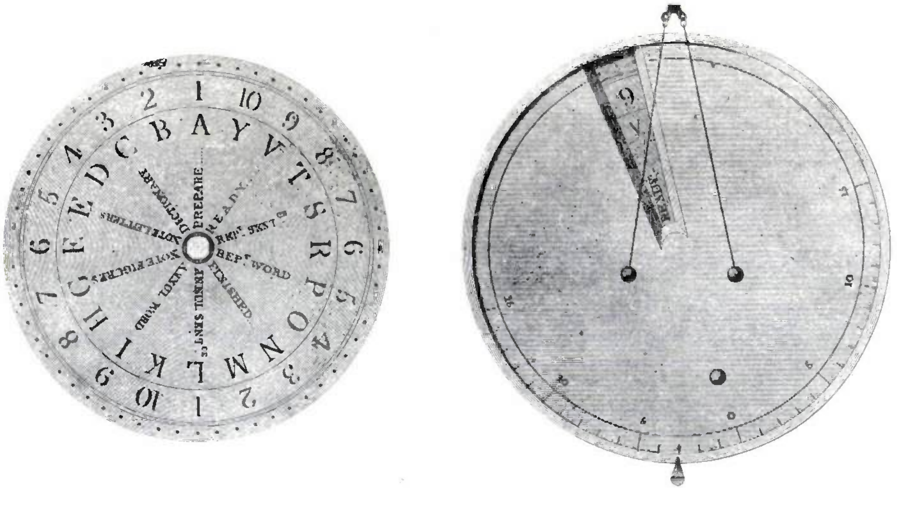


Fig. 5. Proposed form of the indicator of Ronalds' electrostatic telegraph, based on time-coding. Both the sender and the receiver have a disc exhibiting letters, numerals and orders, rotating synchronously at both stations behind a fixed window. When the desired letter or numeral appeared in the window, the sender discharged a Leyden jar and this was indicated at the receiver by an electroscope.

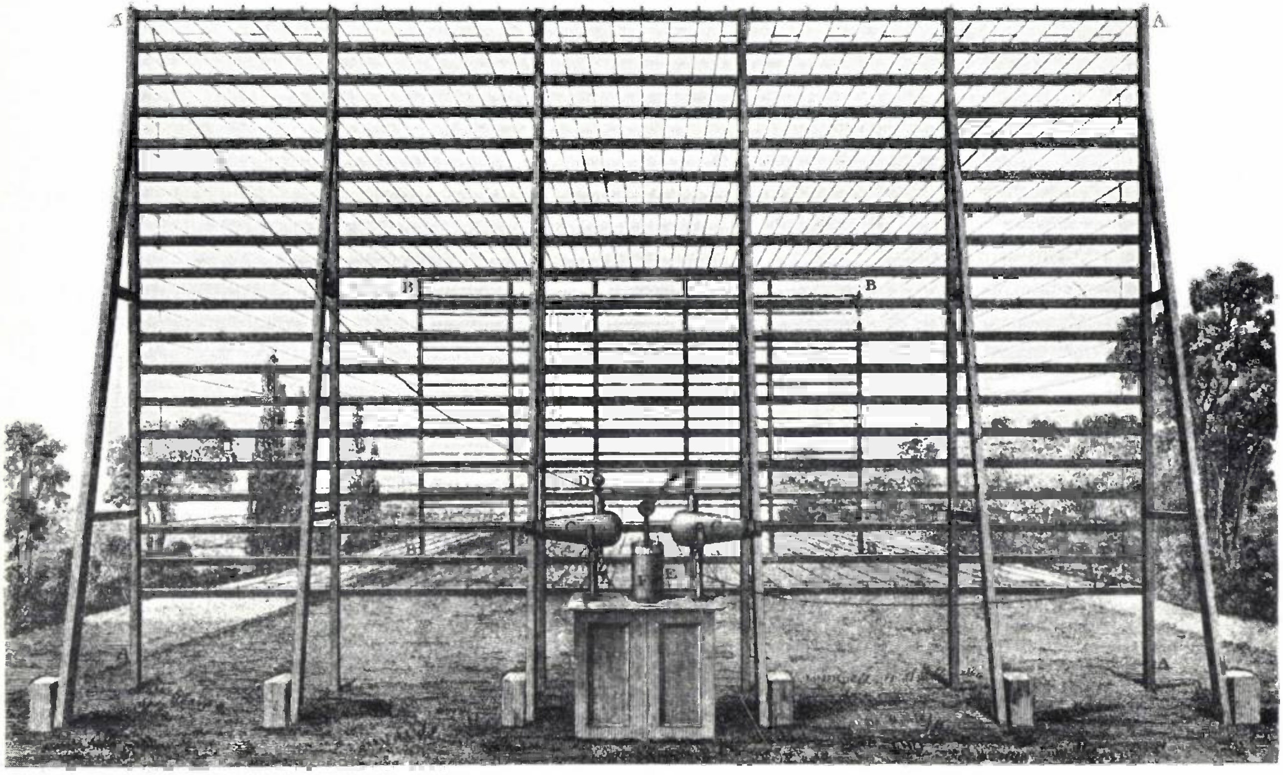


Fig. 6. Experimental arrangement of Ronalds' electrostatic telegraph in the garden of his house at Hammersmith, London. Between the large frames, Ronalds had suspended 8 miles of wire.

was a capricious "fluid", fleeting in its way and almost literally as unreliable as the weather (with which in fact it is not entirely unconnected). This may have been one of the reasons why Ronalds experimented with an underground conductor: a copper wire was fitted in a glass tube surrounded with wax and pitch in a wooden trough, and he actually laid a length of such a "cable" over a distance of nearly 200 yards in a trench dug in the garden of his house. (Part of it has been found in recent years and can be seen in the Science Museum in London.)

Ronalds made strenuous efforts to interest the public and the Admiralty in his scheme. When the question was raised how to deal with those who commit wilful damage to telegraph wires, Ronalds retorted: " . . hang them if you can catch them, damn them if you cannot, and mend it immediately in both cases". The Admiralty did not even raise this question but simply replied (the year was 1816, when wars had ceased) "that telegraphs of any kind are now wholely unnecessary and that none other than the one now in use will be adopted." The one in use was of course the visual telegraph.

The electrochemical telegraph

Static electricity had not proved a very suitable means for telegraphy — its very name seeming to proclaim its unsuitability for propagation and communication. But new discoveries were made and new vistas opened for inventors. Galvani in 1791 had discovered the muscular contraction which occurs when certain metals are brought into contact with animal tissues, and this had led Volta in 1800 to the invention of his "pile", which constituted the first electric battery. With this work a new era began: electricity was now available in a low-pressure form as easy to control as static electricity had been difficult. Insulation of the connecting wires was no longer a major problem.

The use of Volta's pile in the same year led to the discovery by Nicholson and Carlisle that water could be decomposed into hydrogen and oxygen by the passage of an electric current. (The same effect had been obtained by Deiman and Paets van Troostwijk at Amsterdam in 1789 by the use of a frictional machine but the volume of the gases generated had presumably been very small, so that little attention was paid to their observation.) This discovery provided a means of *detecting* the current — it should be realized that hardly any other means was known at the time.

Knowledge of Volta's pile and of its use for decomposing water came in 1804 to the notice of Don Francisco Salvá of Barcelona, who had some years previously taken a prominent part in advocating schemes for an electric telegraph based on static electricity. Salvá realized that Volta's pile "yields more fluid than the electric machine, and could be well applied to telegraphy, as the force can be obtained more simply and more steadily than in the static form". In a paper read to the Academy of Science in Barcelona in 1804, Salvá gave a full description of an electrochemical telegraph in which bubbles of hydrogen and oxygen were used for indicating purposes. Unfortunately, his paper was not published at the time and in 1809 the same scheme was re-invented by Soemmerring, a member of the Academy of Science at Munich, to whom the credit for this invention is more commonly accorded.

Soemmerring had been invited by the Bavarian government to devote his attention to the matter of a telegraph, the Bavarians having been greatly impressed by the role the Chappe telegraph had played in their speedy relief from the Austrians by Napoleon in April 1809. Ten or fifteen years earlier this would probably have incited the adoption of a similar system of visual telegraphy — as we have seen in England and other countries, but apparently in 1809 the electrical phenomena had already caught the imagination to such an extent that visual telegraphy, despite its effectiveness, must have appeared old-fashioned to scientists. Soemmerring immediately conceived and in a very few days completed a telegraph using a number of conducting wires, each of which terminated in a pin which projected through the base of a glass vessel filled with acidulated water. By connecting any pair of the wires to the extremities of a Volta pile, he was able to cause bubbles of gas to rise from the appropriate pins. Soemmerring found that he could send reliable signals over a distance of several hundred yards and a few weeks later he exhibited a more elegant instrument to the Academy of Science in which he made use of a glass trough containing 35 gold pins, each with a connecting wire, to correspond with 25 letters of the alphabet and the 10 numerals (*fig. 7*). Later he reduced the number of wires to 27 and he added an ingenious contrivance for giving an alarm to call the attention of the operator. This arrangement was demonstrated in action over a distance of nearly two miles.

Although his demonstrations, given in 1809-1812, were quite successful, Soemmerring's telegraph was never put to practical use. With its large number of connecting wires it resembled C.M.'s proposal of 1753 and it would have been utterly uneconomic. Soemmerring's experiments nevertheless were of importance because their published description encouraged a number of others to devote serious thoughts to the development of electric telegraphs of various forms. Friends of Soemmerring's played an important role in the creation of the electromagnetic telegraph, as will be seen in the next chapter, but systems of electrochemical telegraphy continued to be proposed long after the electromagnetic telegraph had appeared on the scene. In 1828 Dyar in the United States carried out experiments based on sparks passing through a strip of chemically treated paper and producing discolorations forming a dot and dash pattern. Dyar was years out of date in employing *static* electricity for his experiments but his dot-and-dash code was a precursor to the Vail system to be discussed in a later section. A practical electrochemical telegraph using a similar code was devised by Bain as late as 1846 and for some time found considerable use in America.

Electromagnetic telegraphs [3]

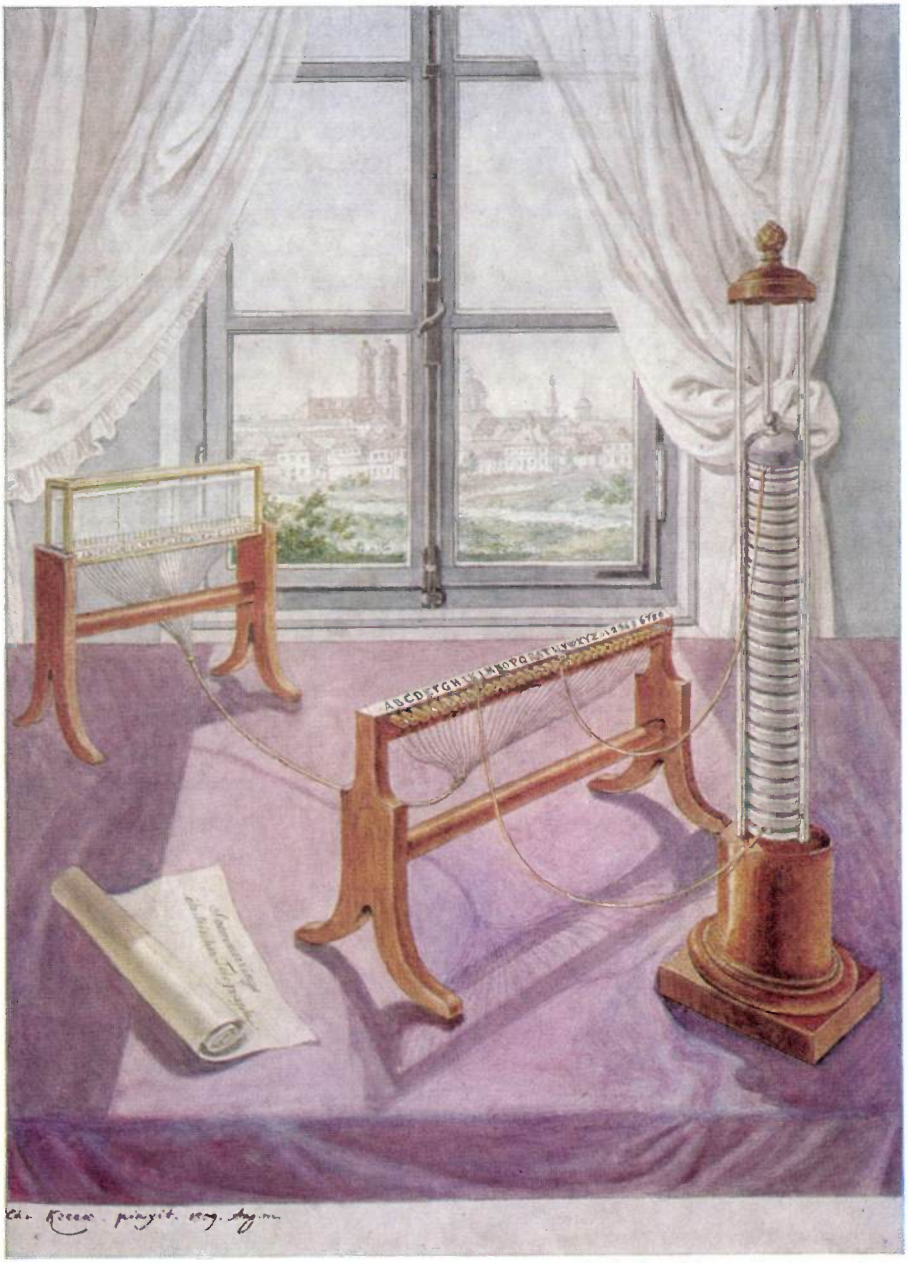
For a great many years there had been persistent speculation as to the possible connection between electricity and magnetism but it fell to Oersted at Copenhagen in 1820 to make the essential observation that a magnetic needle was deflected by a current in a nearby wire. Within a very few weeks, his experiments were being repeated by scientists in many countries, and even in the same year Ampère proposed to base a telegraph on Oersted's discovery. From the point of view of coding, Ampère's proposal was similar to C.M.'s proposal of 1753 and to Soemmerring's experiments: Ampère proposed to move a separate needle and therefore to use a separate wire for each letter.

Very soon others took up the idea in different forms. Baron Schilling, a member of the staff of the Russian Embassy at Munich, who had become acquainted with Soemmerring and who had tried to interest the Russian government in Soemmerring's electrochemical telegraph, was informed by Soemmerring of Oersted's discovery and also of the important development due to Schweigger. The latter had found that the deflection of the magnetic needle produced by the current in a wire could be multiplied by arranging the wire in a coil around the needle. Schilling saw the possibility of employing Schweigger's "multiplier" as the indicating element for an electric telegraph and from 1822 onwards carried out many experiments. A contemporary account describes one of his telegraphs as consisting of five separate "multipliers" (or galvanometers, as they were called later). To the suspensions of the needles were fixed small paper discs painted black on one side and white on the other. The system employed six conducting wires, each letter or numeral being signalled by a particular combination of black and white discs simultaneously visible to the receiving operator. With this system Schilling was the first to apply the important principle of the binary code; each indicator could assume one of two positions (made visible by the white

[3] Much information on the early history of the electromagnetic telegraph can be obtained from near-contemporary books such as: H. Schellen, Der elektromagnetische Telegraph etc., Brunswick 1850, and L. Turnbull, The electromagnetic telegraph, Philadelphia 1853.

and the black side of the paper disc) and with five simultaneously operated indicators Schilling could thus achieve $2^5 = 32$ different combinations. Thus he could make do with five wires, plus an earth return, instead of thirty-two.

Schilling was unsuccessful in his attempts to interest others in the use of his telegraph and it was not until 1837 that he received a firm commission to set up a telegraph between St. Petersburg and Cronstadt. The task was not completed, since Schilling died in the same year, but his invention of the more sophisticated system of coding mentioned above which so considerably reduced the number of wires was instrumental in the further development of practical telegraphy. This idea in fact is found in a nearly identical form in the needle telegraphs of Cooke and Wheatstone which will be discussed in a later section and which were directly inspired by Schilling's apparatus.



Senckenberg Library, Frankfort-on-Main

Fig. 7. Soemmerring's electrochemical telegraph (original version, 1809), depicted by Soemmerring's assistant Chr. Koeck. The drawing shows the Frauenkirche and the Theatinerkirche of Munich in the background.

We have now reached a point in our description when the history of the telegraph can no longer be traced as a simple succession of ideas. From 1830/35 onwards different lines of development are tending to emerge simultaneously in several countries. We shall briefly sketch the development in Germany, England and America.

Development in Germany

In Germany, important work in the field of telegraphy was done by Gauss, the famous mathematician and physicist, and his colleague Weber. In 1832/33, Gauss was engaged upon a study of the earth's magnetic field at the University of Göttingen. He was already well acquainted with the idea of an electric telegraph as he had been friendly with Soemmerring for more than 20 years and he had been one of those who tarried in Munich in 1810 to see the Soemmerring telegraph in operation. Early in 1833, Gauss and Weber converted one of their instruments into a needle galvanometer — the "needle" of which was a bar magnet weighing about a pound — with the object of testing the validity of Ohm's work on circuits. They set up a double copper wire between the astronomical observatory where the work on the earth's magnetism was being carried out and the University laboratory, a distance of about 1½ miles. They soon found their wires useful for purposes other than the checking of Ohm's work. At first they were used for synchronizing the clocks in the two buildings but by Easter of 1833 the wires were part of a communication system for sending words and phrases.

A remarkable feature of Gauss' and Weber's communication system was the use of a code with *successive* binary units. In C.M.'s, Soemmerring's and Ampère's systems each letter was marked by a change in one of 25 or 35 different indicators at the receiver — each indicator requiring a separate communication channel. Schilling, by introducing the binary code, had reduced the required number of indicators to five, but in Gauss' and Weber's system use was made of a *single* indicator capable of two positions (viz, left or right deflection of a single needle) and one group of *successively* executed movements represented a letter or numeral. This required only a single channel for transmission. Apparently Gauss and Weber had adopted this idea because, as distinct from previous inventors, they happened to approach the problem of telegraphy the other way round: having installed a single "channel" for their measurements, the afterthought of using it for telegraphy forced them to find a code which could be used with such a system. It is true that economy of wire was obtained at the cost of signalling speed, since several swings of the needle were required for each letter. The large weight of the needle (this was very heavy because the instrument had been originally designed for quite a different purpose) also helped to slow down the process. This became even worse when in 1834 Gauss and Weber modified their apparatus by incorporating a 25 pound "needle", whose small deflections had to be observed through a telescope. Not surprisingly, the speed of signalling seldom exceeded seven letters per minute.

In 1836 Gauss and Weber added a second interesting feature to their telegraph: they applied the phenomenon of induction — discovered by Faraday in 1831 — to replace the chemical battery which they had used in their earlier experiments. Their inductor consisted of a coil of wire which could be moved up and down a pair of large bar magnets. Movement of the coil in one direction or the other produced a current which deflected the "needle" to right or left (*fig. 8*).

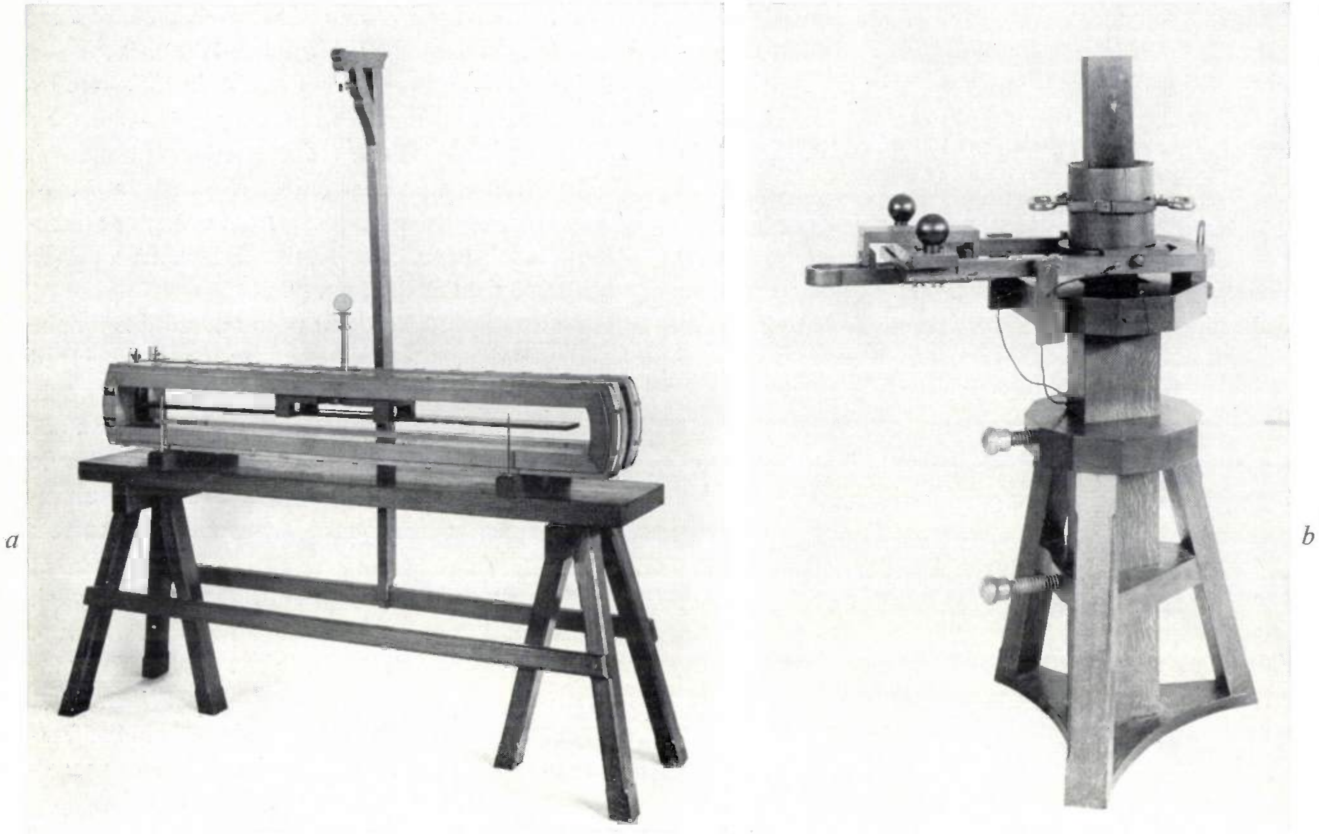
The years from 1830-1850 were remarkable for the considerable extent to which railways began to spread throughout most of the countries of western Europe. The first important line was opened in 1830 (to connect Manchester with Liverpool) and the use of an electric telegraph for giving information regarding the movement of trains at remote points seemed likely to prove a useful application.

Among the first to consider the use of a telegraph were the directors of the Leipzig-Dresden Railway who, in 1836, expressed an interest in the Gauss-Weber instruments. The low rate of working of these instrument, however, was a serious handicap and Gauss himself, realising the deficiencies, invited Steinheil (another member of the Munich Academy of Science) to develop a more simple and practical form of telegraph. Steinheil transformed the cumbersome Gauss and Weber receiver into a well proportioned instrument in which signals were received *acoustically* by means of two small bells of different pitch, struck by the right or left deflection of the needle carrying a little clapper. (From the letter in fig. 4 it can be seen that C.M. had already considered the use of bells of different pitch, viz, different for each of the 25 letters — rather an exacting requirement on the musical ear of the receiver!) As an alternative to this arrangement Steinheil adapted his instrument to record the signals on a moving paper tape by means of dots. The code in both cases was similar to Gauss' and Weber's.

Steinheil's receivers show a definite transition from mere "philosophical toys" to practical apparatus. The same is true of his transmitters which, though basically similar to Gauss' and Weber's, embodied principles of mechanical design that remained in vogue during many decades: pulses of current in one direction or the other were produced by turning a crank.

The first telegraph system established by Steinheil at Munich consisted of four stations up to 5 miles apart, which could be connected in any required combination by means of a simple switching device. The instruments worked very well at the rate of about 6 words per minute.

In 1838 a Steinheil telegraph was installed along the Nürnberg-Fürth railway for a distance of about five miles. Gauss had suggested that in railway applications it might be possible to do away with the two wires by



Deutsches Museum, Munich

Fig. 8. Gauss' and Weber's telegraph.
a) Receiver (replica of the second version, of 1834). In effect, it is a needle galvanometer, in which the "needle" consists of a 25 pound iron bar suspended by a wire from the vertical support.
b) Transmitter (replica of version of 1836). This was an inductor consisting of a large coil of wire which by means of a hinged lever could be moved up and down a pair of vertical bar magnets. The operator had to grip the two knobs and to move his end of the lever up and down.

using the railroad tracks as conductors. Steinheil soon found that the insulation between the two tracks was insufficient, but he decided to use the tracks at least as a substitute for *one* of the wires. He then discovered that his installation continued to work even when the track was interrupted and concluded (as others had similarly established) that the earth itself acted as a conductor. He thus introduced the use of the "ground return", cutting down the cost of his own and all future telegraph lines.

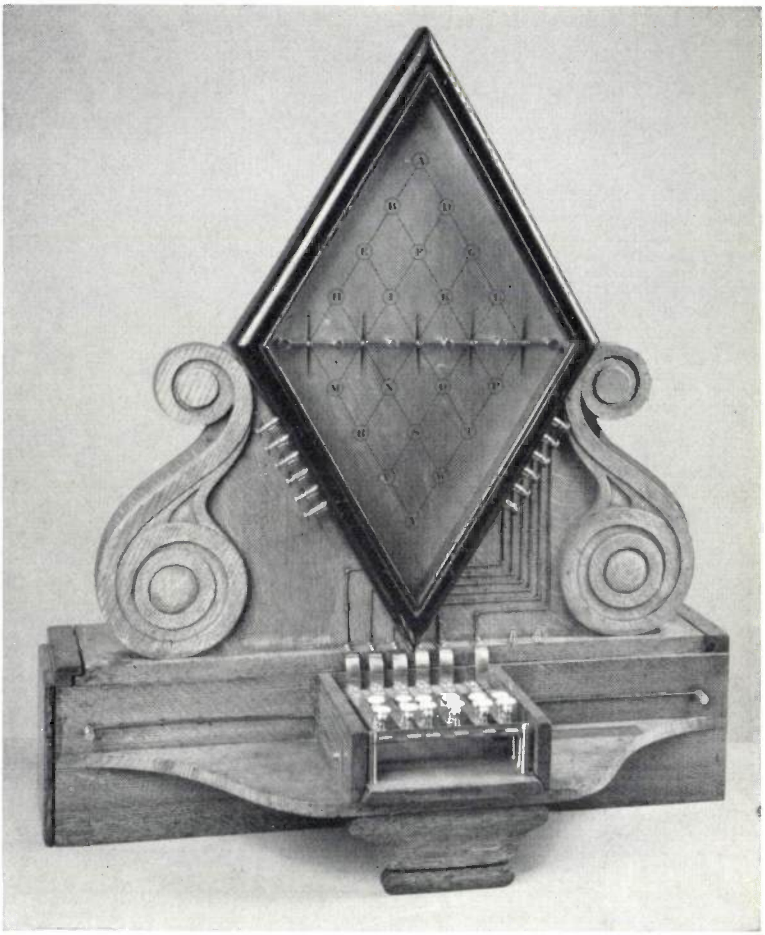
Installation of his telegraph was considered also for the Munich-Augsburg railway line but, despite the economy achieved, officials decided that the expense would not be justified. The proposal was dropped and development of the telegraph in Germany retarded for a number of years. It was later revived by the reports originating from England and America where the advantages of electrical communication were more quickly appreciated.

Development in England

Development of the electric telegraph in England and its practical application was almost entirely due to the efforts of W. F. Cooke and Charles Wheatstone. When visiting Heidelberg early in 1836, Cooke had witnessed a demonstration of one of Schilling's instruments and he became deeply impressed with the commercial possibilities of an electric telegraph. Returning to England, he tried without much success to construct several forms of telegraph and soon felt that he should seek scientific guidance. He consulted Faraday, who referred him to Wheatstone, professor of natural philosophy at King's College, London. Wheatstone himself had been conducting experiments with a form of electric telegraph and in view of their mutual interests in the project they decided to enter into partnership.

The first successful telegraph developed by Cooke and Wheatstone contained five vertical needles pivoted on horizontal axes and arranged across a diamond-shaped dial marked with the letters of the alphabet (*fig. 9*). Each of the five needles carried on its axis a small magnet placed in a coil of wire behind the dial, and keys in front of the instrument were so connected that each needle could be deflected at will to right or left, the corresponding needle in the distant receiver thereby undergoing the same deflection. The signalling of any given letter was achieved by the deflection of two of the needles in opposite directions, the intersection of their directions indicating the required letter. This was an ingenious combination of the binary code, which used only two motions (left or right), and the multi-place code as proposed by C.M., Soemmerring and Ampère using separate indicators for each letter: it required five line wires (plus an earth-return) as had Schilling's apparatus, and moreover it afforded quick and direct readability of the letters signalled so that its use required but little skill and anyone could learn to work it in a few minutes.

Two of these instruments were made during August 1837 and demonstrated to the directors of the London-Birmingham Railway on a line of about 1½ miles in length. Although the demonstrations were very successful, some of the directors remained unconvinced of the need for such communication and the installation was dismantled. In the following year, however, Cooke succeeded in persuading the directors of another railway, the Great Western, to adopt the telegraph and in July 1839 a line was completed between Paddington Station in London and West Drayton, a distance of about 13 miles. The same pair of 5-needle instruments was employed on this line and performed so well that the line was extended to Slough in 1843 (One of these original instruments, shown in fig. 9, is now in the possession of the Science Museum; the whereabouts of the second instrument had remained unknown until early in 1964 when it was discovered in the possession of the Postmuseum in East Berlin.)



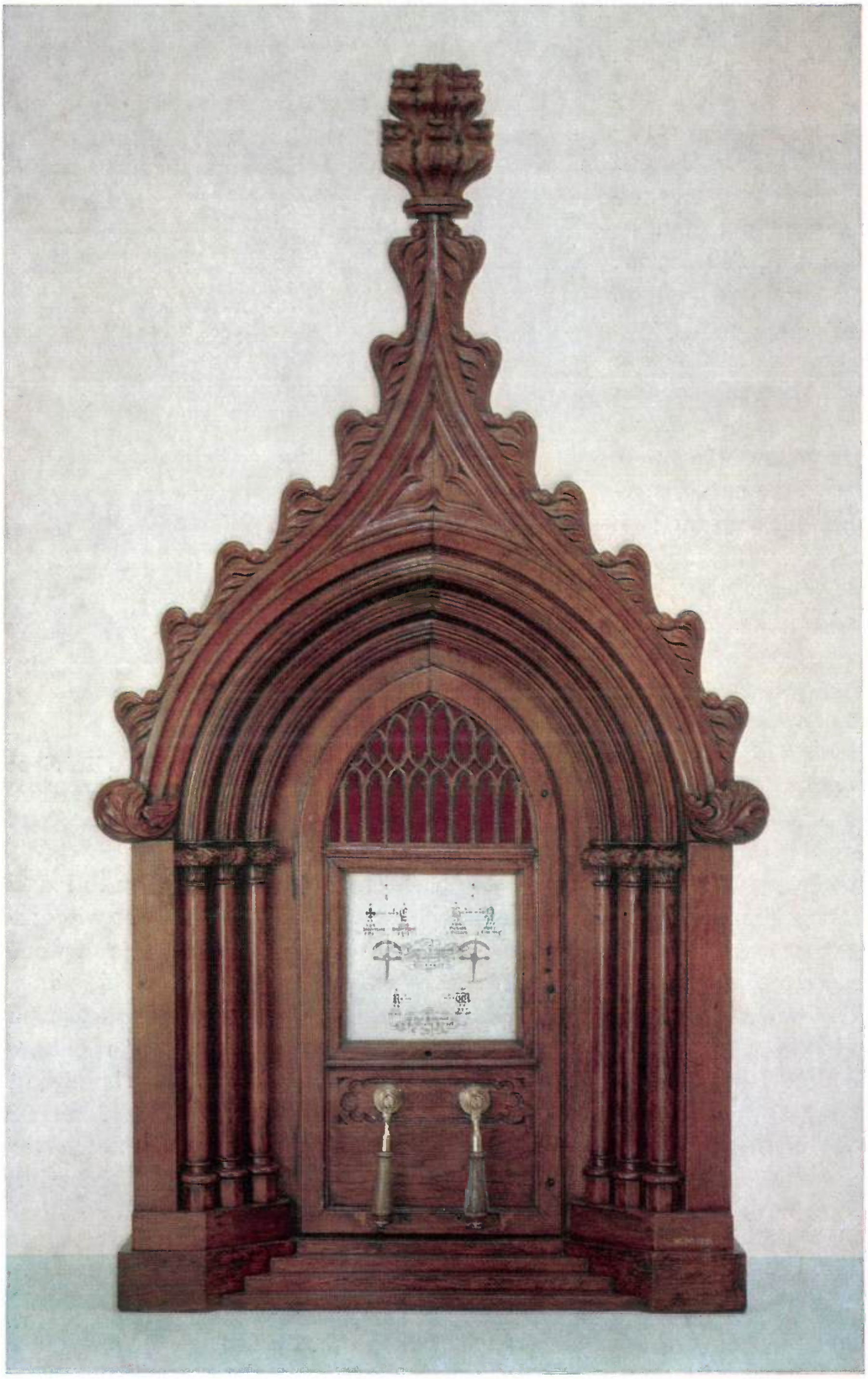
Science Museum, London

Fig. 9. Cooke and Wheatstone's 5-needle telegraph. This is one of the original pair of instruments made in 1837 for demonstration and subsequently used from 1839 to 1843 on the Great Western Railway.

We have described the 5-needle instrument as a combination of economy of wire on one hand and simplicity of code on the other. With experience of telegraph working it began to be realised that the simplicity of read-out was not really essential and that a higher rate of signalling could be achieved by experienced operators using a seemingly more complicated code and a system with only one or two indicators. Following this line of thought, Cooke in 1843 withdrew the 5-needle instruments on the Great Western Railway and replaced them by a design using only *two* needles. Each letter was signalled by a sequence of momentary deflections to left or right of one or both the needles in a manner very similar to that employed earlier by Gauss and Weber. The code, of course, necessitated some training of the operators but the system was soon shown to be far superior: in addition to reducing the capital outlay (by reason of the fewer wires and more simple instruments), it enabled a rate of working as high as 22 words per minute to be achieved.

The success of the 2-needle telegraph on the Great Western Railway encouraged rapid progress in other parts of England. Cooke had already in 1840 been commissioned to install a telegraph on the Blackwall Railway; in 1843 a line was laid between Norwich and Yarmouth, London was connected with Gosport and Southampton in 1845 and in 1846 a line was completed between London and Dover.

The electric telegraph now began to attract the interest of the public and in 1845 an event occurred which greatly contributed to a popular understanding of the service which the telegraph could render to the community. A woman was murdered in her cottage in the outskirts of Slough and her murderer was seen to board a train at Slough Station en route for Paddington. A message was immediately sent by the electric telegraph to Paddington where, when the train arrived, the man was recognized and later arrested. His trial attracted tremendous interest and when the accused man, John Tawell, was eventually convicted and hanged, it was widely said that "John Tawell had been hanged by the electric telegraph".

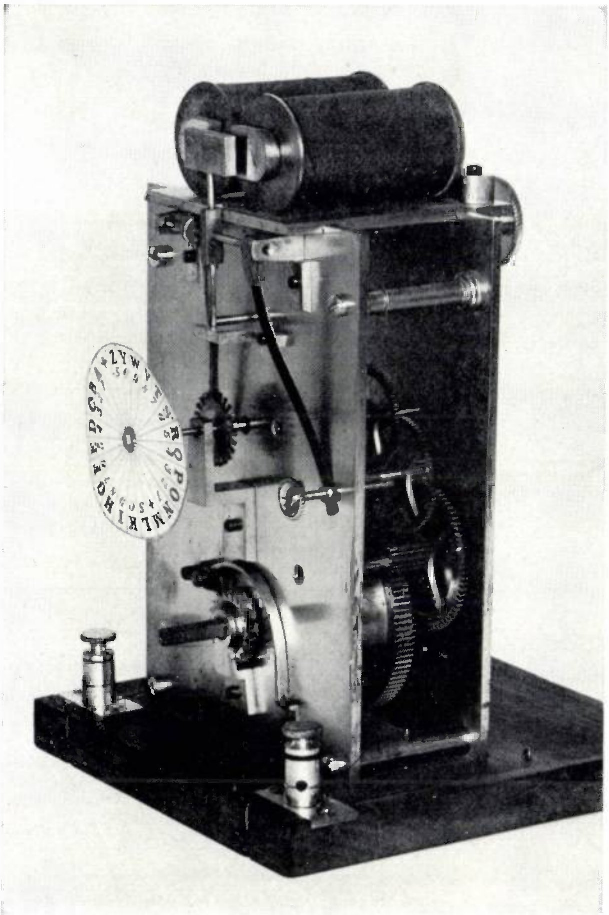


Science Museum, London

Fig. 10. Two-needle telegraph, carrying Gothic letters and mounted in a richly decorated cabinet, which was installed in the Octagon Hall of the House of Commons in 1846.

Less spectacular but perhaps even more important was the application which the telegraph soon found in political life. A 2-needle telegraph, whose appearance was adapted to conform with the surrounding architecture, was installed in the Octagon Hall of the House of Commons by the Electric Telegraph Company in 1846 and was used for the sending and receiving of messages relating to parliamentary business (*fig. 10*).

With the electric telegraph thus firmly established in England, we ought to leave this scene, but one interest-

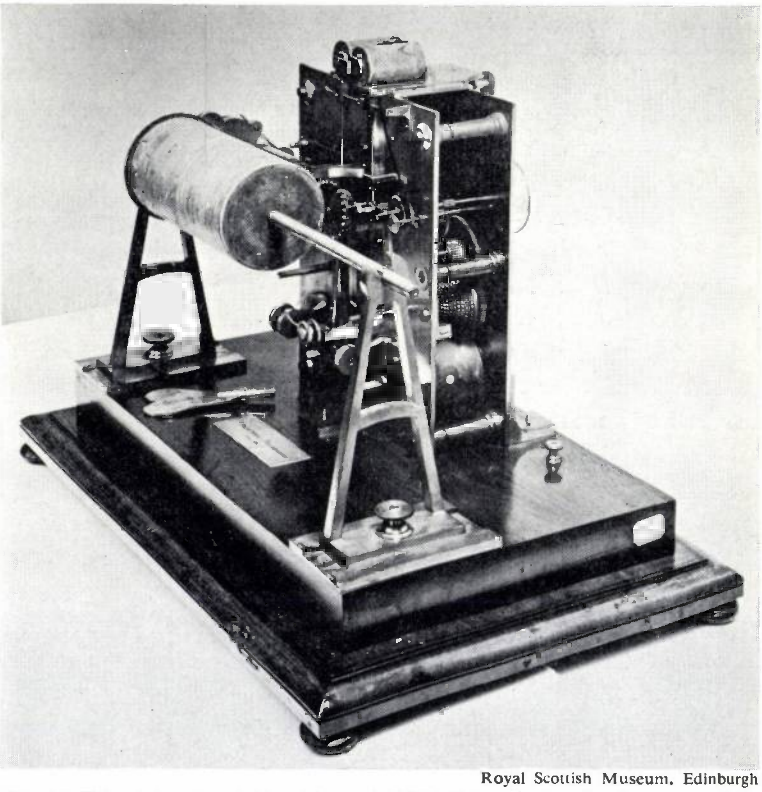


Science Museum, London

Fig. 11. Wheatstone's "A.B.C." telegraph, of 1840, based on his electromagnetic escapement. An electromagnet energized by pulses of current from the sender causes a step-by-step rotation of the clockwork-driven paper disc bearing letters and numerals.

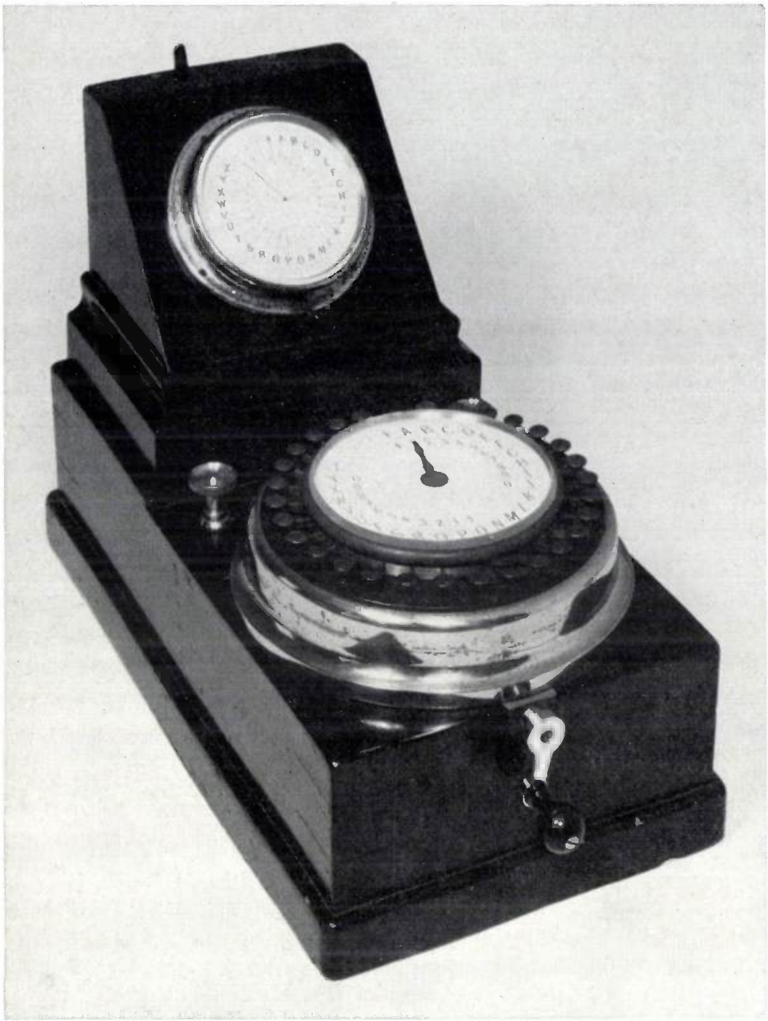
ing off-shoot of the development should not go unmentioned. Whilst Cooke had realised at an early stage that easy interpretation of the signals was of lesser importance than economy and not essential for a high signalling speed, his partner Wheatstone seems to have remained prejudiced in favour of the letter-indicating form of telegraph. In 1840 he devised what he called an "A.B.C." telegraph (*fig. 11*). It required a single wire line, along which pulses of current were sent to energize an electromagnet in the receiver. This electromagnet controlled a clockwork escapement whereby a paper disc bearing around its periphery the letters of the alphabet was made to rotate, step by step, behind a small window, a given letter being signalled by allowing the disc to pause for a moment while the required letter appeared in the window. Except for the fact that the time scale was arbitrary and irregular, this system was essentially similar to the "time-coding" applied by Chappe and by Ronalds (page 271), with which it had in common the disadvantage of what computer people would now call a long "access time". In the following year, in 1841, Wheatstone went even further and transformed the instrument into a printing telegraph! The "letter disc" in this instrument was replaced by a type wheel, i.e. a circular arrangement of flexible reeds each carrying a type-face, and the window was replaced by a hammer behind this wheel (*fig. 12*). Having reached the correct position for a given letter, a separate signal was used to release the hammer to strike the type face against a sheet of paper placed on a drum, the latter making a sliding and rotatory movement as in our modern type-writers. Although the apparatus was highly ingenious, there was no demand for a printing telegraph at this date. Only two instruments were ever built and they remained forgotten for more than a hundred years in a store-room at King's College, London, until recognised by the present writer a few years ago.

Many years later — in 1858 — Wheatstone patented a completely new design of A.B.C. telegraph. These instruments, (see *fig. 13*), though intricate, were beauti-



Royal Scottish Museum, Edinburgh

Fig. 12. Wheatstone's printing telegraph, 1841. It is a further development of the step-by-step telegraph shown in fig. 11. (An even better engineered specimen, which is in working order but which is less suited for showing the mechanism, is in the Science Museum, London.)



Science Museum, London

Fig. 13. Wheatstone's "A.B.C." telegraph, 1858. Like that shown in fig. 11, it also employed a step-by-step mechanism, but one of greatly improved design. While transmitting, the operator turned the crank continuously, thereby generating a continuous series of current pulses. To transmit any desired letter, the operator pressed the corresponding button and the current pulses then caused the pointer — in his own instrument as well as in the distant receiver — to advance in a clockwise direction, coming to rest opposite the selected letter. This form of instrument was still in use in remote country districts of the British Isles as late as 1920.

fully constructed; they were very reliable and they remained popular for many years on private telegraph lines and on public circuits where the volume of traffic did not warrant the employment of highly trained operators. A few were still in use as recently as 1920!

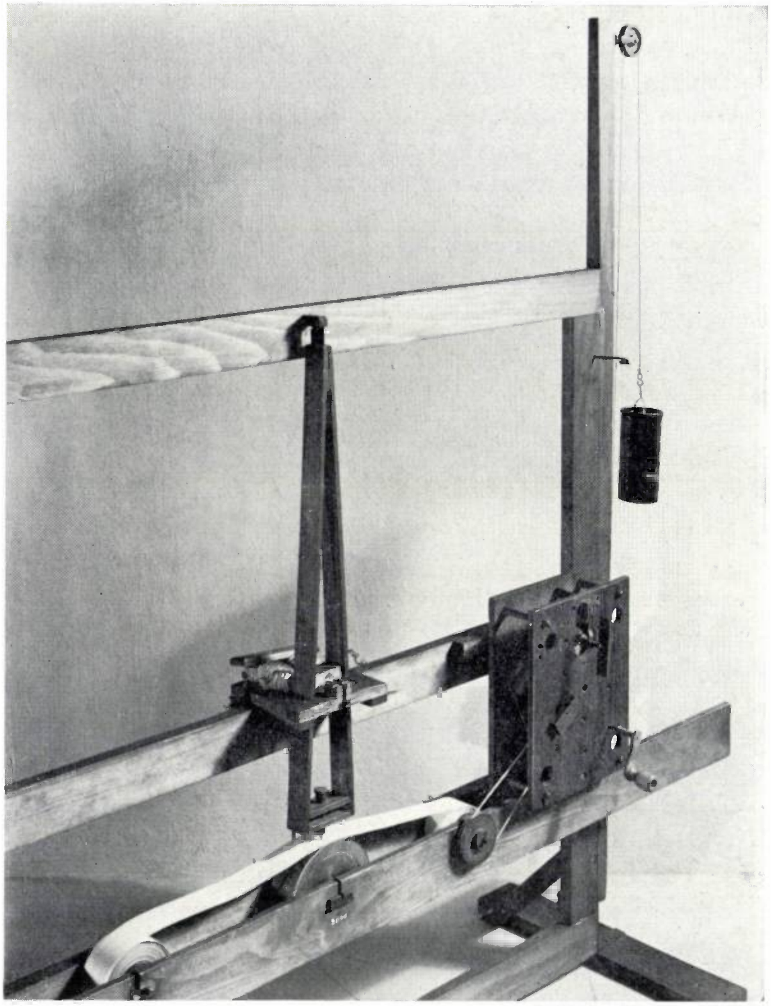
Development in America

The development and introduction of the electric telegraph in America is usually ascribed to one man, Samuel Morse, but as will be seen in the following paragraphs, this is an over-simplification which is not strictly justified.

Morse, who was an artist by training and profession, first conceived his idea of an electric telegraph while returning from a voyage to Europe in 1832 when he was nearly 50 years of age. A replica of his first crude instrument which he constructed in 1835 and for whose frame he is said to have used part of an artist's easel is illustrated in *fig. 14*. A wooden pendulum carrying an iron bar and a writing pen was suspended from the framework by means of a spring and was intended to be moved to and fro by an electromagnet energized by current pulses from a battery, the pen thereby recording a zig-zag line on a paper strip drawn continuously over a roller by a weight-driven mechanism. The instrument did not work well, Morse having been severely handicapped by his inadequate understanding of electromagnetism and his complete lack of mechanical skill. It exhibited one significant difference, however, between the working of a telegraph as conceived in America and in England. In England it was accepted that a telegraph receiver should offer a visual presentation of the received signal, relying on the *simultaneous* interpretation and recording by a skilled operator, and this principle was adhered to even for some time after Wheatstone had demonstrated the feasibility of a printing telegraph. In America, on the contrary, it was from the outset thought necessary for the receiving instrument to make a record of the signals on paper tape for subsequent transcription by semi-skilled operators. The limitation of signalling speed due to the rate of read-out was avoided by this separation of functions (cf. also page 277).

Towards the end of 1835 Morse became acquainted with Leonard Gale, professor of chemistry at New York University. Assisted by Gale and Joseph Henry, to whom Gale had introduced him and who contributed to a clarification of his concepts of electricity, Morse made a little progress but it was his meeting with a young man, Alfred Vail, in 1837 that led to ultimate success. Vail, who was a well-educated man and a competent mechanic, completely redesigned the telegraph and, by the early part of 1838, it was capable of signalling through a line of about three miles. Vail's instrument recorded dots and dashes, a principle which was adhered to in all later models (*fig. 15*).

The system of coding was of course again an essential detail of such a system. Morse had started work on a code-book dictionary in using which he conceived that messages would be conveyed by the transmis-

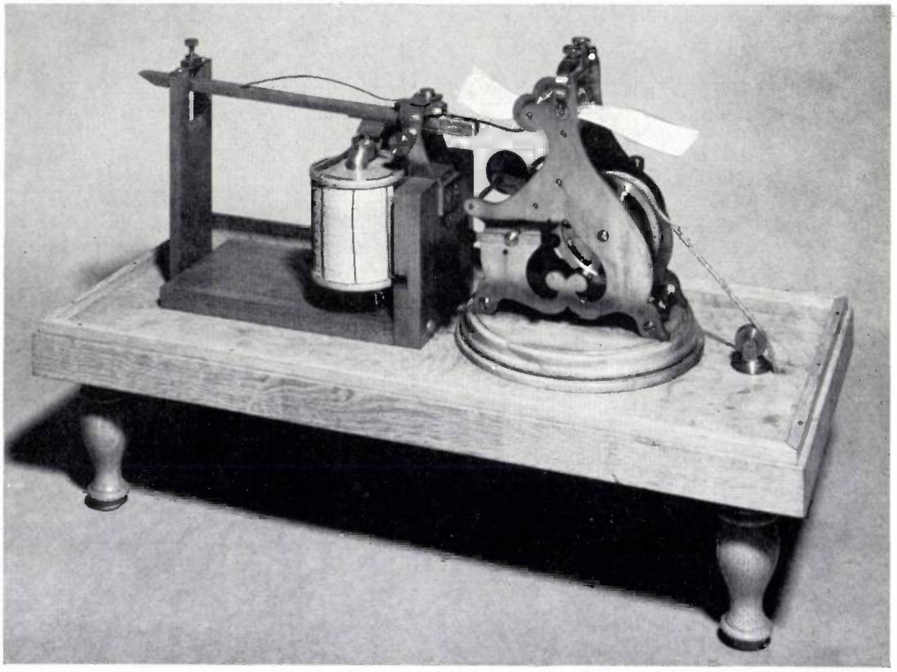


Deutsches Museum, Munich

Fig. 14. Morse's electromagnetic telegraph. First model of the receiver, 1835 (replica).

sion of groups of numerals — a very ancient idea which had been propounded and elaborated many times, for example with Chappe's optical telegraph. Vail regarded such a procedure as too restrictive and cumbersome; he preferred the use of a code in which a succession of symbols represented particular letters and thus followed a path trodden by Gauss, Steinheil and Cooke. Vail, however, in elaborating his code did better. He visited a local printer to ascertain the relative frequency with which each letter was used in the English language and he determined on the use of the shortest symbols for the most commonly used letters — a single dot for an "E", a single dash for a "T", and so on. Thus the Morse code was born: an economical code but one which should in fact more correctly be termed the "Vail code".

The widespread acceptance of the name of Morse as the inventor of the electromagnetic recording telegraph and of the famous code almost certainly stems from an agreement made by Morse, Gale and Vail that all inventions and developments in their work on the telegraph should be ascribed to Morse. There were undoubtedly good commercial reasons for such a course but it happened to accord all too well with what must have been an overweening desire on the part of Morse to receive public acclaim. The same unfortunate trait evidently prevented Morse from giving due credit to the important role that Henry's advice had played in the ultimate success of their telegraph system. The problem of operating an electromagnet by a battery through very long wires was by no means easy to solve, as the concept of impedance matching was as yet unknown. Henry had demonstrated in 1830 that in such a



Science Museum, London

Fig. 15. Morse telegraph, as redesigned by Vail. Model of 1845 (replica). This form was used on all the early Morse telegraph lines in America.

situation it was useless to employ a "quantity battery" (one with large plates but few cells) together with a "quantity magnet" (one with only a few turns of thick wire), but that an "intensity battery" combined with an "intensity magnet" could do the job. By such considerations Henry could advise Morse and Vail on the proper dimensioning of their coils and batteries in order to increase the distance over which their telegraph signals could be received. Henry also invented the *relay*, which was soon adopted by Morse and which when inserted at suitable intervals enabled the telegraph signals to be transmitted over much larger distances than had been attempted previously.

Unlike the situation in England where the early railways had played such an important part in the establishment of a telegraph network, the telegraph in America was dependent on political patronage and subsidy before any promise of commercial support could be obtained. During the five years between 1838 and 1843 Morse made repeated attempts to procure government assistance to establish the telegraph but it was not until March 1843 that Congress was persuaded to allocate 30 000 dollars for the purpose. These funds were employed to establish a line between Washington and Baltimore, a distance of about forty miles, the first message being transmitted on 24th May 1844 at a speed of about six words per minute. The line was opened for public business on the 1st April 1845, as it happened the very day upon which John Tawell, the murderer of Sarah Hart, was executed iu England the first criminal to suffer retribution for his crime as a direct consequence of the use of the electric telegraph in a public service.

The value of the telegraph having been demonstrated, the earlier apathy was displaced by unbridled enthusiasm and Morse, who had been hitherto regarded by many as a tiresome crank, was now praised on every side. A political incident, quite trivial in itself but in which the telegraph played an important role, enhanced his fame and a line connecting New York with Philadelphia was quickly projected. In the course of the next ten years the States bordering on the Atlantic coast were covered with a vast network of telegraph lines under the control of nearly forty different companies which were amalgamated in 1856 to form the organization which is known to this day as "Western Union".

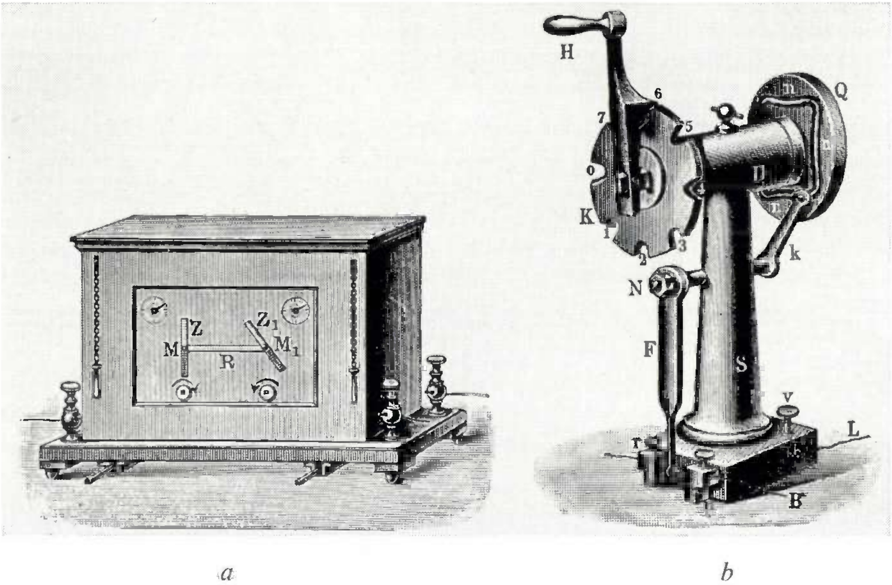


Fig. 16. Electromagnetic telegraph designed by Bréguet for the French government in 1845. It was based on the use of two electromagnetic escapements (cf. fig. 11), each capable of 8 positions. In the receiver (*a*) the two pointers rotated by these escapements simulated the indicator of a Chappe (optical) telegraph (cf. fig. 1) and the Chappe code was to be used. The sender contained two crank units, one of which is shown in (*b*), and these also were positioned to form a Chappe "répétiteur". (From Th. Karrass, Geschichte der Telegraphie, Brunswick 1909.)

Further developments

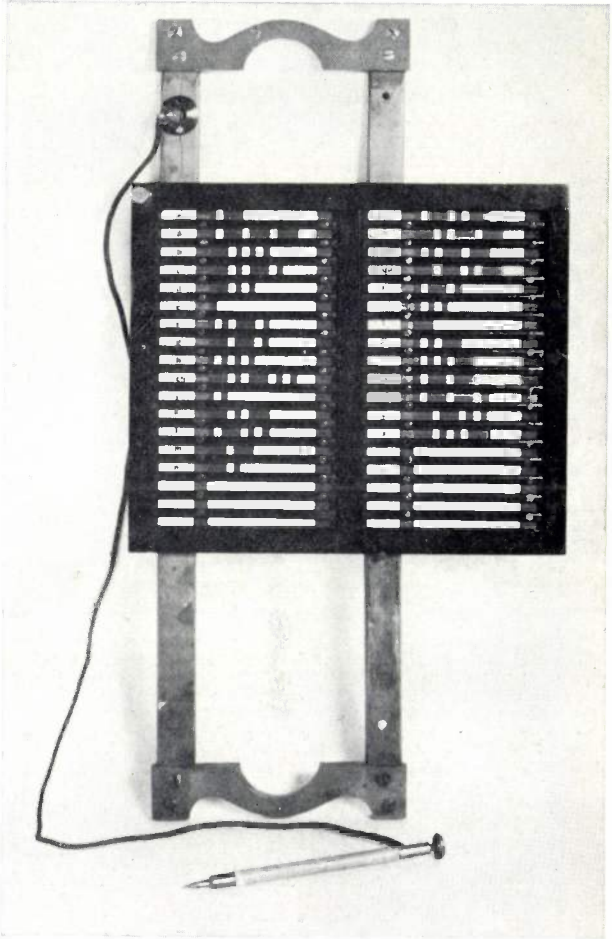
By the middle of the century, America had taken the lead in utilizing the electric telegraph as a modern means of communication. According to a review of national telegraph facilities made in 1853, there were then more than 24 000 miles of telegraph lines in the States, about 2500 miles in England, 3000 miles in Germany (where a vigorous development of the telegraph had been resumed from about 1848 onwards [4]) and 750 miles in France.

The growth of an electric telegraph network in France had been delayed by the existence of the large and well-organized network of optical telegraph lines of the Chappe type which had seemed to satisfy the current needs and which had required so much investment and so much training of its operators that it was psychologically difficult to disband it. When Foy, the director of this Chappe system, finally considered the establishing of an experimental electric telegraph line in France in 1843, he insisted that the apparatus should be capable of giving the same visual indications as used in the Chappe system, so that there would be no need for the operators to be trained in a new code! This was achieved by Bréguet, who in 1845 completed his apparatus based on a combination of two electromagnetic escapement systems similar to Wheatstone's step-by-step telegraph and whose receiver indeed showed a configuration identical with the Chappe "répétiteur" (*fig. 16*). In view of this strange requirement, it is not surprising that the Bréguet instrument, despite its ingenuity, necessitated some sacrifice of efficiency and economy.

In the Netherlands the first electromagnetic telegraph was installed in 1845 for the Amsterdam-Haarlem railway line, the system being based on a needle telegraph similar to Cooke and Wheatstone's. In 1851 the Morse telegraph was introduced on one line and gradually displaced the other types. A similar trend is seen in other countries, when with increasing traffic the importance of having a record of each signalled message became evident. For the same reason, regulation by law became necessary and a law on public telegraphy ("Wet tot regeling der gemeenschap door elektromagnetische Telegraphen") was passed in the Netherlands as early as 1852.

From the mid-century onwards, the practice of telegraphy expanded in many directions. Telegraph lines were erected over ever-increasing distances and, in 1850 a significant milestone was reached by the laying of a cable under the sea to link England with the continent of Europe. New instruments were invented and speeds increased by the early steps towards mechanization (see e.g. *fig. 17*) and by the introduction of automatic systems employing *punched tape*, an invention for which Charles Wheatstone was mainly responsible in 1858 and which, today, forms such an important element in both communication and computer systems.

Incidentally, Wheatstone was also among the first to apply the principles of electric telegraphy for *telemetering* purposes: In 1843 he designed a "telegraph thermometer" which when carried aloft by a balloon connected to a ground station by a double copper wire instantaneously indicated at this station the temperature of the upper air. Another interesting sideline was the use of telegraph systems for the transmission of *time signals*. This possibility had already been recognized by Gauss and Weber (page 275) and by about 1850 electric clocks constituted an important branch of the art of electric telegraphy, as witnessed by a book published in that year (see *fig. 18*).



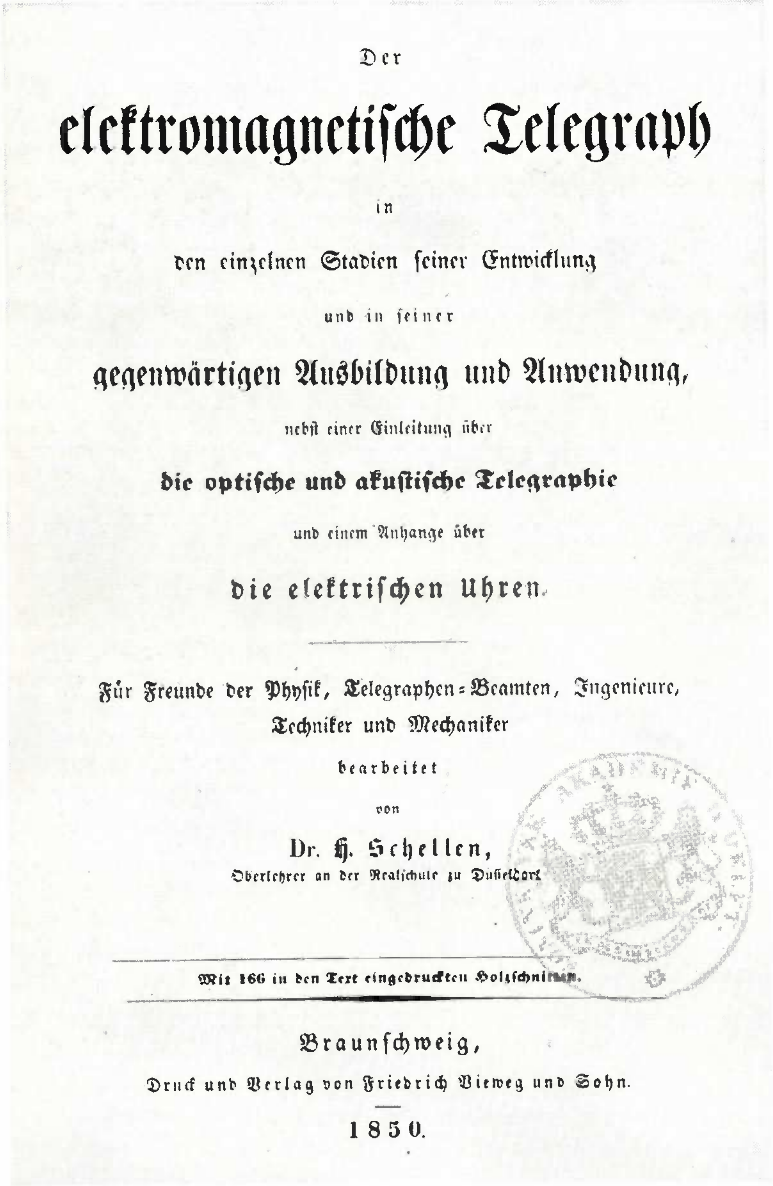
Post and Telegraph Museum, Vienna

Fig. 17. Writing tablet for mechanically producing the current pulses of the Morse code (1856): for each letter to be signalled the contact pin seen below was drawn over the corresponding row of contacts.

A final development which dates from the mid-century years is that of the *printing telegraph* of which Wheatstone had been the pioneer in 1841. In the hands of House, Brett, Hughes, Baudot and many others this

[4] An account with special reference to the use of the telegraph for railway systems was given in: W. Fardely, Der elektrische Telegraph, mit besonderer Berücksichtigung seiner practischen Anwendung für den gefahrlosen und zweckgemässen Betrieb der Eisenbahnen, etc., Mannheim, about 1855.

development resulted in the teleprinter of today, an achievement which has made possible the International Telex System now so widely used by industry and commerce.



Der

elektromagnetiſche Telegraph

in

den einzelnen Stadien ſeiner Entwicklung

und in ſeiner

gegenwärtigen Ausbildung und Anwendung,

nebſt einer Einleitung über

die optiſche und akuſtiſche Telegraphie

und einem Anhange über

die elektriſchen Uhren.

Für Freunde der Phyſik, Telegraphen=Beamten, Ingenieure,
Techniker und Mechaniker

bearbeitet

von

Dr. H. Schellen,
Oberlehrer an der Realſchule zu Duſſeldorf

Mit 166 in den Text eingedruckten Holzſchnitten.

Braunſchweig,

Druck und Verlag von Friedrich Vieweg und Sohn.

1850.

Fig. 18. Title page of Schellen's book on the electromagnetic telegraph, published in 1850 and containing an appendix dealing with electric clocks.

Recommended literature

In addition to the books and articles mentioned in footnotes [1]-[4] and in the subscripts to fig. 3 and fig. 16, the following may be mentioned:

J. Priestley, The history and present state of electricity, etc., 1775.
R. Sabine, History and progress of the electric telegraph, 1869.
J. J. Fahie, A history of electric telegraphy to the year 1837, London 1884.
R. Hennig, Die älteste Entwicklung der Telegraphie und Telephonie, Leipzig 1908.
E. Feyerabend, Der Telegraph von Gauss und Weber, Berlin 1933.
W. J. King, The development of electrical technology in the 19th century, 2. The telegraph and the telephone, Bull. 228 Smithsonian Institution, Washington D.C. 1962.

Summary. The first proposal for an electric telegraph, based on static electricity, dates from 1753; experiments with several systems of this kind were carried out in the late 18th and early 19th centuries. In the meantime the pressing need for rapid communication in times of war had favoured the establishing in France and other countries of a large network of optical telegraphs (Chappe 1792). After some attempts at telegraphy based on electrochemical phenomena, rapid development started when the fundamental discoveries of electromagnetism were made (Oersted 1820, Faraday 1831). A brief survey is given of developments in Germany (Gauss and Weber, Steinheil), England (Cooke and Wheatstone) and America (Morse, Vail and Henry). Coding was a central problem in all telegraph systems proposed, as a compromise was sought between economy of capital investment, signalling speed and ease of operation. In the description of different systems special attention is paid to this aspect.

Volume 26, 1965, No. 8/9 pages 205-284 Published 21st April 1966

PHILIPS TECHNICAL REVIEW

VOLUME 26, 1965, No. 10

In memoriam Dr. E. Oosterhuis

On 26th January, 1966, Dr. Ekko Oosterhuis passed away at the age of seventy-nine. The special place he occupied in the Philips Research Laboratories in Eindhoven makes appropriate a review of his life and work.

Dr. Oosterhuis read natural sciences at the University of Groningen. After receiving his doctorate, he spent a few years as assistant to Zeeman in Amsterdam and to Kamerlingh Onnes at Leiden. Early in 1914 he was invited to come and work as a physicist at the Philips Research Laboratories, which were then only a few months old. Here the writer was able to work in the closest cooperation with him for nearly twenty years. Dr. Oosterhuis was at the Laboratories for about thirty-five years, and, during this time, made very significant contributions to their development.

In 1914, Philips was still concerned only with the manufacture of incandescent lamps. The basis for manufacture was almost totally empirical. On the introduction of the gas-filled lamp, however, it appeared desirable to strengthen this basis, and much of the research required was carried out by Dr. Oosterhuis. The investigations included many related subjects, such as gas-discharges and the thermionic valve, and in the end they covered all possible types of discharge tube, both with and without incandescent cathodes. There was every reason for such an extensive research programme. Bohr's theory, then recently developed, and the experiments of Franck and Hertz, among others, made it possible to obtain some idea of the processes that took place inside these tubes. Many technical applications for them were devised.

For some time, the thermionic valve was the most important of these subjects. It was found that to make radio valves that were best fitted to their purpose it was necessary to have a thorough knowledge of the way they work in the various circuits. Dr. Oosterhuis paid a great deal of attention to the matching of valve and circuit. He became the leader of the group engaged on research into radio receivers and, later, television receivers. In this capacity he worked with a large number of other people, and in such work his human qualities were well in evidence.

Never content with the results obtained, always on the search for improvement and new developments, he made our products in this line equal to the best that could be obtained anywhere in the world. His sound judgement and ability to pass constructive criticism, his amiability and modesty, made him an ideal leader. He was able to communicate his reliability and strength of mind to many of his fellow-workers. His evenness of temperament helped to give him that rare quality of bringing people together into a close-knit team. Many of them will look back with gratitude and pleasure on the time they spent working with Dr. Oosterhuis. They all learned a great deal from him. Everyone who came into contact with him became in some way the better for it.

The results of his research are to a large extent embodied in the eighty or so patents that carry the name of Oosterhuis as inventor or co-inventor. With these, he made a very effective contribution towards making Philips' patent position in the world what it is.

Finally, a few words on Dr. Oosterhuis' contributions to various periodicals. After the first World War there was no suitable publication in the Netherlands for reporting the results of straightforward research. As there were basic objections to publishing in foreign periodicals only, Dr. Oosterhuis, together with one or two others, took the initiative and started the magazine "Physica", of which he also became an editor. Until 1938 he was an editor of the "Nederlands Tijdschrift voor Natuurkunde", an offshoot of "Physica" started in 1934.

In 1935, Dr. Oosterhuis was one of the founders of the Philips Technical Review and, from the very beginning, assisted in editing it. After the second World War, and until he retired in 1952, he was the editor-in-chief of this periodical. Even after his retirement, Dr. Oosterhuis placed his experience at the disposal of the "Tijdschrift van het Nederlands Elektronica- en Radiogenootschap"; he worked for this journal without wishing to be mentioned as an editor.

In all these ways Dr. Oosterhuis made a very considerable contribution to the advancement of physics and electronics in the Netherlands.

G. HOLST

A positive rod or piston seal for large pressure differences

J. A. Rietdijk, H. C. J. van Beukering, H. H. M. van der Aa and R. J. Meijer

621-762.649

The moving seal between the gas spaces represents one of the most difficult design problems in hot-gas engines and gas refrigerating machines based on the Stirling cycle. It now appears that this problem can be satisfactorily solved, through the discovery that a specially shaped rubber diaphragm can be rolled up and down more than 2×10^9 times before failing — provided the rolling is done in a suitable way. This discovery and the device based upon it may well also prove to be of great value in other types of machine.

From the beginning of the work by Philips on the Stirling cycle, much attention has been given to the investigation of moving seals. Stirling machines consist of a driving mechanism and a piston-fitted cylinder system in which a working agent goes through a continuous cycle of temperature and pressure changes; two basic design variations appear in *fig. 1.* At the moment, hydrogen and helium under high pressure are virtually the only practicable working agents. Working agent pressures may be as high as 140 atm in hot-gas engines, and as high as 60 atm in gas refrigerating machines. The mode of functioning of these machines has been described in earlier articles in this review [1] [2], and a brief account of the principle of the Philips-Stirling gas refrigerating machine may be found in the present issue [3].

In the design appearing in fig. 1*a* two moving seals are necessary to shut off the space containing the working agent from the space containing the driving mechanism. These seals have to satisfy stringent requirements. Helium is expensive and hydrogen is inflammable, so neither must be allowed to escape, the more so as leakage of the working agent would lower the efficiency of the machine. Conversely, the seals must prevent any leakage of lubricating oil or other contaminants into the working space, where they might seriously interfere with the functioning of the machine by, amongst other things, blocking of the regenerator. A certain amount of lubrication, however, is essential, to prevent wear on the bearing surfaces of pistons or rods and their guides.

The design in fig. 1*b* has two rod seals whose function is as described above. In addition, there are two seals that might be described as "internal"; their function is not to shut off the space containing the driving mechanism, but to separate spaces which each contain working agent. Any leakage past these points likewise has an adverse effect on the performance of a Stirling machine, and the need to keep the working spaces uncontaminated makes the lubrication of the bearing surfaces a particularly difficult problem.

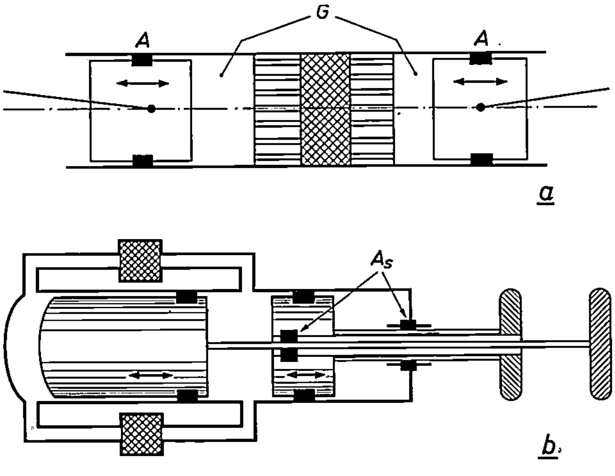


Fig. 1. *a*) Stirling machine working on the two-cylinder principle. Two seals A are necessary between the working space G and the driving mechanism.
b) Stirling machine working on the displacer principle. Two "external" seals A_s for the piston rods are necessary to isolate the working space; in addition, two further seals are required for separation of spaces at the same mean pressure.

Dr. Ir. J. A. Rietdijk, Ir. H. C. J. van Beukering, H. H. M. van der Aa and Dr. Ir. R. J. Meijer are research workers at Philips Research Laboratories, Eindhoven.

[1] J. W. L. Köhler and C. O. Jonkers, Fundamentals of the gas refrigerating machine; Construction of a gas refrigerating machine, Philips tech. Rev. **16**, 69-78 and 105-115, 1954/55.
[2] R. J. Meijer, The Philips hot-gas engine with rhombic drive mechanism, Philips tech. Rev. **20**, 245-262, 1958/59.
[3] A. A. Dros, An industrial gas refrigerating machine with hydraulic piston drive, Philips tech. Rev. **26**, 297-308, 1965.

It would be too much to attempt to deal with all the types of seal that are in principle suitable for the above functions. The chief types can however be enumerated — piston rings, O-rings and sleeves, "small gaps" (with or without labyrinth), and glands (only suitable for rods). None of these types fully meets the above requirements; all, in fact, can be regarded as mere "leakage limiters". Their compromise nature has become more and more keenly realized in the course of the development of hot-gas engines and gas refrigerating machines, one reason having been the increasing requirements for good operating life.

A device of what has been called the "positive seal" type has been under development in our laboratory since 1960. This appears also likely to be of value in machines other than those based on the Stirling cycle. The present article is a detailed account of the new seal and of the manner in which it is applied in Stirling machines. Its application in oil-free compressors will also be discussed. These compressors play an important part in many cryogenic and other processes, and figure in sequential systems embodying the Philips-Stirling gas refrigerating machine [4]. In these compressors it is absolutely essential that there should be no contamination of the compressed medium.

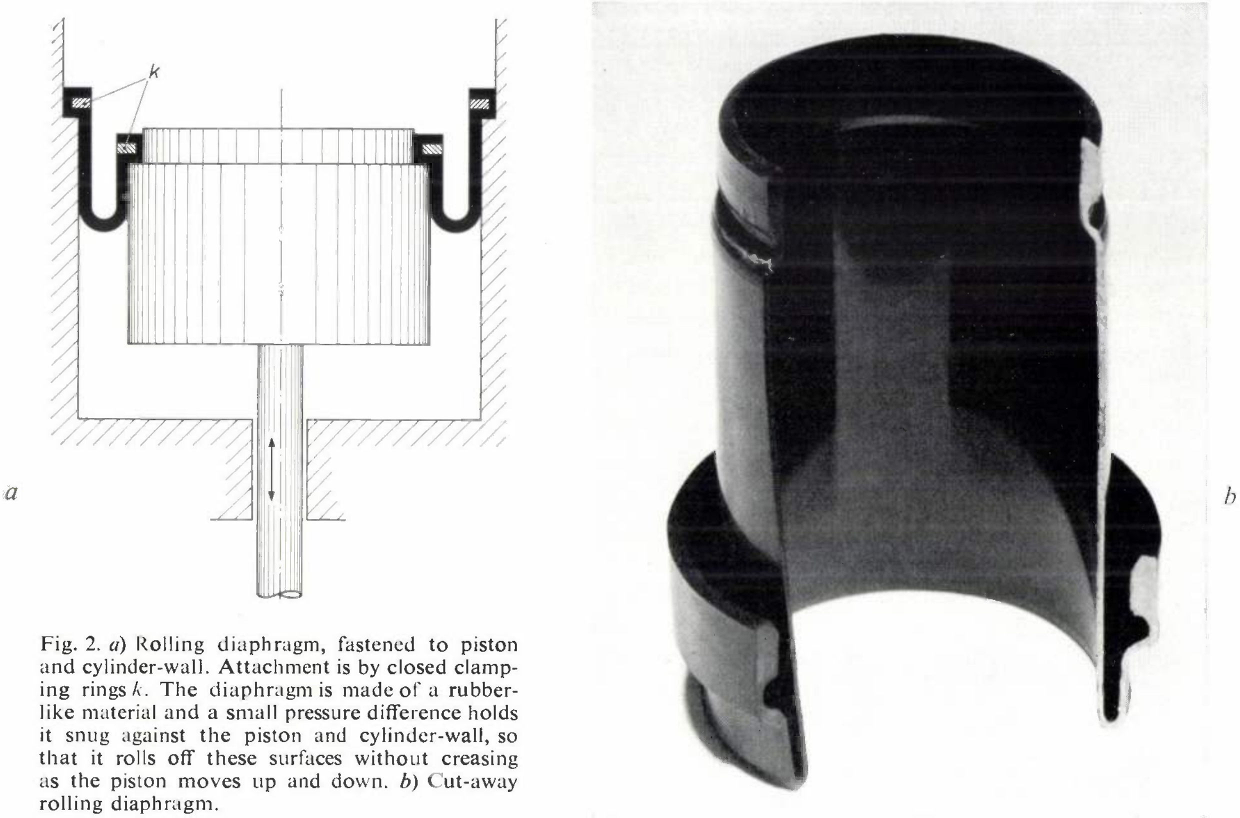


Fig. 2. *a*) Rolling diaphragm, fastened to piston and cylinder-wall. Attachment is by closed clamping rings *k*. The diaphragm is made of a rubber-like material and a small pressure difference holds it snug against the piston and cylinder-wall, so that it rolls off these surfaces without creasing as the piston moves up and down. *b*) Cut-away rolling diaphragm.

Rolling diaphragms

The new seal makes use of a long-known principle, that of the rolling diaphragm. A seal of this kind is shown in *fig. 2a*; fig. 2*b* is a photograph of the diaphragm itself. The diaphragm, which has to be of a rubber-like material, is alternately unrolled from the piston and the cylinder-wall [5]. The movement it performs is rather like that of a stocking or sock; and indeed in the laboratory, the term "sock" has become commonplace. The *guide arrangements* for a piston fitted with one of these "socks" are much the same as for a conventional sealed piston.

Up till now rolling diaphragms have only been suitable for comparatively slow movement, and for *small* pressure differences, such as those in hydraulic and pneumatic control and regulating equipment, pneumatic springs, and so forth. A certain minimum difference of pressure is necessary across the diaphragm in order to prevent creasing. To raise the upper limit of the permitted pressure difference, which was about 5 atm, the diaphragms have generally been reinforced with some kind of fabric that is inextensible except around the circumference, and which performs much the same function as the canvas in a car tyre. Diaphragm life, in terms of the number of strokes before failure; has however been limited, particularly for reinforced diaphragms.

In the following we shall see how, with comparatively simple methods, a rolling diaphragm can be applied at *very high* pressure differences. It has been found that, with a suitable choice of the conditions under which the diaphragm operates, remarkably good life can be achieved.

Rolling diaphragm supported on a fluid cushion

The problem of adapting the rolling diaphragm for high-pressure seals was solved by supporting the diaphragm with a fluid cushion — see *fig. 3.* This solution implies acceptance of the limitations of the rolling diaphragm in regard to permissible pressure difference; across the diaphragm itself there is only a small pressure difference, of about 5 atm. The function of the diaphragm is confined to separating the oil in the cushion from the working agent [6] above the piston. The real difference of pressure (which may range between 50 and 100 atm) is borne by a second seal of conventional type, which separates the oil-filled space under the diaphragm from the space containing the driving mechanism.

To safeguard against rapid wear of the diaphragm, and fatigue due to alternating stresses, it is necessary to ensure that the diaphragm is at all times in close contact with the fluid and *has the same length* throughout its rolling cycle. Once these conditions are realized the reinforcement is no longer necessary and indeed has been found to shorten the life of the diaphragm.

The requirement of constant length is equivalent to one of *constant pressure difference* across the diaphragm. The graph in fig. 3 shows the desired variation in the cushion fluid pressure, for a given pressure variation in the gas. The case illustrated here is that of a "concave diaphragm" occurring when the fluid pressure is lower than the gas pressure. Alternatively, the direction of the pressure difference across the diaphragm can be reversed, so that the fluid pressure is higher than the gas pressure. The diaphragm is then called a "convex diaphragm". This is illustrated in *fig. 4*; the surface from which the diaphragm is rolled is not lubricated here, and this arrangement has in fact been found to be not so favourable.

In general, the tendency will be to employ the lubricating oil of the machine as fluid for the supporting cushion, especially since it has to lubricate the surface of the piston. Accordingly, the cushion will from now on be referred to as an "oil support". Other fluids may in special instances be suitable.

The requirement that the difference of pressure across the diaphragm is kept constant can be satisfied in two ways: by means of the stepped system, or by means of a floating piston-cap. The two methods will now be discussed in detail.

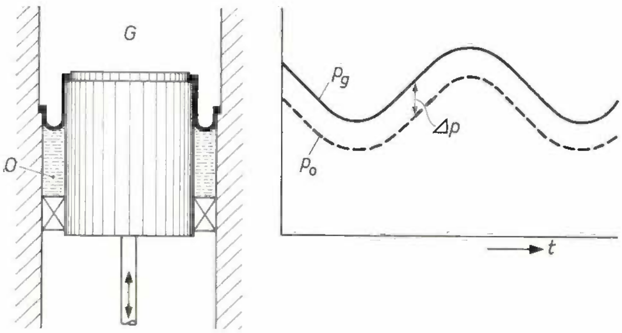


Fig. 3. Rolling diaphragm supported on a cushion of fluid. The fluid O takes up the whole of the pressure exerted by the working agent G except for a small difference in pressure Δp carried by the diaphragm. The pressure p_0 of the fluid must at all times be lower than the pressure p_g of the working agent by a small constant amount Δp. The diaphragm shown here is of the concave type. In the diagram on the right the pressures just referred to are shown as functions of time t.

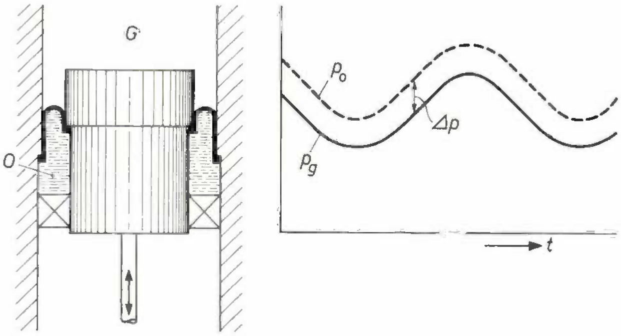


Fig. 4. A rolling diaphragm of the convex type. Here the pressure of the supporting fluid must at all times be *higher* by a certain small amount than the pressure of the working agent to be sealed off. The diagram on the right shows pressures as functions of time.

Methods of keeping the pressure difference constant

Stepped system

The oil space below the diaphragm can be shaped in such a way that the volume it encloses remains constant throughout the cycle for constant diaphragm length. This is ensured by *stepping* both piston and cylinder wall as in *fig. 5*; the area of the step in the piston must be equal to that of the step in the cylinder wall. This condition will be satisfied if the piston and cylinder diameters satisfy the relationship:

$$d_2^2 = \tfrac{1}{2}(d_1^2 + d_3^2) . \qquad (1)$$

[4] G. Prast, A gas refrigerating machine for temperatures down to 20 °K and lower, Philips tech. Rev. 26, 1-11, 1965 (No. 1), particularly pp. 10-11.

[5] The new seal is suitable for rods or for fitting to a piston in a cylinder. Only the cylinder and piston combination will be discussed in this article, but what is said may be considered to apply equally well to a rod passing through a cylinder base.

[6] The medium above the diaphragm will normally be a gas. However, the seal is also suitable, without modification, for use in a pump handling a liquid, for example. In the rest of the article it will be assumed for the sake of simplicity that the space above the seal is occupied by a gas.

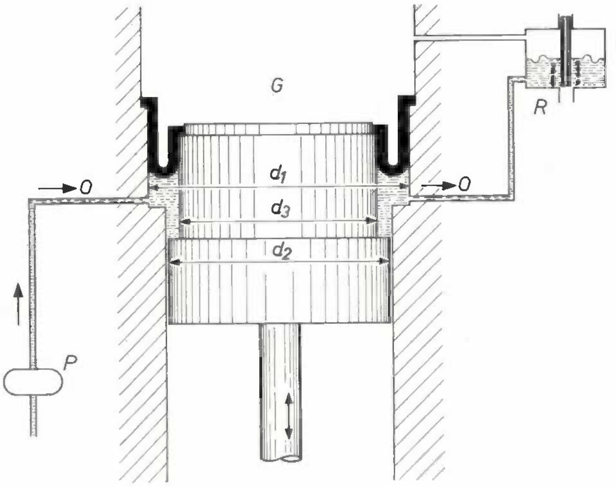


Fig. 5. Rolling diaphragm with fluid support in stepped system. Provided the amount of oil does not change as the piston moves, the length of the rolling diaphragm will not change. Oil *O* is fed by a pump *P* and removed by a control device *R*. The device embodies a spring-loaded metal diaphragm which senses the difference between the mean gas and mean oil pressures, and controls two valves. One is the safety valve which releases gas when the pressure difference becomes too large; the other is an oil escape valve which, in normal operation, is at the "cut-off point". This method of keeping the diaphragm seal length constant is suited to both the convex and the concave types of diaphragm.

A proof that the enclosed volume remains constant can be obtained by considering a certain displacement of the piston. If the length of the diaphragm remains the same, the fold in it travels half this distance; the step in the piston — which occupies half the area of the gap between piston and cylinder — travels the *whole* distance. Clearly, then, there has been no change in the overall volume enclosed. The constant pressure difference requirement is therefore met if the space is filled with a quantity of oil such as to produce the required stressing of the diaphragm, which may be convex or concave.

In practice there will be slight deviations from the ideal behaviour just described. Amongst these are the compressibility of the oil, the fact that the walls of the space are not perfectly rigid, and within-tolerance deviations in the various diameters. There will, therefore, be some slight fluctuation in the length of the diaphragm. The above deviations can be kept within limits close enough that there is no difficulty with non-reinforced rolling diaphragms, despite the small changes in their length. In reinforced diaphragms, however, very severe stresses would be set up. In short, the stepped system can only be employed in conjunction with *non-reinforced* rolling diaphragms.

The oil support itself calls for some special measures. Oil will escape from the cushion space past the oil-seal, which has to be one of the leakage-limiting types already referred to. The loss is made up by feeding in an excess of oil by means of a miniature high-pressure pump [7]. The excess is removed via a *control device* that maintains a constant difference between the oil and gas mean pressures.

The control device can take the form of a regulating valve which is located away from the diaphragm and which also serves as a safety valve. An arrangement of this kind is shown in fig. 5; its mode of action is explained in the caption. Another possibility is to use the diaphragm itself as a control element, as in *fig. 6*. Here an escape orifice is provided in the cylinder wall; this is uncovered during the short interval in which the diaphragm is in its topmost position. In this position the rate of loss is strongly dependent on very small changes in the length of the diaphragm. This arrangement makes it possible to remove oil at a varying rate while the length of the diaphragm — and hence also the pressure difference — remains very nearly constant.

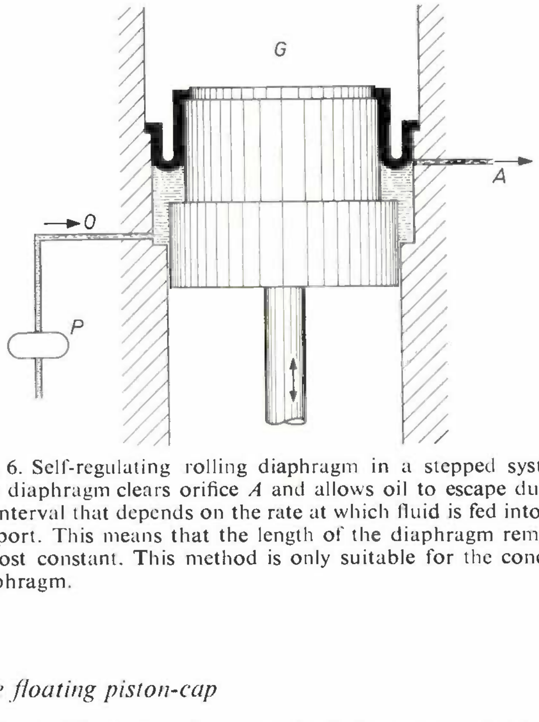


Fig. 6. Self-regulating rolling diaphragm in a stepped system. The diaphragm clears orifice *A* and allows oil to escape during an interval that depends on the rate at which fluid is fed into the support. This means that the length of the diaphragm remains almost constant. This method is only suitable for the concave diaphragm.

The floating piston-cap

Fig. 7 illustrates the second of the two methods, in which the primary object is to maintain a constant difference of pressure across the rolling diaphragm. It is only suitable for piston seals. The piston in question is equipped with a cap that is free to move in the axial direction. The cap is connected by a spring to the body of the piston. The rolling diaphragm (which can be convex or concave, as before) extends between the cap and the cylinder wall. The part of the piston body on which the cap rides is so constructed as to allow a free passage of oil between the space under the diaphragm and that between the piston and the cap.

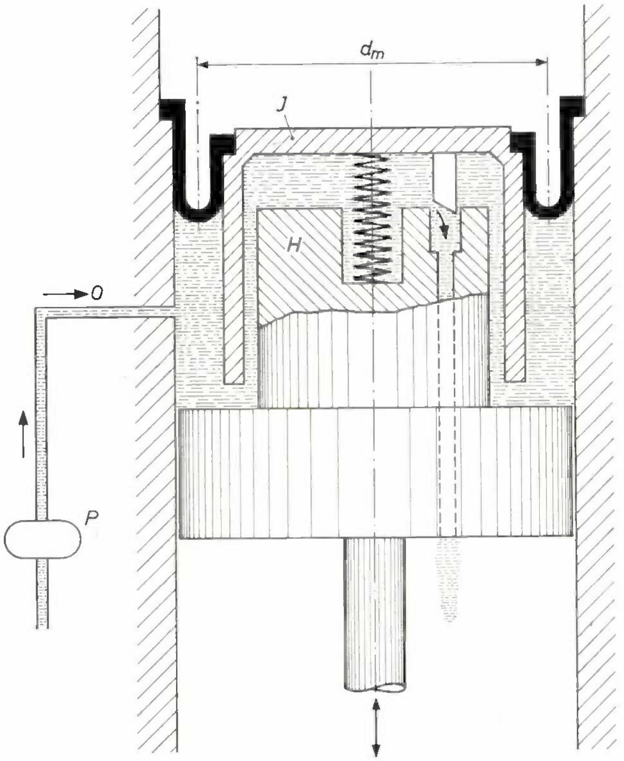


Fig. 7. Rolling diaphragm fitted on a piston with a "floating cap". The free cap *J* and the main piston *H* act together as a regulating system. The difference of pressure across the diaphragm is kept constant by means of a spring between piston and cap. A spring is chosen whose stiffness is so small that the force is more or less unaffected by small movements of the cap relative to the body. This construction is suitable both for concave and convex diaphragms.

Oil is supplied to the support in exactly the same way as in the first method. The excess is removed by an arrangement that comes into action as soon as the cap has reached a certain peak height above the piston body. The arrangement can consist either of a port in the piston wall or a dipping pin in the base of the piston, as in fig. 7. The spring is so dimensioned that the force *K* it exerts in its topmost position satisfies the relation:

$$K = \Delta p \frac{\pi}{4} d_{\mathrm{m}}^2, \qquad \qquad (2)$$

where Δp denotes the desired difference of pressure across the diaphragm. Thus the piston-cap combination functions at the same time as a regulating valve. There is no need here for the steps cut into the piston and cylinder wall to satisfy closely the condition expressed by eq. (1), since a small periodic displacement of the cap with respect to the piston body is enough to keep the oil volume constant. Provided the spring is not too stiff, the extensions it undergoes will not be accompanied by any appreciable variation in force *K* or, therefore, in the difference of pressure across the diaphragm. If a sufficiently weak spring is employed, the stepped geometry can be done away with altogether, as in fig. 7. At the same time this widens the choice of oil feed arrangements; a pulsating supply can be used if desired, the oil being fed in at a high rate by a plunger-type pump.

When discussing the oil-free compressor we shall see that this method also enables simple automatic control of the dead space volume in piston machines.

Both methods imply circulation and hence renewal of the oil, and this has two advantages. Gas may diffuse through the diaphragm and into the oil support, which as a result ceases to be incompressible; this may interfere with its action, especially in the stepped system. By circulating the oil the diffused gas is removed. At the same time the oil acts as coolant, and this lengthens the life of the seal.

It may also be noted that the oil supporting the rolling diaphragm ensures as near perfect lubrication of the piston surface as possible.

The foregoing will have made it clear that the rolling diaphragm has reduced the problem of sealing off the gas-filled space to one of sealing off an oil-filled space from which a certain amount of leakage is permissible. Conventional seals can be employed for this, and these are quite capable of standing up to pressure differences of hundreds of atmospheres.

Results

It will be clear that the rolling diaphragm seal depends completely on the availability of a suitable rubber-like material (elastomer). To be suitable, the material must have a high fatigue strength and creep resistance and not be subject to chemical attack by the oil or the working gas.

Fortune was with us when we tried out the first material, a polyurethane rubber recommended by the Philips Plastics Laboratory, and which proved to satisfy these requirements quite well. While investigating the material we discovered that its endurance is largely dependent on three parameters — temperature, the pressure difference across the diaphragm, and the ratio δ between $(2d_0)$, the double thickness of the diaphragm, and the piston-cylinder wall clearance (s). Some of our results are displayed in *figures 8* and *9*.

It will be seen from fig. 8 that the life of the diaphragm falls off sharply with rising temperature. This effect may be connected with the marked loss of the

[7] An elegant solution to the oil feed problem is also offered by the "pumping ring" invented by H. J. Verbeek of Philips' Industrial Equipment Division. The primary function of the ring is to seal off the oil space, but its shape is such that the piston movement sets up a hydrodynamic pump action, which pumps oil directly into the support space from the space containing the driving mechanism.

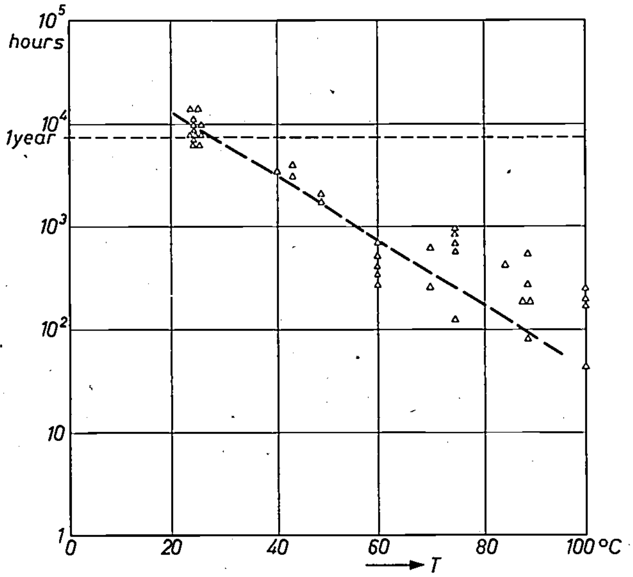


Fig. 8. Life in hours of rolling diaphragms made of polyurethane rubber, as a function of temperature T. Pressure differences across the diaphragms thus tested ranged from 4 to 6 atm. At zero pressure difference the diaphragms had a thickness of d_0 = 0.5 mm; the piston-cylinder clearance was s = 2 mm, the shaft speed 1500 r.p.m., and the piston stroke 65 mm. Raising the temperature claerly shortens the life of the diaphragms markedly. At room temperature almost all the diaphragms tested lasted for more than a year.

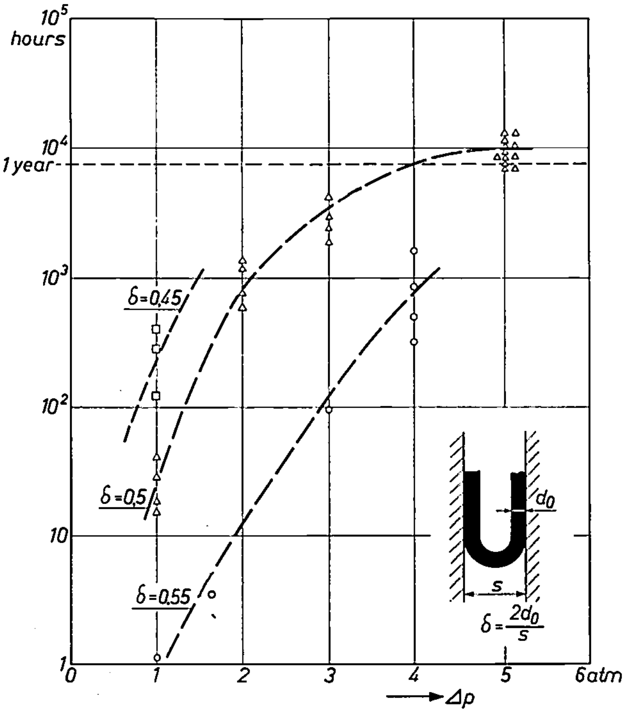


Fig. 9. Life in hours as a function of pressure difference Δp and ratio δ. Both these quantities affect the amplitude of the periodic elongation which the material undergoes during the cycle (see appendix) and hence the life of the diaphragms.

tensile strength of the material at higher temperatures. At 100 °C it has only 20% of the tensile strength it possesses at room temperature. Fig. 9 reveals that up to a certain limit, the life of the diaphragm at a given temperature can be lengthened by stepping up the pressure difference Δp. It will also be seen that for a given pressure difference, the life of the diaphragm is strongly dependent on δ. The life of the seal has been found to be independent of the absolute pressure level.

With a view to explaining the influence of Δp and δ we endeavoured to set up a simple theory and carry out calculations that would cast light on the variation in the stresses set up in the rolling diaphragm, and on the deformation it undergoes. This theoretical study is reported in the *appendix* to this article.

The results displayed in figures 8 and 9 relate to rolling diaphragms fitted to a piston with a 65 mm stroke. The machines on which the diaphragms were tested had a shaft speed of 1500 r.p.m. The longest-lived diaphragm so far tested stood up to 13 000 hours of operation (about 18 months), during which the rubber was flexed more than two thousand million times.

The abandonment of reinforcement of the diaphragms has removed a limitation on the ratio between stroke and diameter, a limitation inherent in conventional rolling diaphragms. The dimensions of the diaphragms so far tested have ranged from 6 mm diameter and 0.1 mm thickness, for a 0.4 mm piston-cylinder clearance, to 175 mm diameter and 0.7 mm thickness for a 4 mm clearance.

Research is now in progress on materials which stand up to various chemically aggressive substances and to high temperature operation. Both properties are especially important in compressors.

Applications

Gas refrigerating machine

First mention should be made here of the large gas refrigerating machine that has been developed by the Cryogenics Department of the Philips Industrial Equipment Division. This machine, known as type C, combines the rolling diaphragm seal with hydraulic piston drive in a highly elegant way; it is discussed at length in another article appearing in this number, which has already been referred to [3]. This machine has the stepped oil space system.

Hot-gas engine

A single-cylinder hot-gas engine equipped with rolling diaphragms on a fluid support has been built in our laboratory. The stepped system was again used. The engine is shown schematically in *fig. 10*. The thermodynamic part of the machine, i.e. the working space in which the Stirling cycle takes place, is shut off by a piston from a buffer space filled with gas at a pressure equal to the mean pressure of the working agent. The purpose of the buffer space is to take up as much as possible of the thrust on the driving mechanism.

The engine incorporates four rolling diaphragm seals, two of which can be classed as "external". One of these is fitted to the displacer rod, the other to the piston rod. The other two are fitted to the piston itself, as shown in fig. 10. As can be inferred from the diagram all these seals operate at room temperature. The displacer rod seal and the upper piston seal are acted upon by the pressure of the working agent, which varies during the cycle; the one on the piston rod and the lower piston seal are both exposed to the pressure of the gas in the buffer space. The four seals thus fall into two groups of two, and there is communication between the two oil spaces in each group. Each group further has its own oil-pump and regulating valve.

In this particular design the oil spaces belonging to the two piston seals are adjacent but separate; transference of oil from one to the other has to be prevented because there is a periodically alternating difference of pressure between the two oil supports, the peak value of which is in the region of 40 atm. The seal between the two spaces is provided by a piston ring. Although there are special problems, which we will not go into here, in the use of piston rings in this application, the arrangement is acceptable.

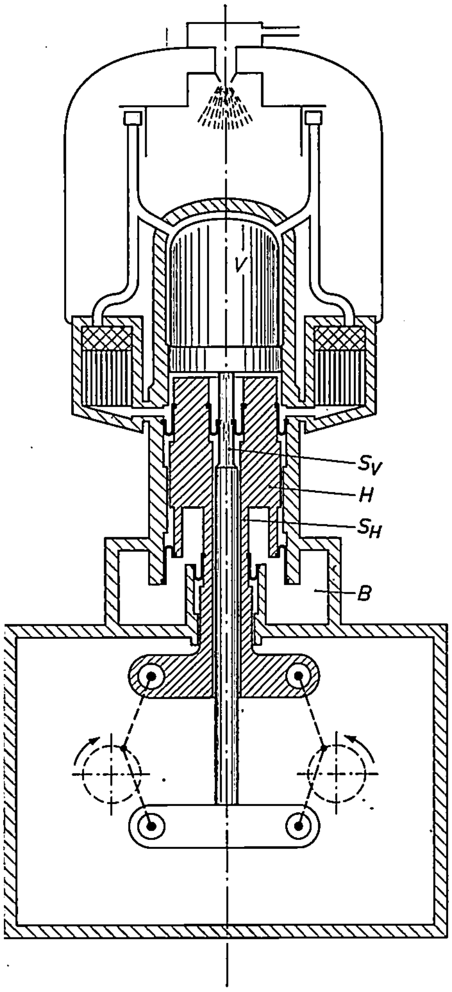
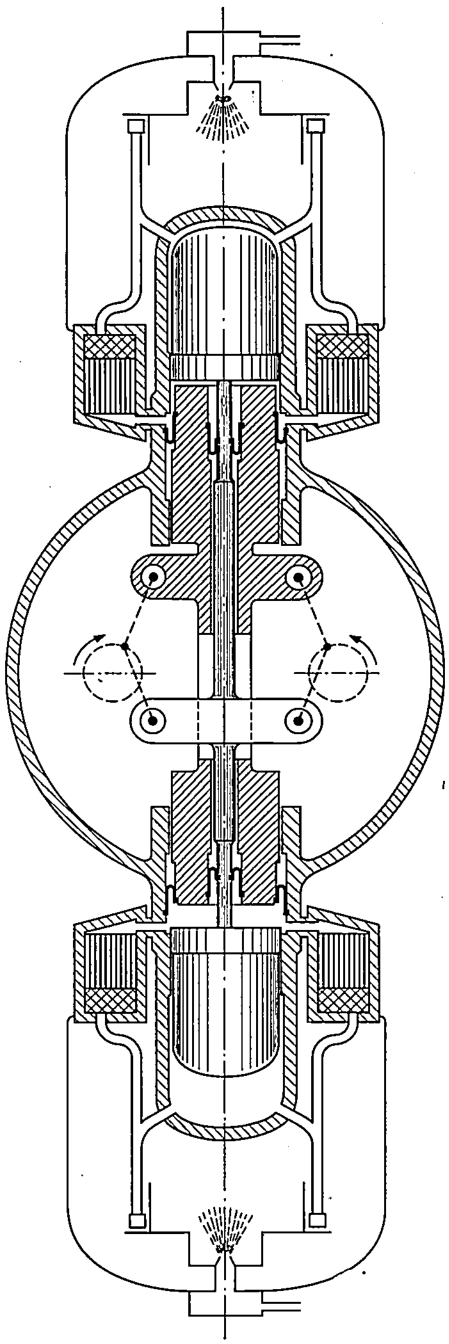


Fig. 10. Experimental model of a hot-gas engine embodying four rolling diaphragms and developing a shaft power of 80 h.p. at 1500 r.p.m.

During the cycle the working agent (hydrogen) attains a peak pressure of 140 atm. *H* piston. *V* displacer. S_H piston rod. S_V displacer rod. *B* buffer space.



→

Fig. 11. Basic design of a horizontally-opposed type of hot-gas engine. Here the function of the buffer space in fig. 10 (which is to take up the mean thrust on the piston) has been taken over by a second working space; the cycle taking place in the second working space is exactly the same as that in the first. The pistons are mounted on the same rod. The number of rolling diaphragm seals per cylinder is reduced to two.

The buffer spaces in hot-gas engines are generally difficult of access, and in fact represent an undesirable complication. For this reason work has begun on a new type of engine in which the function of the buffer space will be performed by a second working space. The two pistons of this "horizontally opposed" engine are fitted to a common piston rod, so that each compensates for the mean thrust of the other. The arrangement is shown in *fig. 11.* The number of sealing diaphragms per working space has been reduced to two, and all are easily accessible. This design would not be practicable were it not for the positive seals fitted to the two pistons; a conventional type of seal would allow too much leakage of gas into the space containing the driving mechanism. Conventional seals only allow small diameter constructions in which the pistons have to be linked to the driving mechanism by relatively thin rods.

Oil-free compressors

The rolling diaphragm seal can also be applied in oil-free compressors, and in view of the importance of these machines, particularly in cryogenics, a great deal of attention is being paid to this application of the new seal. Two experimental piston-type compressors, fitted with rolling diaphragms, have been built. One is a single-stage machine supplying a pressure of 4 kg/cm^2 and having an output of about 45 normal m^3/h, and the other is a three-stage machine supplying a pressure of 150 kg/cm^2 and having an output of about 30 m^3/h. The employment of convex diaphragms makes both machines suitable for a subatmospheric gas intake, and both embody pistons of the floating-cap type. They differ only in constructional details and the performance data given above. Accordingly, we shall only deal here with the single-stage compressor, which is shown schematically in *fig. 12*.

Piston *1* is equipped with a cap *2* which, in this particular case, has a large displacement with respect to the piston. The freedom of movement possessed by the cap makes it possible, during operation of the machine, to regulate the volume of the dead space (the free space remaining in the cylinder when the piston is at the top of its stroke), and hence the output of the compressor. The cap is linked to the piston body by a spring *3*. Oil is fed into the intervening space by a plunger-type pump *4* which is built into the piston body. It flows along the inside of the cap towards the space under the rolling diaphragm, and this gives very effective cooling of the cap. At *5* there is an overflow which opens when the cap has reached a certain maximum height above the piston body. Alternatively, an excess of oil can escape by duct *6* to control device *7*. The action of this device is governed in the first instance by the pressure $\hat{p}$ in the compression vessel *8*, the valves closing as soon as pressure $\hat{p}$ falls below a certain value which is adjustable, and opening as soon as this value is exceeded. The effect of this is that the volume of the dead space is continuously adjusted as required. This control action must not interfere with maintenance of the correct pressure difference across the rolling diaphragm, and control device *7* therefore also contains an arrangement that prevents the pressure difference falling below a certain value (the latter function is performed by the upper part of *7*). It is naturally possible to arrange for the control device to respond to some quantity other than pressure. The device is to be regarded as a refinement that can be left out of the design if the regulating function is not required.

Immediately the machine is stopped, oil starts to leak past the piston body into the crankcase, and the cap gradually drops back into a starting position as a result of the pressure exerted by spring *3* and that of any gas present in the compression space. An O-ring *9* is so dimensioned that it clears the cylinder wall when the machine is running; when the machine is at rest the ring is pressed flat, and seals off the gap between the piston body and cylinder wall. This means that enough oil remains in the support space to keep the diaphragm distended, so that the machine is ready for the next start. The machine therefore starts at maximum dead space, i.e. more or less unloaded.

This single-stage compressor has so far been running continuously for a good 4000 hours at a speed of 1500 r.p.m. A life-test of the three-stage compressor has only just begun.

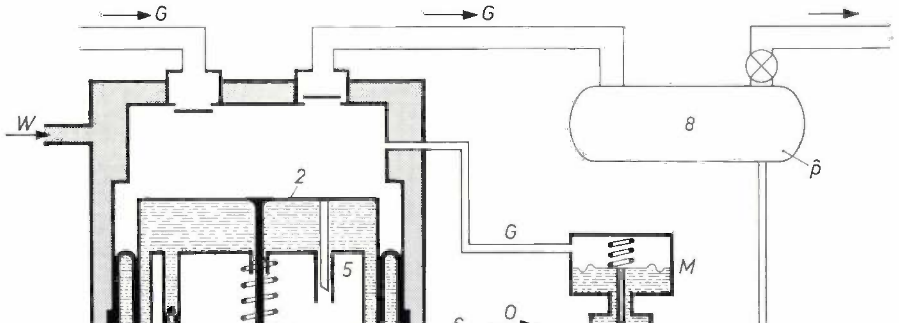


Fig. 12. Single-stage compressor embodying a piston with a floating cap, fitted with a convex rolling diaphragm (schematic). The control system responsible for maintaining a constant pressure across the diaphragm is designed to provide at the same time automatic adjustment of the dead space volume. The latter control function is performed by oil escape valve *7* which is governed by the pressure $\hat{p}$ in compression vessel *8*. The pressure difference across the rolling diaphragm is limited on one hand by the escape of oil through overflow *5*, and on the other by an extra valve actuated by the metal diaphragm *M* in the control device. The compression cylinder is water-cooled, the water entering and leaving at the points marked *W*.

Appendix: stresses in the rolling diaphragm

From the fundamental theory of elasticity any stressed state may be represented by a system of principal stresses acting along orthogonal axes, as in *fig. 13*. We shall take σ_x, σ_y and σ_z to denote the principal stresses, the subscripts referring to the directions in which they act. Here we define stress as force per unit area of an element of the body undergoing deformation; the planes on which the principal stresses act are distinguished by their lack of shear stress. We shall take λ_x, λ_y and λ_z to denote the relative deformations in the x, y and z directions (i.e. the dimensions after deformation divided by the original dimensions). If we started with a unit cube before loading, then λ_x, λ_y and λ_z will represent its new dimensions under load.

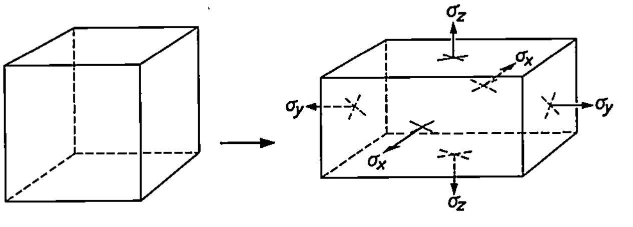


Fig. 13. Deformation of unit cube acted on by principal stresses σ_x, σ_y, σ_z.

The behaviour of a stressed elastomer can be described, to a fair degree of approximation, by the following equations:

$$\lambda_x \lambda_y \lambda_z = 1, \quad \ldots \ldots \ldots \ldots \quad (3)$$

$$\left.\begin{aligned} \sigma_x &= C e^{\lambda_x} - p, \\ \sigma_y &= C e^{\lambda_y} - p, \\ \sigma_z &= C e^{\lambda_z} - p. \end{aligned}\right\} \quad \ldots \ldots \ldots \quad (4)$$

Eq. (3) takes account of the extremely low compressibility of this kind of material. C is a material constant which in our case has a value of 9.2 kg/cm^2 [8]. p is taken as the "hydrostatic" pressure of the material, and is a function of σ_x, σ_y and σ_z. When $\sigma_x = \sigma_y = \sigma_z = 0, p = Ce$.

We shall start off by considering the stresses in the straight part of a rolling diaphragm supported by a fluid in which a pressure of p_2 prevails, the other side of it being acted upon by a higher pressure p_1 (see *fig. 14*, which shows the dimensions of the system and the axes along which the stresses act). We shall assume that $s \ll D$, in order to be able to neglect circumferential extension. (The results of the calculation will show that this is justified, for in practice the circumferential extension is very much less than that in other directions.) Hence we may put $\lambda_z = 1$. Now let us consider a bar of unit width, perpendicular to the plane of the drawing. If the vertical forces are in equilibrium we can write:

$$2d\sigma_x + p_2 s = p_1 (s - 2d),$$

or

$$\Delta p = p_1 - p_2 = \frac{2d}{s}(\sigma_x + p_1) . \quad \ldots \quad (5)$$

Here it has been assumed that the wall exercises no shearing force on the rolling diaphragm. It is not difficult to see that the principal stresses in the straight portions of the rolling diaphragm act in the x, y and z directions shown in the figure and, in the case under consideration, that $\sigma_y = -p_1$. We can also infer that $d = \lambda_y d_0$, where d_0 is the initial or unloaded thickness of the diaphragm. Using (3) and (4), taking $\lambda_z = 1$, then (5) becomes:

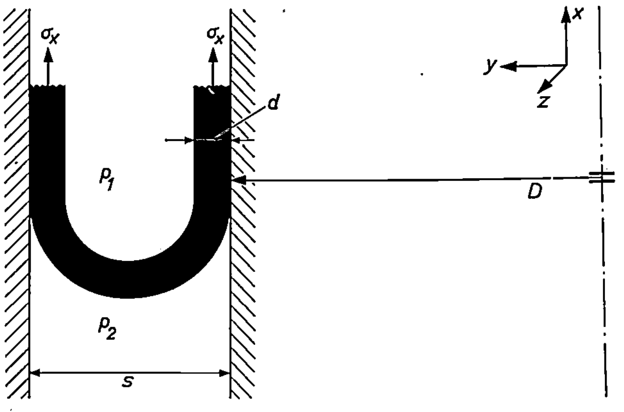


Fig. 14. Diagram to illustrate the derivation of stresses and strains operative in the straight part of a rolling diaphragm.

$$\Delta p = \delta C \frac{e^{\lambda_x} - e^{1/\lambda_x}}{\lambda_x}, \quad \ldots \ldots \quad (6)$$

where δ, as already defined, is

$$\delta = \frac{2d_0}{s} . \quad \ldots \ldots \ldots \quad (7)$$

Measurements on a rolling diaphragm at room temperature, where $\delta = 0.5$, have confirmed that (6) corresponds closely with practical results.

Only the pressure difference figures in (6), not the absolute pressure level. The situation has been found rather analogous for deformation of the curved part of the diaphragm, which we shall now consider. Thus the elastic behaviour of a rolling diaphragm does not depend on the pressure level at which the machine or engine cycle is taking place — a conclusion that has been confirmed by the experiments.

In dealing with the curved section of the diaphragm (see *fig. 15*) we shall make two assumptions. Firstly, we neglect the circumferential extension, as before ($s \ll D$). We shall further assume that the curvature is strictly circular. The main axes x, y and z, along which the stresses are concentrated, will therefore be as shown in fig. 15. We can again say that $\lambda_z = 1$, so that $\lambda_x \lambda_y = 1$. Let us now consider a small volume whose dimension perpendicular to the plane of the drawing is unity, and whose other dimensions are r dφ and dr. If forces in the y direction (r direction) are in equilibrium, we can write:

$$\sigma_x \mathrm{d}r \, \mathrm{d}\varphi = \frac{\mathrm{d}(\sigma_y \, r \, \mathrm{d}\varphi)}{\mathrm{d}r} \, \mathrm{d}r,$$

or

$$\sigma_x - \sigma_y = r \frac{\mathrm{d}\sigma_y}{\mathrm{d}r} . \quad \ldots \ldots \ldots \quad (8)$$

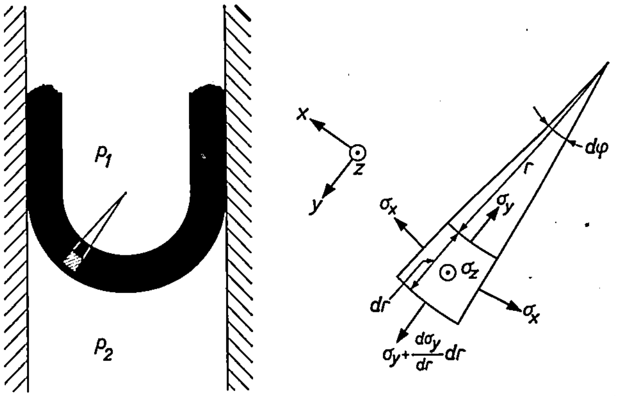


Fig. 15. Diagram illustrating the derivation of stresses and strains in the curved part of a rolling diaphragm. The volume element in the sketch on the left is shown on a larger scale on the right. The principal stresses are along the main axes x, y, z.

[8] The above theoretical approach does not take thermal effects and creep into account. The figures quoted relate to room temperature. A different version of eq. (4) is to be found in the literature, namely $\sigma_i = C' \lambda_i^2 - p$ (where $i = x, y, z$). This formula is only true for strains up to $\lambda = 2$ (i.e 100% extension).

This being so:

so that

$$\left.\begin{aligned}\lambda_x &= \pi r/l_0\,,\\ \lambda_y &= l_0/\pi r\,.\end{aligned}\right\}\quad\ldots\ldots\ldots\ldots\quad(9)$$

The quantity l_0 is the initial "fibre-length" of the material in the curved portion; it is independent of r, and in fact it depends only on the boundary conditions of the problem. From (4), (8) and (9), we get:

$$\frac{\mathrm{d}\sigma_y}{\mathrm{d}r} = \frac{C}{r}\left(\mathrm{e}^{\pi r/l_0} - \mathrm{e}^{l_0/\pi r}\right). \quad\ldots\ldots\quad(10)$$

The boundary conditions are:

$$\left.\begin{aligned}\sigma_y &= -\,p_2 \quad\text{for}\quad r = \tfrac{1}{2}s,\\ \sigma_y &= -\,p_1 \quad\text{for}\quad r = r_t,\end{aligned}\right\}\quad\ldots\ldots\ldots\quad(11)$$

where r_t is the radius of curvature of the inside surface. In the simple situation represented in fig. 14, $d_0 = d/\lambda_y$, but this does not now apply; instead:

$$d_0 = \int_{r_t}^{\frac{1}{2}s} \frac{1}{\lambda_y}\,\mathrm{d}r = \int_{r_t}^{\frac{1}{2}s} \lambda_x\,\mathrm{d}r\,,$$

and from this and (9) and (7) it follows that

$$r_t = \tfrac{1}{2}\,s\sqrt{1-\frac{4\delta l_0}{\pi s}}\,. \quad\ldots\ldots\quad(12)$$

We are now in a position to calculate a number of unknowns by numerical integration of (10), using (9), (11) and (12). In this way we can calculate λ_x and λ_y, and if desired, σ_x, σ_y and σ_z, as functions of Δp, δ and r.

Strains λ_x in the x direction calculated in this manner for the straight part of the diaphragm, confirmed, as we have already stated, by measurements, and also the strains for the inside and outside surfaces of the curved part, have been plotted in *fig. 16* as functions of Δp for two values of δ. As the diaphragm rolls up and down, a volume element on the inside of the diaphragm will undergo a strain (extension) in the x direction whose magnitude swings periodically between the curve appropriate to the straight part and the curve appropriate to the relevant point in the fold. An analogous situation applies for any volume element on the outside of the diaphragm. Thus the spacing of these curves determines any fatigue effect manifested by the rolling diaphragm (volume elements *inside* the diaphragm in every case undergo less deformation than those at the surface).

The first conclusion we can draw from fig. 16 is that the smaller δ is made, the smaller will be the periodic deformation undergone by both the inside and outside surfaces of the diaphragm at a given difference of pressure Δp. This conclusion is in accord with the experimental evidence displayed in fig. 9 that diaphragms with smaller δ last longer.

As Δp increases, the periodic deformations undergone by the outside surface gradually become less, while those undergone by the inside surface either do not change at all or increase slightly. This theoretical inference must be linked up with two experimental facts. Firstly, rolling diaphragms subjected to endurance tests have always failed at the *outside surface* first; irrespective of whether the diaphragm has the convex or the concave form, the first tear appears on the outside and the plane of the break is always perpendicular to the x direction. Secondly, the life of the diaphragm increases with Δp, as is apparent from fig. 9. We must conclude that *the deformation limits do not form the only factor determining the life of the diaphragm.* The truth of this can be seen as follows. At a certain small pressure difference $(\Delta p)_1$ the strain in the outside surface varies between the extremes indicated by the double arrow on the left of fig. 16. In due course, a tear will form in this surface. At a much greater pressure difference $(\Delta p)_2$ the strain in the *inside* surface varies between almost the same limits, but *no* tear will form here within the same period of time; the life of the diaphragm at this Δp value is much longer, and failure will finally occur at the *outside* surface, as before.

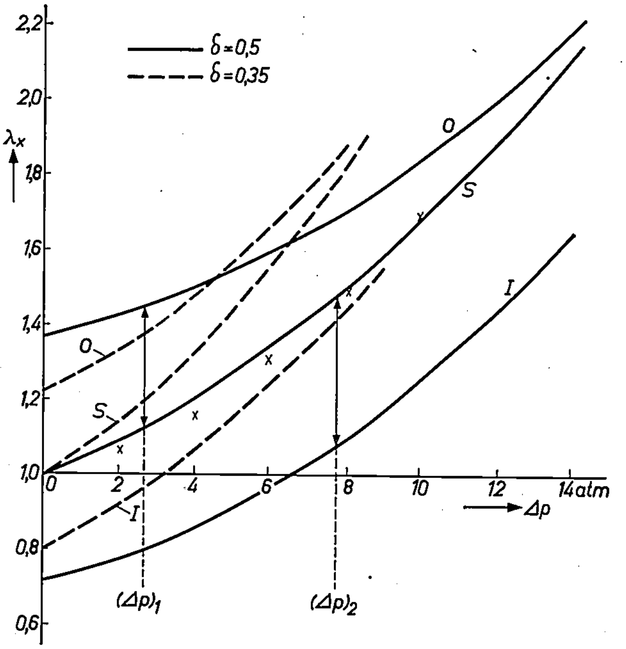


Fig. 16. Theoretical strains in the x direction as a function of pressure difference Δp for two values of δ. The curves relate to deformation λ_x in the straight part (S) and at the inside and outside surface (I and O respectively) of the curved part of a rolling diaphragm at room temperature. The crosses represent measured values obtained from the straight part of a diaphragm with $\delta = 0.5$.

What really governs the life of the rolling diaphragm is not yet known with certainty. The fact that failure invariably occurs on the outside surface first would suggest that a part is played by the difference in the *character* of the deformation. In the course of the rolling cycle, the permanent strain at the outside surface is increased for a brief interval; that at the inside surface is briefly diminished. There is yet another point of difference between the outside and inside surfaces. For the outside surface, the periodic changeover from the straight to the curved condition and back again is associated with a making and breaking of contact with the piston or cylinder wall. This inevitably involves some local sliding movement of the diaphragm over the wall, which will certainly have an adverse effect upon its life. The greater the periodic change in the strain value, the greater will be this sliding movement with respect to the wall. As fig. 16 shows, the movement decreases as Δp increases, and this is in agreement with the results displayed in fig. 9: the greater Δp, the longer is the life of the diaphragm.

Summary. It has been found that the rolling diaphragm, a familiar type of positive seal for pistons or rods, can be employed at large pressure difference when supported by a cushion of fluid. The seal must be designed in such a way that the length of the diaphragm remains very nearly constant during the rolling cycle. Two possible designs are described, one embodying a stepped oil space, the other a floating piston-cap. Both are equipped with control devices that keep the pressure difference across the diaphragm (and hence also its length) constant to within very close limits. Diaphragms made of polyurethane rubber, used in seals of this type, have proved to have extremely long life, standing up to more than 2×10^9 rolling cycles at a rate of 1500 cycles per minute. The influence of various parameters is investigated and a theoretical study is given. Applications of this kind of seal in a gas refrigerating machine, a hot-gas engine and an oil-free compressor are dealt with. The last-mentioned machine is described in rather greater detail.

An industrial gas refrigerating machine with hydraulic piston drive

A. A. Dros

621.573

The range of Philips gas refrigerating machines has been extended by the addition of a new machine that combines a high refrigerating capacity with great reliability and very high efficiency (some 1.3 times that of the smaller models). Its improved performance stems from the application of new constructional ideas which have given the machine a rather unusual appearance. The article below explains the logical development behind this revolutionary design.

Gas refrigerating machines working on the Stirling principle have been manufactured by Philips for a good many years. The small single-cylinder machine (type A) and the four-cylinder model (type B) are in use in laboratories throughout the world for generating cold at very low temperatures.

In recent years industry has also shown a growing interest in the machines, whose compactness, quiet running and ease of operation are exceptional. The industrial user demands however a great deal more than the scientist in efficiency, low maintenance costs, and reliability under continuous operation. In addition, higher capacities per unit are usually needed.

Some years ago Philips embarked on the development of a large gas refrigerating machine that would meet these more stringent industrial requirements. The new machine, known as type C, has like the four-cylinder type B been designed and developed by the Industrial Equipment Division (PIT).

During the development it became clear that design features proved in small gas refrigerating machines were not necessarily applicable in a large one. More was now known about the Stirling cycle, and this deeper theoretical knowledge, together with new ideas in machine design, opened up completely new paths. The result is a machine perhaps somewhat revolutionary in comparison with type B; see *figs 1* and *2*. The new machine has a substantially higher capacity and efficiency: its refrigerating capacity at 77 °K is 20 kW, as against 3.8 kW for type B, and its relative efficiency (relative, that is, to the Carnot cycle) is over 41 % as against about 30 % for type B. Further data are given in the caption to fig. 2. Delivery of the first production run of type C machines has been completed.

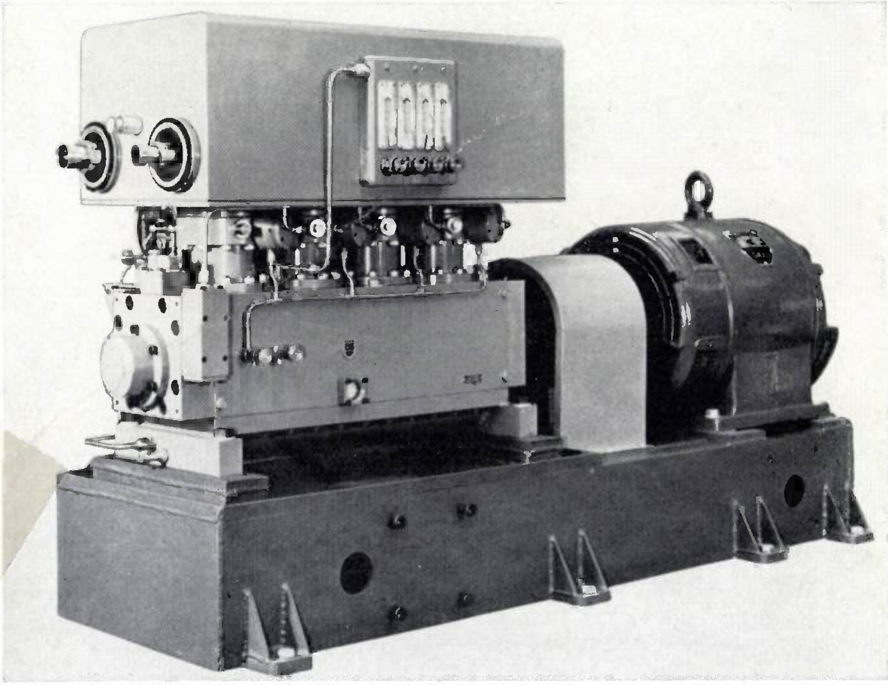


Fig. 1. The Philips type B gas refrigerating machine, a four-cylinder machine widely used in large laboratories. It has a refrigerating capacity of 3.8 kW at 77 °K when cooled with water at 15 °C. Under these conditions its efficiency relative to the Carnot cycle is $\eta/\eta_C = 30\%$. (See also the photograph of a liquid nitrogen plant in Philips tech. Rev. **25**, page 341, 1963/64.)

Ir. A. A. Dros is on the staff of the Philips Industrial Equipment Division, Eindhoven.

To aid the correct understanding of the construction and functioning of the new machine, to be described below, it is useful to start with a brief recapitulation of the Stirling principle, explained in terms of a schematic cycle [1].

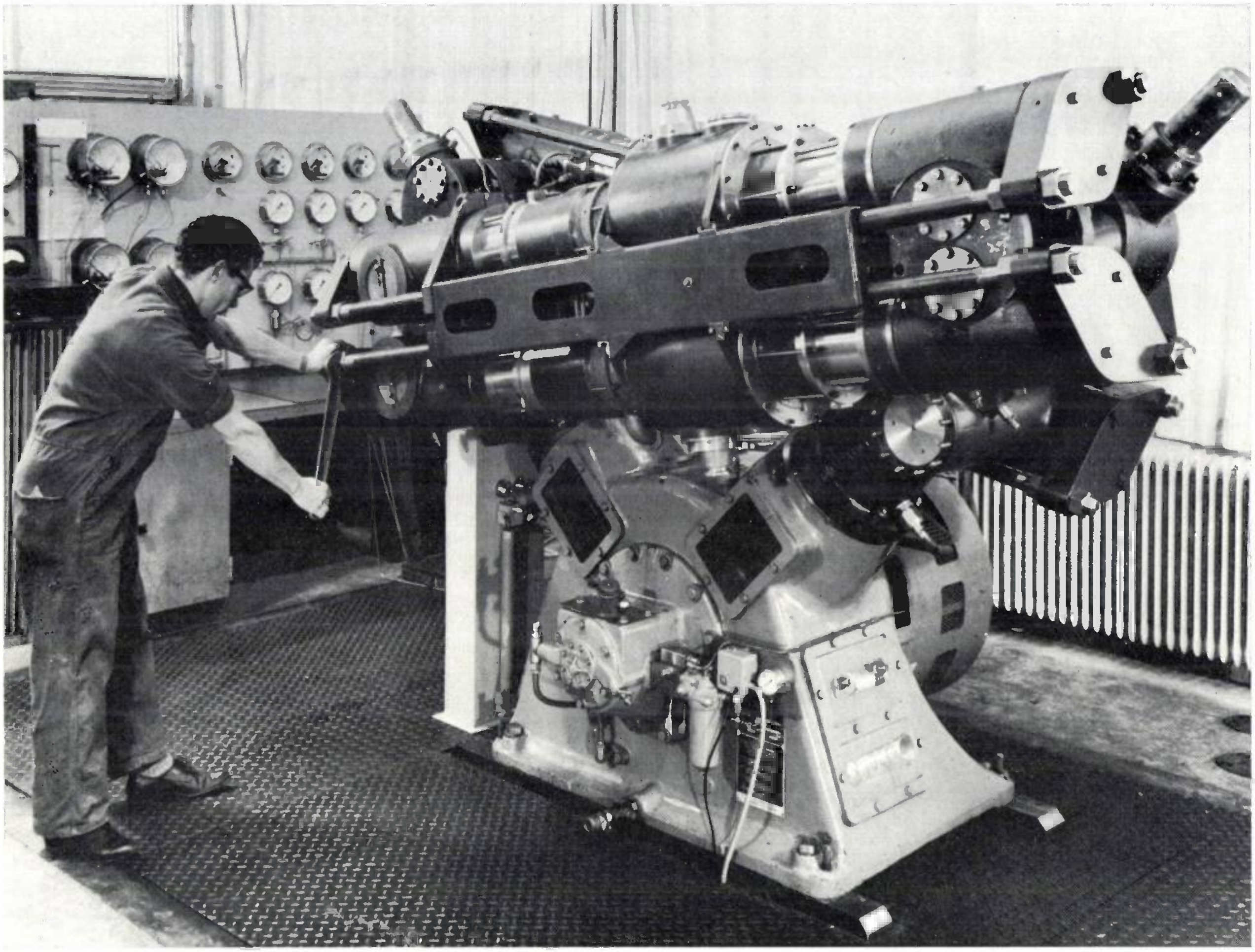


Fig. 2. The new gas refrigerating machine type C. It has a refrigerating capacity of 20 kW and a relative efficiency of over 41 % at 77 °K, when cooled with water at 15 °C. Other technical data are: maximum shaft power 134 kW, speed 585 r.p.m., cooling water consumption 20 m^3/h, weight about 6000 kg.

The Stirling cycle in brief

We consider the schematic cycle as taking place in a gas-filled cylinder fitted with two pistons, as shown in *fig. 3a*. A high temperature T_C prevails in the left-hand half of the cylinder, a low temperature T_E in the right-hand half. The cycle falls into four phases. Phase *I* occupies the interval between the two situations numbered *1* and *2*, when the left-hand piston is compressing the working gas at temperature T_C. The heat of compression is removed from this part of the cylinder so that the process is isothermal. During phase *II* the gas is transferred to the part of the cylinder at lower temperature T_E (situation *3*) passing through a regenerator *Reg* to which it gives up its heat. During phase *III* the gas expands at temperature T_E (situation *4*). The expansion process is again isothermal, as the resulting cold is removed from the cylinder. The cycle is completed by phase *IV*, during which the gas is returned to the left of the cylinder, at the same time taking up the heat stored in the regenerator. This brings the system back to situation *1*.

In the course of each cycle we see that the gas flows back and forth between two spaces at temperatures T_C and T_E. In the diagram at fig. 3*b*, the schematic cycle is represented by two isotherms *I* and *III* and two isochores *II* and *IV* (isobars could be used instead of isochores [2]). When the cycle is to be applied for refrigeration, the temperature T_C of the warm space is kept close to room temperature and the output of the machine will be the cold produced during phase *III*.

We shall not digress here on the essential importance

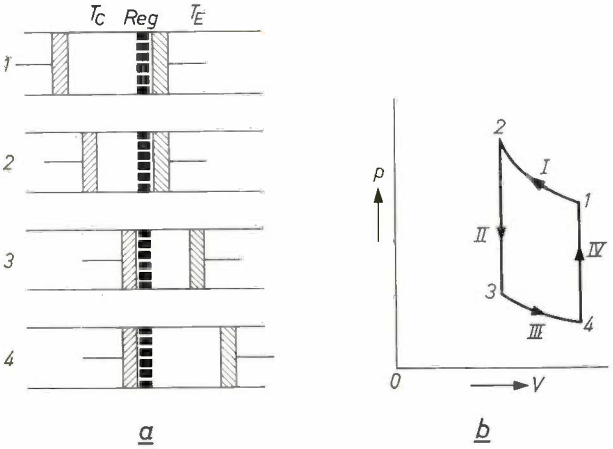


Fig. 3. Illustrates the brief account in the text of the schematic Stirling cycle as it takes place in a gas refrigerating machine.
a) Piston positions between the four phases of the cycle.
b) The relevant *p*, *V* diagram which consists of two isotherms *I* and *III* and two isochores (lines of constant volume) *II* and *IV*. (Isobars could be used instead of isochores.)

of the regenerator in the efficiency of the cycle. We should however note that care must be taken in the design to ensure that the gas flowing through the regenerator recovers from it as much as possible of the heat (or cold) it has previously given up to it. Quite small deficiencies in this recovery may be fatal to the efficiency of the machine.

It is not really practicable to build a system as described, with discontinuously moving pistons. However, there is no great loss of effectiveness if the pistons operate in simple harmonic motion. A phase difference between the movements of the two pistons is required such that the variation in volume of the expansion space is in advance (by 90°, say) of that of the compression space.

Practical Stirling machine design

Traditional approaches

There are many possible approaches to the design of machines making use of the Stirling cycle. Some of these will now be considered in the course of a step-by-step explanation of the construction of our machine.

Fig. 4 represents, in bare outline, a very old design of Stirling machine. In this the regenerator *Reg*, with heat exchangers fitted on each side, lies between the compression piston *C* and the expansion piston *E*. The one on the left is the *cooler*: water circulates through this, removing the heat generated by compression of the working agent. The cold resulting from the expansion is collected by the *cold-exchanger* (or "freezer"), on the right of the regenerator. The two pistons are driven through connecting rods and pivoted levers by a common crankshaft whose cranks are offset by 90°.

Although the thermodynamic part of the machine — the gas space, pistons and heat-exchangers — could hardly be simplified further, the transmission with its many jointed members is only barely acceptable from the constructional viewpoint. Here there is yet another appreciable disadvantage. If a machine based on the Stirling principle is to produce a reasonable amount of cold per litre of swept cylinder volume, then working agent pressures must be high [1], varying in the course of a cycle from say 25 to 50 atmospheres. This means that the pistons are always subject to a strong force tending to push them apart. The resulting load on the driving members must be balanced as far as possible and this could be achieved by enclosing the drive in a crankcase filled with gas at the mean pressure of the working agent. Little imagination is needed to recognize the difficulties in building a crankcase to hold a transmission like that in fig. 4, and which will withstand a pressure of (say) 35 atm with no more than very slight deformation (so that, amongst other things, the bearings do not become misaligned).

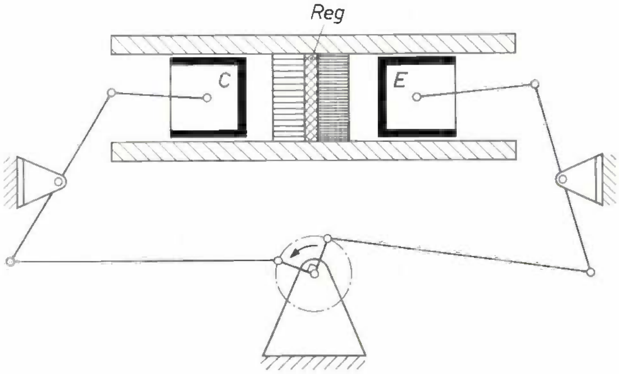


Fig. 4. One of the earliest designs proposed for a machine making use of the Stirling cycle. The compression piston *C* and the expansion piston *E* are actuated through a rod-and-lever system by a crankshaft whose two cranks are offset by 90°. The working space is between the pistons, and it is partly occupied by a regenerator *Reg* sandwiched between heat-exchangers.

The arrangement shown in *fig. 5a* looks more promising. It has the Vee configuration met with in many machines, for example compressors. The drive is extremely simple, and taking up the thrust of the working agent is less of a problem. As the crankcase is cylindrical and relatively small, it may easily be dimensioned to stand up to high pressure; alternatively (preferable in a big machine) a crosshead construction could be adopted for the transmission, and gas-filled buffer

[1] For more detailed treatment of the Stirling cycle, as made use of in refrigerating equipment, see J. W. L. Köhler and C. O. Jonkers, Philips tech. Rev. **16**, 69-78, 1954/55, and J. W. L. Köhler, Progress in Cryogenics **2**, 41-67, 1960.

[2] See the second article quoted in footnote [1].

spaces could be provided behind the two pistons, as in fig. 5*b*. However, from the thermodynamic point of view the designs of fig. 5 are inferior to that of fig. 4 in two respects. In the first place the distance between pistons *C* and *E* has been greatly increased, so that the dead space has been considerably enlarged. Secondly, the complete working space no longer possesses rotational symmetry. There is now a danger that the gas flowing back and forth through the regenerator will be unevenly distributed over the cross-section, with adverse effects on the efficiency of recovery of the stored heat or cold.

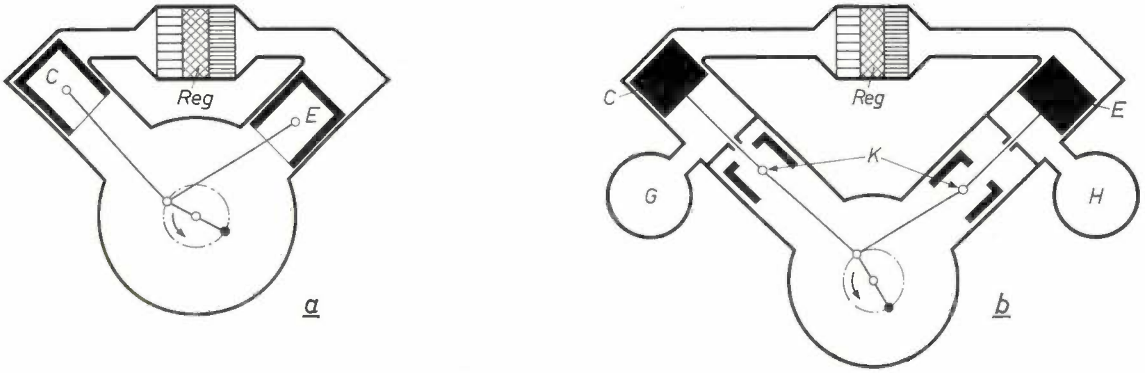


Fig. 5. (*a*) Stirling machine with a Vee drive.
(*b*) Buffer spaces *G* and *H*, filled with gas at the mean pressure of the working agent, have here been provided to take up the thrust on the pistons. The Vee drive is equipped with crossheads *K*.
This layout of working space and pistons is less favourable thermodynamically than the one shown in fig. 4.

The familiar arrangement embodying a displacer, which offered a good approach to the design of the *small* gas refrigerating machines, is again not so favourable as that of fig. 4. This will be dealt with below, on p. 306.

A new principle: hydraulic piston drive

Carrying on from the crosshead version of the Vee configuration, we came upon an idea combining the mechanical simplicity of this drive with the ideally shaped thermodynamic working space of the one we first of all considered. The principle is shown in *fig. 6*: pistons *C* and *E* are not driven by mechanical linkages, but *hydraulically, by means of oil columns*. The oil columns are driven by the plungers *D* and *F* mounted on the piston-rods of the crosshead drive. These oil columns, imprisoned between piston and plunger, can be regarded as flexible piston-rods. The non-hydraulic part of the transmission is the same as in fig. 5*b*, and the working space is the same as in fig. 4.

Elaboration of this idea led us to take two further steps. First, we saw there was a neat way of eliminating the two buffer spaces. The gas in these serves to exert a counter-pressure on the undersides of the plungers *D* and *F*, to relieve as much as possible of the downward thrust on the drive. The counter-pressure can however equally well be supplied by the working agent in a *second cylinder* in which the same cycle is taking place with a 180° phase shift. We arrived thus at the arrangement shown in *fig. 7*.

The final step was the removal of an imperfection still present in the arrangement shown in fig. 7. This has two cylinders and four pistons actuated by a drive with two *double-acting plungers*, each of which

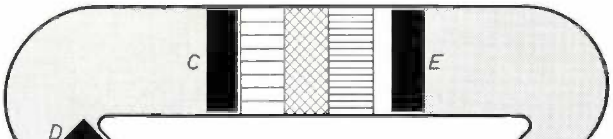


Fig. 6. A Stirling machine equipped with "hydraulic piston rods". The pistons *C* and *E* are located as in fig. 4, and are driven through oil columns by means of the plungers *D* and *F*. The plungers are actuated by a mechanism exactly the same as in fig. 5*b*.
Generally speaking, for optimum functioning it is necessary to arrange that the compression piston *C* has a longer stroke than the expansion piston *E*. If the two plunger strokes are equal, this can easily be achieved by giving the compression plunger *D* a rather larger diameter than the expansion plunger *F*.

drives two oil columns: one from the lower face and one from the upper face. A quick check on the phase relationships between the various volume variations (see the graph in fig. 7) makes it clear that the two compression spaces must lie on the same side. That is to say, plunger D has to drive two compression pistons while plunger F takes up the power delivered by two expansion pistons. The two arms of the Vee drive are therefore very unequally loaded, for both power and operating forces. If, for example, the crankshaft were delivering mechanical power at the rate of 100 kW, the compression arm would have to transmit about 120 kW, only about 20 kW being returned through the expansion arm, resulting in a rather low load on this arm. To remedy this defect we adopted the arrangement shown in *fig. 8*, which has *four working cylinders* instead of two; the layout is essentially the same as that of the new machine now in production. The two piston-rods in the arms of the Vee drive now each carry two double-acting plungers in tandem; one for compression and the other for expansion. In the case just considered, the two arms are now equally loaded with 50 kW each.

Layout of machine

As may be seen in fig. 8 (and in the photograph fig. 2) there are two cylinders at the front of the machine and two at the back. The two pairs are mounted with opposite slopes. This layout has been arrived at in order to keep both arms of the Vee as similar as possible: in each arm the compression plunger is the lower one and the expansion plunger the upper one. As it was desirable to keep the oil connections short, the compression pistons had to be close to the upper plungers and the expansion pistons had to be close to the lower ones, so that the sloping position of the cylinders followed automatically. The inclination of the cylinders has the further important advantage that it assists drainage of the liquid air or other condensate that collects on the pipes in the cold-exchanger. The film that forms on these pipes is therefore thinner, so that the transfer of cold is more efficient than in a strictly horizontal array.

In a closer examination of fig. 8 the reader will notice that the oil connections on the expansion side of the two rear cylinders are crossed, whereas the correspond-

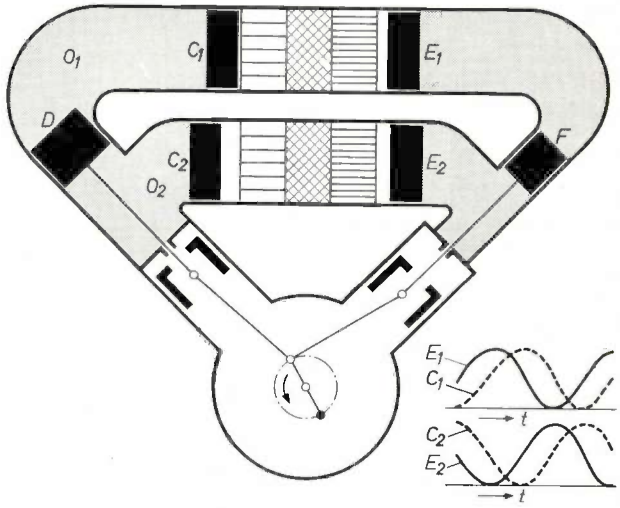


Fig. 7. The design of fig. 6 with the addition of a second working cylinder. The two compression pistons C_1 and C_2 are driven by oil columns O_1 and O_2, to which motion is communicated by the upper and lower faces of a *double-acting* plunger D. The expansion pistons E_1 and E_2 are driven in a similar way. This does away with the need for buffer spaces.

The displacements of the four cylinders have been plotted as a function of time in the inset. To obtain the correct phase difference between compression piston and expansion piston it is necessary that the two compression pistons should lie on the same side of the machine. The cycle taking place in one cylinder is in opposite phase to that in the other.

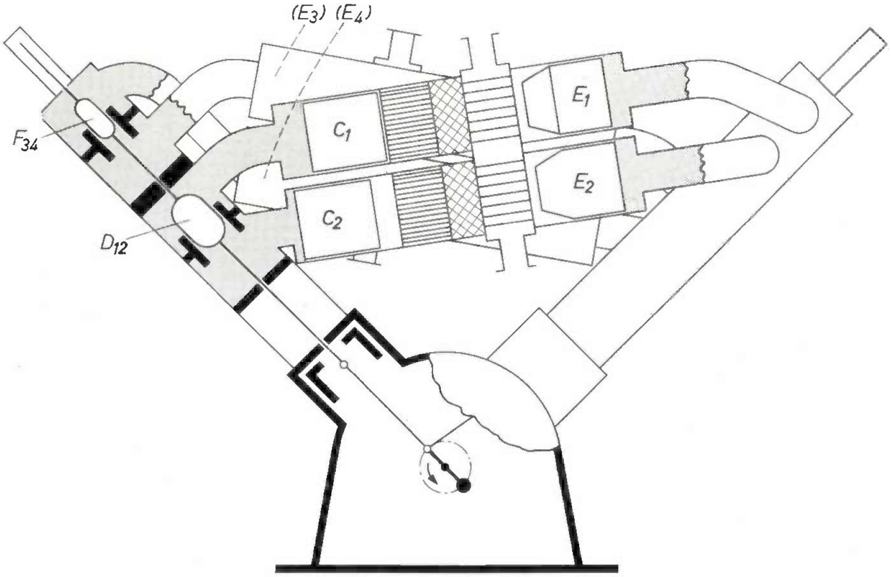


Fig. 8. Array comprising two of the cylinder pairs shown in fig. 7, one pair being placed at the front and the other at the rear of the machine. Each arm of the Vee transmission contains a compression plunger that drives two compression pistons and an expansion plunger that drives two expansion pistons. In the arm shown in longitudinal section, the compression plunger D_{12} drives C_1 and C_2, and the expansion plunger F_{34} drives the two expansion pistons of the rear cylinder pair (E_3 and E_4, not visible in the drawing).

In this array the two arms of the Vee transmission are equally loaded and exactly similar in design. This is the layout used for the new gas refrigerating machine (type C).

ing connections to the front cylinders are not crossed. The explanation is as follows. In the front pair of cylinders, whose compression side lies to the left, the correct phase difference between expansion and compression is obtained when the crankshaft rotates in the sense indicated in the figure, piston E then being 90° in advance of piston C. In the rear pair the expansion pistons are on the left, and here the same sense of rotation of the crankshaft would evidently give wrong phasing without the crossover in the oil connections. Because there is a 180° difference in phase between the upper and lower cylinders of each pair, the crossover re-establishes for both cylinders the correct phase difference between compression and expansion pistons.

Choice of operating speed

It is not intended in this article to deal at length with the choice of the various basic parameters of the design. The choice of operating speed does however call for some explanation because it was partly a rethinking of this question that led to the new directions of development referred to in the introduction.

If one sets out by specifying a certain refrigerating capacity per cylinder from a gas refrigerating machine operating under specified conditions, this determines the product speed × swept volume. When the Stirling cycle, a 19th century invention, was taken up again by workers in Philips Research Laboratories some twenty years ago, the starting point was the choice of speeds far greater than had been used in the old-fashioned hot-gas engines: this enabled reductions in the swept volume and hence in the bulk and weight of the machine. These developments were made feasible by advances in heat-exchanger and regenerator design. The theory of the Stirling cycle, which was then developed, had shown that compression ratios must not be too small if good efficiency was to be obtained. The inference was that not only the swept volume but the whole of the dead space, including heat-exchangers, would have to be reduced. If now the heat-exchangers, while occupying less space, were still to present a large enough superficial area for the efficient transfer of heat, then they would have to be honeycombed much more finely with ducts for the working agent. The essential point in the development was that structures were found which did not cause excessive flow losses, in spite of the proportionately higher operating speeds. Machines were successfully constructed for speeds of 1500 r.p.m. and higher. All the Stirling machines made by Philips since about 1950 used this speed.

For industrial gas refrigerating machines, that law of the Medes and Persians requiring that size must be reduced, no matter how, does not apply. "Smaller" has no significance for the industrial user unless it means "cheaper" and "more efficient" or, possibly, a more adequate compromise between these two attributes. In fact a compromise is relevant in the present case, as may be seen from the following simple argument. Let us first consider a small high-speed gas refrigerating machine. Now the same refrigerating capacity will be obtainable from a bigger and slower machine which, because of its larger swept volume, can safely have larger dead spaces. This means that the heat-exchangers can be simpler in design, i.e. have a coarser structure, without prejudicing the efficiency of the heat transfer process or the overall efficiency of the machine. Because of its simpler construction such a machine would be no more expensive to make than the small one, even though it would require more material. From here, however, one could push the performance of a large machine to the utmost by using more finely-divided heat-exchangers as in the small machine. The cost price would then be rather higher, but the efficiency would be better than that of the small machine.

Where such a compromise has to be found for an industrial machine, efficiency carries rather more weight than a saving on purchase price. Furthermore, the description "for industrial applications" is enough in itself to suggest a fairly heavily constructed and perhaps a fairly large machine running at a moderate speed suitable for good mechanical efficiency and long life.

The problem of operating speed was examined afresh in the light of the above considerations. Optimization calculations were carried out by computer, account being taken of a whole series of factors, including the thermal capacity of the regenerators — a critical point in the thermodynamic behaviour of cold-producing machines. The results indicated that larger gas refrigerating machines could with advantage be run at lower speeds than small ones. It was decided on these grounds that the new machine should run at about 600 r.p.m. This is the usual operating speed for compressors of comparable shaft power, so we were able to base our machine on a Vee drive of a commercially available type.

The main features of the machine will now be discussed in some detail.

Constructional features

Hydraulic system

The oil column, enclosed between each working piston and the appropriate plunger, can be regarded as a rod capable of being bent into any desired shape. This "flexible piston rod" not only allows the designer a great deal of freedom in the positioning of components, but it has further excellent properties. It will transmit strong forces almost frictionlessly, the hydraulic losses (flow, leakage and fluid friction losses) being reducible to about 5 % of the shaft horsepower. It does away with alignment problems. The ratio between the plunger and piston displacements can be varied within wide limits by appropriate choice of diameters.

Needless to say, some extra provisions are necessary: as in all hydraulic systems, the fluid must be filtered, cooled and degassed. Apart from this the "flexible piston rod" presents only one real problem, that of the seal between the hydraulic fluid and the working agent. It is of the highest importance that the gas space should remain clean and free from oil; moreover, any leakage of gas from the cylinder would greatly prejudice the efficiency (and other features) of the machine.

During the early stages of work on the hydraulic system, in which the sealing problem was particularly intractable, it happened by a fortunate coincidence that a new kind of *positive seal* was developed in the Research Laboratories, and this proved to be the answer to the difficulty. This seal consists of a rolling diaphragm supported on an oil cushion, and is fully described elsewhere in this number of the Review [3]. Requirements for satisfactory operation of this seal were found to fit in very well with those for the hydraulic drive. The next section, dealing with the oil/gas seal arrangements, will illustrate this. For more details about the principle of the rolling diaphragm seal and its various practical forms, see the previously quoted article [3].

Rolling diaphragm seal

The seal is shown schematically in *fig. 9*. The rolling diaphragm is made of special polyurethane rubber with no reinforcement. The outer perimeter of the diaphragm is attached to the cylinder-wall, and its inner edge to the piston, so that it is rolled alternately off the two surfaces. It thus seals off the gas space hermetically. The pressure of the gas above the piston forces the diaphragm against a ring-shaped oil cushion, in which the pressure is a few atmospheres lower. The sliding fit between piston and cylinder-wall lower down (*S*) acts as seal for the oil forming the cushion.

For long life it is essential that the diaphragm keeps a constant length throughout its "rolling" cycle: this is equivalent to the maintenance of constant extension, or therefore of constant pressure difference across the diaphragm. This condition will be satisfied if the *volume* of the oil cushion remains the same throughout the cycle. This is achieved by using a stepped oil space, as in fig. 9 (see the article quoted [3] p. 289). Over a longer time, the amount of oil in the cushion is liable to change slowly as a result of leakage (in or out) through the sliding fit, and this would cause the diaphragm tension to change. All rolling diaphragm seals must therefore also be equipped with a control device that directly maintains a constant difference of pressure between gas space and cushion by regulating the quantity of oil in the cushion.

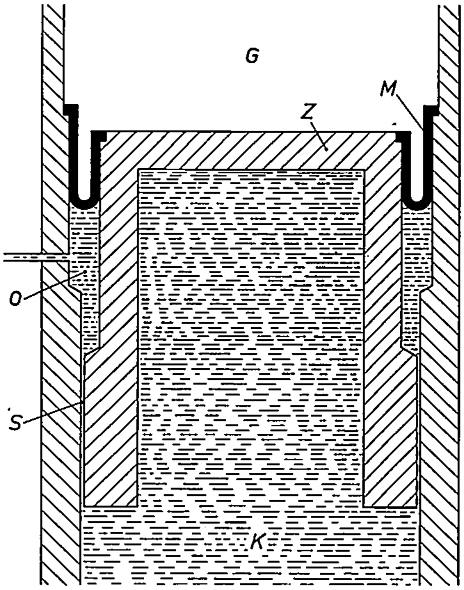


Fig. 9. In each cylinder, the gas *G* is sealed off hermetically from the oil column *K* behind each piston *Z* by a rolling diaphragm *M* supported on an oil cushion *O* (see the article quoted [3]). The pressure difference between *O* and *G* has to be kept constant.

The oil cushion must further be kept as close to room temperature as possible, because the polyurethane rubber tends to lose its excellent properties when heated. Nor must the oil contain any gas, since this would make it too compressible; and as there is always some diffusion of the working gas through the diaphragms, the oil has to be circulated and degassed. Other impurities are removed by filtration.

The cushioning fluid thus calls for the self-same treatment as that in the main hydraulic system — cooling, degassing and filtration.

Obviously, the decision to base the new gas refrigerating machine on the application of rolling diaphragm seals was not taken until it was firmly established that these seals had a long enough life — a matter on which, at the outset, many had their doubts. Life-tests in the Research Laboratories showed that suitably dimensioned seals could reach lives of more than 10 000 hours at 1500 r.p.m., and since the speed of our machine is only 600 r.p.m. we may reasonably expect even longer lives. Again, the larger the diaphragm, the greater its endurance, and we have been able to give the diaphragm

[3] J. A. Rietdijk, H. C. J. van Beukering, H. H. M. van der Aa and R. J. Meijer, Philips tech. Rev. 26, 287-296, 1965.

in our machine the relatively large mean diameter of 175 mm. It has also been possible, in our special case, to make use of a particularly simple and efficient form of the pressure control arrangement referred to above. This will be dealt with shortly.

A prototype of the new machine has undergone extensive proving trials, and the results fully justify our confidence in the rolling diaphragm. No seal failure occurred during a continuous test under full load lasting more than 4000 hours; and when the machine was subsequently dismantled, the rolling diaphragms showed hardly any signs of wear or of ageing in general, and indeed, this was true of all the other parts of the machine.

The rolling diaphragm does not of course have an unlimited life. However, the construction of the machine is such that replacement of a leaky membrane need take little longer than changing a flat type. To minimize the risk of even this short break in service it is recommended that all rolling diaphragm seals be renewed in the course of annual overhaul.

The combination of seal and hydraulic system examined in more detail

Fig. 10 is a simplified diagram showing part of the overall hydraulic system relevant to the operation of a single working cylinder. The piston *Z* is actuated by the double-acting plunger *D* and the oil column *K*. The piston is hollow to take the column of oil. The rolling diaphragm *M*, which seals off the gas space *G* hermetically, rests upon a ring-shaped cushion of oil *O* for which the narrow gap *S* acts as seal. The sliding fits *S* and *T* form the piston guides.

The volume swept by the plunger determines the piston stroke. The length of the "flexible piston rod" is however itself liable to vary: a certain amount of leakage is inevitable at various points in the hydraulic system; and moreover, as has already been pointed out, freshly filtered oil must circulate through the system at all times. Provision has therefore to be made for maintaining the length of the oil column at the correct value, and for this the openings *i* and *u* in the cylinder wall are provided at the sliding fit *T*. If the oil column is too short the opening *i* is clear when the piston is at the *start* of its stroke; the pump *P* can then add oil to the column from the reservoir *Res*. If the oil column is too long the opening *u* is clear when the piston is at the *end* of its stroke, and excess oil can then escape, returning to the reservoir. This simple expedient for regulating the length of the column works excellently — except of course when the machine is at a standstill. We shall come back to this shortly.

As already observed, the extension of the rolling diaphragm, distended against the oil cushion by the gas, must be constant, and for this a regulating system is necessary to keep the difference of pressure between gas space and oil cushion constant. A difference of about 4 atm is desirable. The article quoted [3] describes a simple control arrangement (see fig. 5 in that article) in which a miniature pump feeds oil to the cushion and a valve regulates its rate of flow; the valve is controlled by a spring-loaded flexible diaphragm sensing the gas pressure on one side, and the oil cushion pressure on the other. Our machine makes use of a similar arrangement, but there was no need for any direct connection between the valve and the gas space. In fact the pressure of the oil inside the piston is almost the same as that of the working agent (there is only a small difference due to friction and the momentum of the piston). The valve is therefore made to sense the oil pressure in the column, and not the gas pressure. The resulting cushion control device *R* and its connections are shown in fig. 10. Instead of a flexible diaphragm it contains a spring-loaded plunger *H*. The force exerted by the compressed spring is such that the port *p* opens as soon as the oil pressure in the cushion rises above the desired value, here 4 atm less than the oil pressure in the column. The excess oil then escapes, returning to the reservoir *Res*, and the pressure in the cushion falls. If this should become too low, the plunger blocks the port *p*, and the pressure in the oil cushion increases again because of the continuous leakage past the clearance *S* of oil from the column, which is at a higher pressure.

The slow loss of oil from the column is compensated by the control system described above that corrects its length. Both the fluid in the main hydraulic system and that in the seal-supporting cushion are thus continually being changed; and cooling, filtering and degassing of the oil can take place at the same time. It may fairly be said that the combination of rolling diaphragm seal and hydraulic piston rod make an excellently integrated device.

The electrically driven oil pump *P* runs continuously, and is not switched off when the machine is stopped temporarily. However, the column length correction then ceases, and oil begins to drain away from each column. All eight pistons gradually slip back to the start of their stroke, i.e. the position at which the inlet *i* is clear; since at that point the pump starts to feed in oil. This means that during a stoppage all the column lengths go out of adjustment. Surprising as it may seem, this does not matter in the least; on restarting, the correction system restores each column to its proper length within a few strokes. It is however obvious that an abrupt start-up is not permissible as the pistons would be driven hard against their stops. The procedure is therefore as follows. Mounted on the shaft of the electric motor powering the installation is a starting handle, and the machine is slowly turned over by this means. During the first turn one can definitely feel the pistons run against their stops; but during the next they are already running smoothly. The design of the electrical circuit is such as to prevent the motor being switched on until this is so. Another built-in safeguard defers starting until the working agent in the cylinders has built up to a pressure at which the seals are at the correct tension. This ensures at the same time that the oil columns are under compression throughout the stroke; a tension exerted on the column would induce cavitation in the oil, and the "flexible piston rods" would "snap".

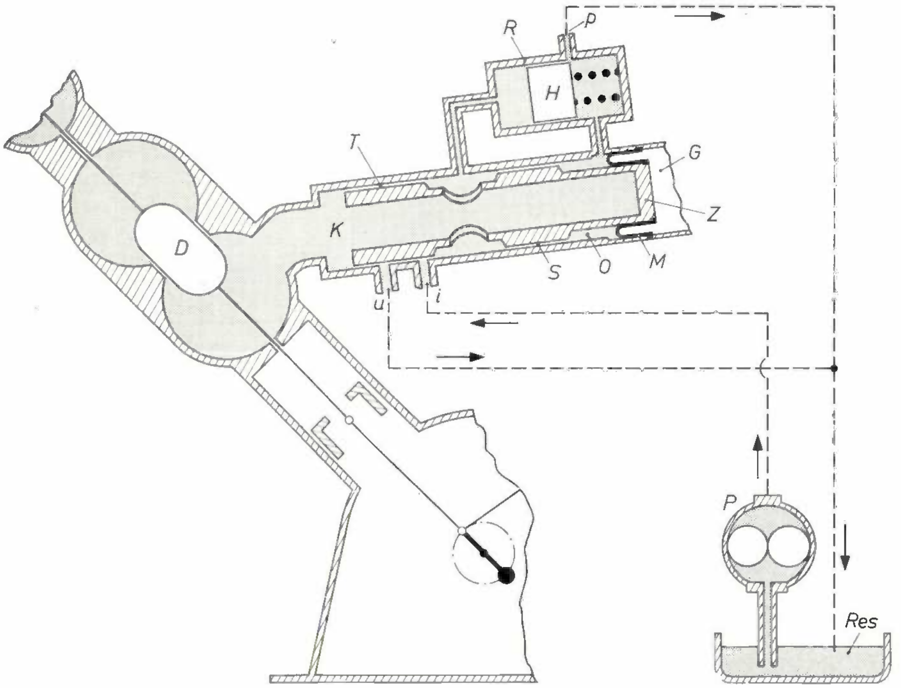


Fig. 10. Systems for controlling the "length" of the hydraulic piston rod that moves the piston Z, and the oil pressure in the cushion support for the rolling diaphragm M. P oil pump, *Res* oil reservoir. The length of the oil column K is adjusted by feeding oil through the inlet i and removing it through the outlet u. Oil is continuously forced into the cushion space O through the small gap S; the excess is removed through the port p in the control device, the rate of loss being governed by the spring-loaded plunger H in such a way as to maintain a constant pressure difference across the rolling diaphragm.

In the event of an unexpected stoppage of the machine and oil pump, due for example to an electricity supply failure, an automatic safety device comes into action and blows the gas from all four cylinders. If this were not done the eight pistons would slip back beyond the normal start position referred to above, and the rolling diaphragms might be destroyed by the gas pressure in the cylinders (30 atm on average).

At the outset, in view of the unusual design of the machine throughout, there was uncertainty on a number of points. For instance, it was quite conceivable that fluid friction in the rapidly reciprocating oil columns might be many times greater than that associated with steady flow of the oil at comparable displacement rates. No useful experience on this was available to us, and the literature as known to us was equally unhelpful. Luckily, the losses in question proved to be quite acceptable; measurements made with the thermodynamic section of the machine out of action revealed that they did not account for more than about 2 kW in all, at an overall shaft horsepower equivalent to 134 kW, and thus were only slightly more than would be expected for steady flow. Other results of measurements on the machine will be given in the final section of this article.

The heat-exchangers

Fig. 11a shows the layout of heat-exchangers in any of the four cylinders; the arrangement is for each cylinder essentially the same as that of fig. 4 (the two-piston arrangement), which, it will be remembered, served as starting point for the design of the new machine.

The re-orientation of our design thinking referred to in the introduction entailed the adoption of this two-piston arrangement and the abandonment — at least

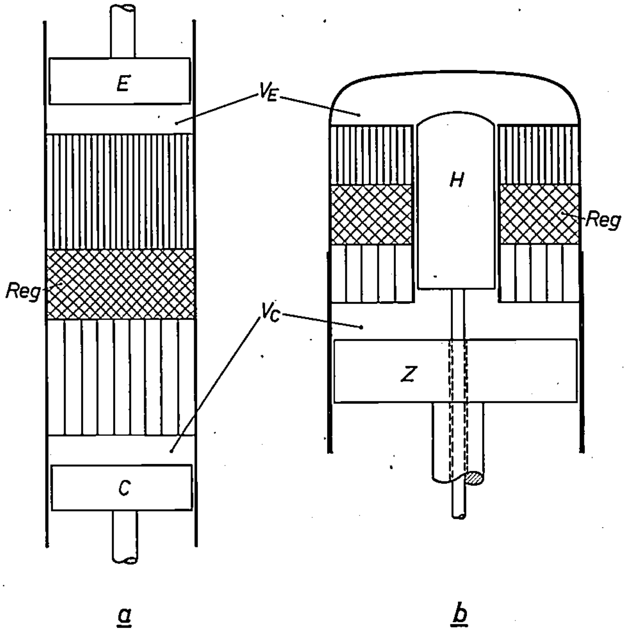


Fig. 11. The disposition of the heat-exchangers in the new machine (*a*) compared with that in a machine working on the displacer principle (*b*). In the latter, Z is the piston and H the displacer. V_C denotes the compression and V_E the expansion space in both machines.

for the large industrial machine — of the *displacer* principle which had hitherto been employed in all Philips gas refrigerating machines. The reasons behind this important decision call for some further explananion.

The displacer arrangement used by Philips appears it fig. 11*b*. The only purpose of the displacer H is to move gas through the heat-exchangers from the compression space V_C to the expansion space V_E and back again. This means that there is only a very small pressure difference across the displacer, due exclusively to the resistance encountered by the gas flow. Since any leakage of cooled gas from the *expansion space* has very adverse effects on cold production and the efficiency of the machine, particular attention has to be given to the sealing arrangements for this space. The types of seal available earlier were not "positive" (they were called "leakage limiters" in the article quoted [3]) but they gave less difficulty in the arrangement of fig. 11*b*, with its small pressure difference, than they would have given in the arrangement of fig. 11*a*, in which there is an alternating pressure difference of high amplitude across the expansion piston E. This was the consideration that tipped the scales in favour of the displacer.

With the advent of the rolling diaphragm, which effectively disposed of the seal problem, the two-piston arrangement as in fig. 11*a* became the better proposition, for large machines at least. It is thermodynamically more favourable, because the gas flow is perfectly evenly distributed over the cross-section of the heat-exchangers. Also, the stresses arising in the cylinder wall are less complicated and more easily overcome.

One or two details of the heat-exchangers now follow. The heat-exchangers can be clearly seen in *fig. 12*, which is a simplified drawing of the complete machine.

The *cooler* consists essentially of an array of tubes opening at either end into a circular plate, and surrounded by a waterjacket. The gas flows through the tubes and the cooling water circulates around them. The tubes are arranged in a pattern chosen so that the gas flowing out distributes itself as evenly through the regenerator as possible. To improve the heat transfer the cooling water flows through interleaved coaxial cylindrical baffles arranged between the tubes. The water passes first through one and then through the other cooler of a cylinder pair.

The *regenerator* is a stack of sintered copper gauze discs in a thin-walled collar made of 18/8 chrome-nickel steel. The copper is very evenly distributed throughout the enclosed space, the gap between the gauze and the cylinder has been made as small as possible, and the end-faces are so designed that the gas flows evenly from the fine structure of the regenerator into the tubes of the cooler and the cold-exchanger.

The *cold-exchanger* is also made up of copper tubes, end-plates and an enclosing jacket; and it is fitted with two perforated partitions. The gas flows through the tubes and the useful cold is given up to their outside surfaces. Each side of the jacket is brazed to an 18/8 chrome-nickel steel cylinder; one is the cylinder enclosing the regenerator and the other is the cylinder in which moves the head of the expansion piston, made of the same steel. These steel parts act as thermal resistances to prevent conduction of valuable cold to warmer regions of the machine.

The cold-exchangers are chiefly constructed for applications such as gas condensation (gases such as methane, oxygen, air, nitrogen etc.). Condensation is possible at pressures up to 30 atm. The cold-exchangers of the two cylinders in a pair are connected in series.

Other constructional details

The way in which the working cylinders are mounted on the machine can be seen in fig. 12. The main mechanical support for the heat-exchanger assemblies of paired cylinders is provided by an insulating jacket enclosing both cold-exchangers. The expansion and compression cylinders and the interposed heat-exchangers constitute a single cylinder whose inside diameter is practically the same throughout its length. This means that the heat-exchangers cannot be subjected to any strong forces due to gas pressure or thermal expansion or faulty assembly. The long cylinder is closed at either end by a cover plate which is, of course, acted upon by axial forces. As the forces on the two plates are roughly equal and opposite, they can be taken up by two tie-rods between the plates. The cylinder itself is not affected by these axial forces because the upper cover is made so that it is a sliding fit in the shell. The oil leaking through this sliding fit seal drains into the main reservoir, *Res* in fig. 10, whence it is fed back to the columns by the oil pump, in the manner already described.

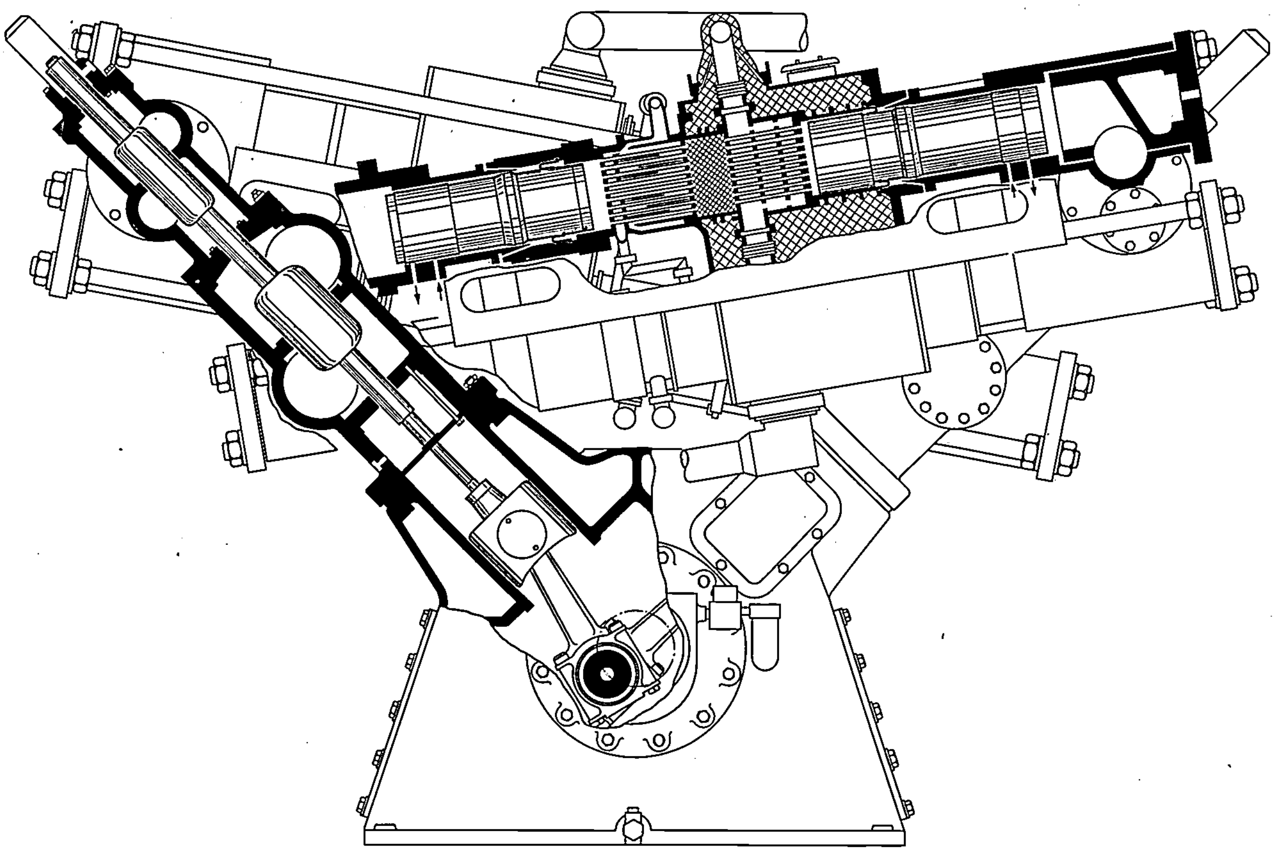


Fig. 12. Section of complete machine, showing constructional details of the working cylinders, heat-exchangers, etc. Note the tie-rods linking the cover plates at either end of each cylinder.

The frame of the *drive gear* complete with crankshaft, connecting rods and crossheads is a normal compressor frame and is obtained by us from a firm specializing in compressors. The crankshaft, which has a single crank, is supported in three roller bearings; it overhangs beyond the third bearing where it carries the rotor of the flange-mounted electric motor. The motor is also of a normal commercially available type. This construction gives very low mechanical losses throughout and the machine occupies very little floor space.

The two Vee-arms housing the chambers with double-acting plungers are mounted on to the body of the machine. We have already referred to the low fluid friction losses. Measurements on a separate Vee-arm have shown that the design chosen, with the plunger sliding in a short seal of a large-bore chamber, involves lower losses than with a conventional hydraulic piston and cylinder.

The machine has been designed for working agent pressure up to a maximum of 60 atm and shaft power up to 134 kW.

Results of measurements

In the course of proving trials on the prototype, exhaustive measurements of refrigerating capacity were carried out at different cold-side temperatures. Results obtained, using hydrogen and helium as working agents, have been plotted in *figs. 13* and *14* respectively; in both cases the machine was run under full load and cooled with water at 15 °C, supplied at 20 m^3/h. One curve represents the power taken up by the shaft, P_m, which does not reach its maximum of 134 kW until the temperature on the cold side has fallen to 110 °K; another the useful refrigerating power output P_E; and a third the efficiency, as calculated from these two quantities and expressed as a percentage of the Carnot efficiency η_C, the theoretical maximum at the temperature in question. It can be seen that remarkably high efficiencies are attained. In *Table I* the efficiency of the

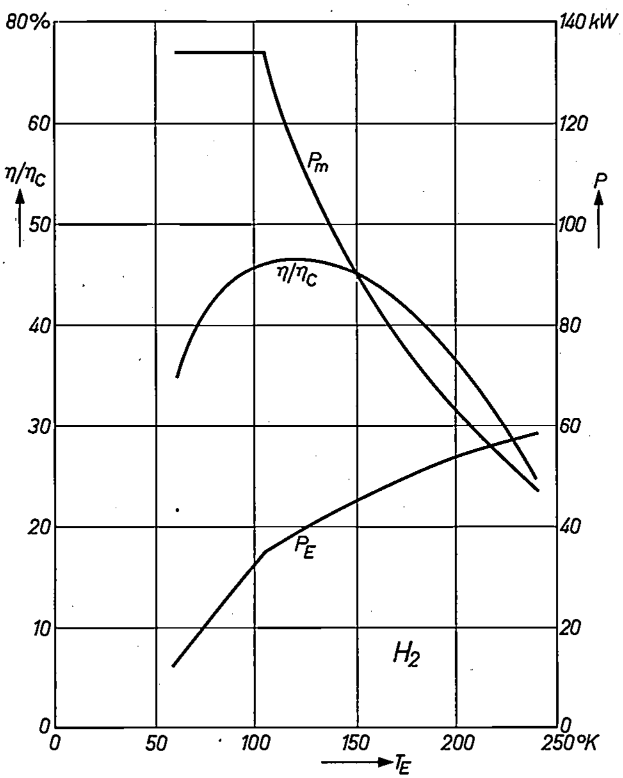


Fig. 13. Performance of the new machine, using *hydrogen* as working agent, and cooled with water at 15 °C at the rate of 20 m³/h. Measured values of refrigerating capacity P_E and shaft power P_m, and also the relative efficiency η/η_C calculated from these two quantities, are shown here as functions of the cold side temperature T_E. The machine was operating under maximum load: at temperatures above 110 °K the maximum working agent pressure had the highest permissible value of 60 atm, which limited the power transmitted by the shaft; at lower temperatures the pressure of the working agent was reduced so that the highest permissible shaft power of 134 kW should not be exceeded.

new machine at 77 °K (liquid nitrogen) is compared with that of the older four-cylinder model type B, for using He or H_2 as working agent. In both cases there is some 30% improvement.

The measurements were at the end of the 4000-hour continuous run already referred to, and the efficiency and refrigerating capacity were found to be unaltered.

Table I. The efficiency η/η_C of the new machine (type C) compared with that of type B. All efficiency figures relate to operation under maximum load at a cold-side temperature of 77 °K and a cooling-water temperature of 15 °C.

	Working agent He	Working agent H_2
machine type B	27.6%	29.4%
machine type C	41.6%	41.4%[*]

[*] The higher efficiency of the new machine when working with helium is attributable to the fact that at 77 °K helium deviates less from ideal gas behaviour than does hydrogen. In the less fully optimized design of the type B machine this advantage is cancelled out by fluid flow losses, which are greater for helium than for hydrogen. — At higher temperatures, at 90 °K for example (boiling point of oxygen), the advantageous property of helium is not yet perceptible in the new machine; as may be seen from figs 13 and 14, the efficiency of type C at this temperature is 42.0% with helium as against 44.2% with hydrogen as working agent.

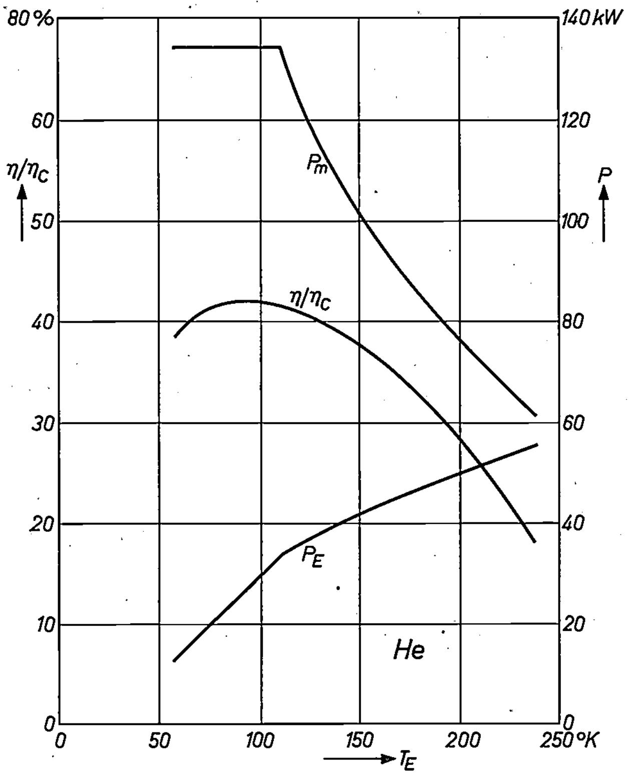


Fig. 14. Performance of the new machine using *helium* as working agent; details as in fig. 13.

The machine described above was ready for production within three years from the start of the development. This was made possible by the concerted effort of an enthusiastic team of workers who cannot all be named here. However, special mention should be made of the important contribution by Ir. H. J. Verbeek.

Summary. An industrial gas refrigerating machine has to be more efficient and reliable and cheaper to maintain than similar equipment for laboratory use. In addition, higher refrigerating capacities per unit are required. A bigger machine, type C, embodying entirely new constructional ideas, has therefore come to supplement the Philips A and B type machines employed in many laboratories. In order to employ the most favourable thermodynamic arrangement of the working spaces while doing away with the need for a complicated system of connecting and driving members which this would otherwise involve, the new machine has been equipped with *hydraulic* piston-drive, the pistons being driven by oil columns which are actuated in their turn by plungers linked to a conventional crosshead Vee drive. The operating speed is about 600 r.p.m.; this is appreciably less than that of the smaller types, which are based on a speed of 1500 r.p.m. The gas spaces are sealed off from the oil columns by means of *rolling diaphragms* which have proved to be admirably suited to use in conjunction with the hydraulic drive. The length of the oil columns and the pressure across these sealing diaphragms are kept constant by simple control systems. The machine has four working cylinders, whose eight pistons are served by four double-acting plungers; by this arrangement the need for buffer spaces to balance the mean pressure in the gas spaces has been avoided. The refrigerating capacity at 77 °K (with cooling water at 15 °C) is 20 kW for a shaft power of 134 kW. This corresponds to an efficiency relative to the Carnot cycle of more than 41%, as compared with about 30% for the type B machine. Hardly any sign of wear was visible in rolling diaphragms or other parts of the new machine after a trial in which it was run continuously for 4000 hours.

The harp cathode, a cathode with low thermal inertia for small transmitting valves

H. G. Gerlach

621.385.1.032.213.1

In mobile transmitter-receiver equipment the receiver is in use for the greater part of the time. A considerable saving of energy can be achieved if the transmitter is completely switched off when not required. In order, however, to answer a call promptly enough in such a situation, the valves used in the transmitter must be capable of almost immediate operation after being switched on. With this in view a cathode has been developed which reaches its working temperature in an extremely short time.

Introduction

In the last twenty years there has been a marked increase in the use of mobile transmitting and receiving units. This trend has been stimulated by the constantly growing need for such communications for police and military purposes, medical services, etc. Technical advances have also created new applications for mobile equipment. For example, it has become possible to use higher frequencies, so that a greater number of communication channels can be in use at the same time; moreover, transistors can now be used in receivers and to some extent in transmitters as well, so that the power supply for mobile units presents fewer problems.

In many cases a listening watch takes up the greater part of the time in the operation of mobile communications equipment. The transmitter is switched on only very occasionally, but it must always be standing by for immediate use. This is no problem if the transmitter is fully transistorized. At present, however, valves are still needed in the last stages of most mobile VHF units. This is particularly the case if an output power of at least a few watts is required at a frequency of a few hundred megacycles. Since the valves suitable for this purpose usually have an indirectly heated cathode, requiring a warming-up period of about 20 seconds, the filaments have to remain switched on the whole time. This means an extra consumption of energy, of about 10 to 20 W in normal transceiving equipment. Since a receiver fitted with valves consumes 15 to 30 W, and a set with transistors as little as 1 to 2 W, a relatively substantial saving of energy can be achieved if the valve filaments in the transmitter section are switched off when the transmitter is not in use. A saving of this nature is important for mobile units, as they are generally powered from dry cells or an accumulator.

What exactly is meant by the statement that the transmitter must be ready "for immediate use"? In practice it has been found that the acceptable delay depends on the manner in which the transmitter is switched on. If this is done by lifting a microphone from its hook, then the transmitter must be ready for use within 1 second. If the transmitter is switched on by pressing a button, the maximum delay should be 0.5 second. A third possibility is for the transmitter to be switched on by the sound of the operator's voice, and in that case no more than 0.1 second should elapse between the moment of switching on and the moment at which the transmitter is put into operation. The warming-up period of the cathodes in the transmitting valves should therefore be no longer than the times mentioned. The term "warming-up period" used here will refer to the time elapsing between the moment of switching on and the moment at which the power supplied to the aerial amounts to half the final value.

In the following we shall first discuss some theoretical considerations relating to the warming-up of cathodes. We shall then describe the harp cathode, developed by Philips. The warming-up period of this cathode is 0.35 second, and special measures can be taken to reduce it to as little as 0.1 second.

The warming-up of a cathode

We shall first calculate the warming-up period of a cathode in the form of a homogeneous rod or wire for the case where a constant electric power P (watts) is supplied. For this purpose we put the heat capacity of the cathode at K (gcal/°K). Further, T (°K) is the temperature, which varies with the time t (in seconds) and T_e is the final temperature.

The power available for warming-up the cathode

Ir. H. G. Gerlach is on the staff of the Philips Elcoma (Electron Tubes) Division, Eindhoven.

is the difference between the power supplied to it and the power it radiates [1] at the temperature to which it is heated. Since the latter power is proportional to the fourth power of the temperature, and since, when the final temperature is reached, all of the power supplied is radiated, we have the following equation:

$$K\frac{\mathrm{d}T}{\mathrm{d}t} = 0.24\,P\left[1-\left(\frac{T}{T_e}\right)^4\right], \quad . \; . \; . \; . \quad (1)$$

After introducing the quantities

$$A = \frac{KT_e}{0.24\,P} \text{ and } y = \frac{T}{T_e},$$

we can write (1) as:

$$A\frac{\mathrm{d}y}{\mathrm{d}t} = 1-y^4 \,. \quad . \; . \; . \; . \; . \; . \; . \; . \; . \; . \quad (2)$$

Integration of this equation gives:

$$\frac{t}{A} = \int\frac{\mathrm{d}y}{1-y^4} = \tfrac{1}{2}\,\text{arc tan}\,y + \tfrac{1}{4}\log\frac{1+y}{1-y} + C, \quad (3)$$

where C is the integration constant. If we suppose that the cathode is warmed up starting from the absolute zero point, then $y = 0$ at $t = 0$ and therefore $C = 0$. The relation thus obtained between y and t/A is represented in *fig. 1*. It can be seen that at low temperatures, at which the radiation is still insignificant, the temperature increases linearly with time; and that beyond the linear region the temperature approaches its final value asymptotically. If the warming-up does not begin at $T = 0$, the same curve applies, starting from the relevant y value.

We shall now confine our remarks to *directly heated* barium-oxide cathodes [2], which at the present state of development are the only ones suitable for our application. (Indirectly heated cathodes have far too high a heat capacity.) Putting the final temperature of these cathodes at 1000 °K, and assuming that the initial temperature is 300 °K, then the appropriate value of y is $y_1 = 0.3$. Experimentally it has been found that a transmitting valve with a barium-oxide cathode delivers 50% of the maximum power when the cathode temperature is 800 °K, that is at $y_2 = 0.8$. From fig. 1 the distance along the horizontal axis between the points corresponding to y_1 and y_2 is $0.89 - 0.3 = 0.59$ The warming-up period is therefore:

$$\tau = 0.59\,A = \frac{0.59\,KT_e}{0.24\,P} = 2460\,\frac{K}{P}. \quad . \; . \; . \quad (4)$$

In reality, the cathode differs from the theoretical one to which the differential equation (1) applies, since the heat is generated in the metal core and the barium-oxide layer has to be heated by conduction and radiation from this core. At conventional oxide-layer thicknesses, it has been found that this increases the warming-up period by about 10%, so that (4) should be written:

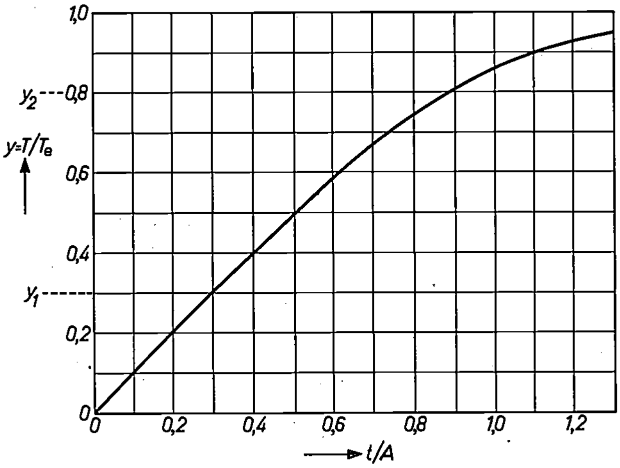


Fig. 1. Temperature curve of a body in the form of a homogeneous bar or wire to which a constant power is supplied. The power loss at the ends due to heat conduction is assumed to be negligible. The ratio y of the absolute temperature T to the final temperature T_e is plotted against the time t, in seconds, divided by the quantity $A = KT_e/0.24P$, where K is the heat capacity in gcal/°K and P is the supplied power in watts.

$$\tau = 2700\,\frac{K}{P}. \quad . \; . \; . \; . \; . \; . \quad (5)$$

Another fact that has not been taken into account above is that the supplied power P is not as a rule constant during warming-up. The resistance of the filament increases as the temperature rises, and since the filament voltage may generally be regarded as constant, the power supplied during the warming-up period is higher than that supplied in the final state. For two materials frequently used in directly heated cathodes, nickel and tungsten, the ratios of the resistance at working temperature to the resistance at room temperature are 5.4 and 5.9 respectively. The corresponding power increase immediately after switching on results in a considerable shortening of the warming-up period.

Where V is the r.m.s. value of the filament voltage and R is the resistance, then $P = V^2/R$, and (2) becomes:

$$\frac{\mathrm{d}y}{\mathrm{d}t} = \frac{0.24\,V^2}{KRT_e}\,(1-y^4), \quad . \; . \; . \; . \quad (6)$$

or:

$$t = \frac{KT_e}{0.24\,V^2}\int_{y_1}^{y_2}\frac{R\,\mathrm{d}y}{1-y^4}. \quad . \; . \; . \quad (7)$$

The warming-up period found from (5), which holds on the assumption that the resistance R is constant and equal to the final value R_e, must therefore be multiplied by a factor

$$F = \frac{\int_{y_1}^{y_2} \frac{R\,dy}{1-y^4}}{R_e \int_{y_1}^{y_2} \frac{dy}{1-y^4}}. \quad . \; . \; . \; . \; (8)$$

This factor can be determined graphically for various materials by using the known relation between resistivity and temperature. For nickel and tungsten at the temperature limits stated above, 300 and 800 °K, the values found for F are 0.6 and 0.5 respectively.

Strip cathodes and wire cathodes

We shall now consider two different forms of cathode. The first is made from a *strip*, of thickness, length and breadth d_1, l and b cm respectively, which is coated with a barium-oxide layer d_2 cm thick (*fig. 2a*). Let s_1 be the heat capacity of the core material and s_2 that of the barium-oxide (both in gcal/cm³); then for a cathode of this type:

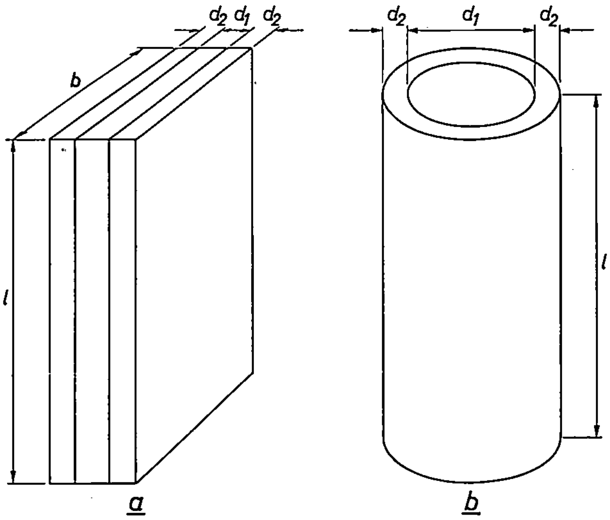


Fig. 2. Two forms of directly heated cathode.
a) Strip cathode of thickness d_1, coated on both sides with an emissive layer of thickness d_2.
b) Round wire of diameter d_1, similarly coated with an emissive layer of thickness d_2.

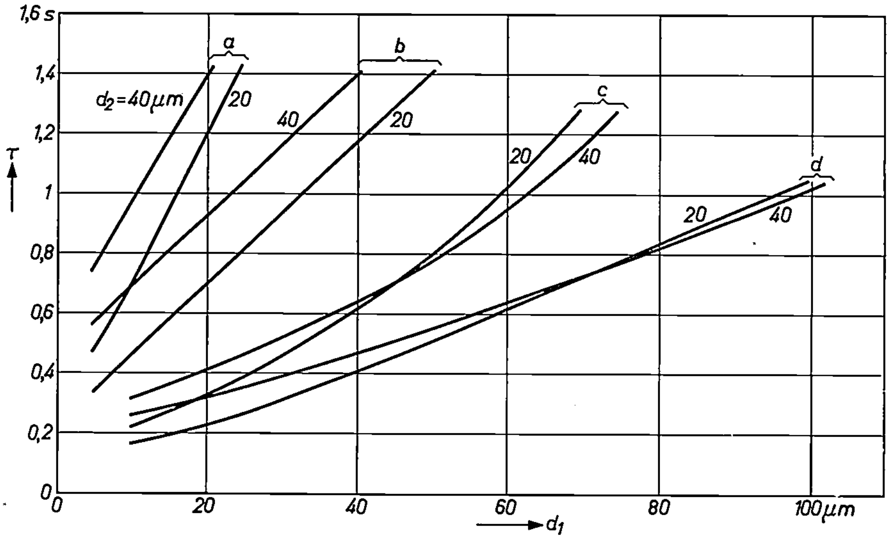


Fig. 3. Warming-up period of the cathodes represented in fig. 2, plotted as a function of the core thickness d_1. *a* nickel strip, *b* tungsten strip, *c* nickel wire, *d* tungsten wire. The two curves for each cathode relate to two different thicknesses of the emissive barium-oxide layer, viz. d_2 = 20 μm and d_2 = 40 μm.

$$K = (d_1 s_1 + 2 d_2 s_2) lb. \quad . \; . \; . \; . \; (9)$$

After the final temperature has been reached, the power supplied to the cathode is equal to the radiated power. At 1000 °K this power for barium-oxide is roughly 1.8 W/cm². In the final state the supplied power is thus:

$$P = 2 \times l \times b \times 1.8,$$

so that, using (5), we find for the warming-up period:

$$\tau = 750\, F(d_1 s_1 + 2 d_2 s_2). \quad . \; . \; . \; . \; . \; . \; (10)$$

The heat capacities s_1 of nickel and tungsten are 0.94 and 0.63 respectively, while s_2 for barium-oxide is 0.15 gcal/cm³. Using these data and the values of F given above, we can now calculate from (10) the warming-up periods of various strip cathodes. *Fig. 3* gives the results of some calculations of this kind, plotted as a function of the thickness d_1 of the strip.

[1] The power lost at the ends by heat conduction can be disregarded owing to the considerable length of the cathodes.

[2] In addition to barium-oxide, conventional oxide-coated cathodes contain a substantial percentage of strontium-oxide, and some cathodes contain calcium-oxide as well.

The curves marked *a* refer to nickel strip, and those marked *b* to tungsten strip; in both cases the calculations were carried out for two different oxide-layer thicknesses d_2, viz 20 μm and 40 μm.

The second form of cathode that we shall consider consists of *round wire* with a diameter of d_1 cm, again of length l and with a barium-oxide layer d_2 cm thick (fig. 2*b*). The heat capacity of this cathode is:

$$K = \frac{\pi}{4}\{d_1{}^2 s_1 + 4d_2(d_1 + d_2)s_2\}\, l, \quad . \; . \quad (11)$$

and in the final state the power supplied and radiated is:

$$P = \pi(d_1 + 2d_2)\, l \times 1.8. \; . \; . \; . \; . \; . \quad (12)$$

From this the warming-up period is found to be:

$$\tau = 375\, F \frac{d_1{}^2 s_1 + 4d_2(d_1 + d_2)s_2}{d_1 + 2d_2}. \quad . \; . \quad (13)$$

The warming-up periods calculated from (3) for various cathodes are also shown in fig. 3 as a function of the thickness d_1 of the metal core. The curves marked *c* refer to nickel wire and the curves marked *d* to tungsten wire. Here again, two thicknesses d_2 have been taken for the barium-oxide layer, viz 20 μm and 40 μm.

From fig. 3 the following conclusions can now be drawn.

1) If a warming-up period of about 1 second is permissible, a cathode can be used which consists of nickel strip with a thickness of about 15 μm, tungsten strip of about 30 μm, nickel wire of about 60 μm, or tungsten wire of about 100 μm. In the QQC 03/14 valve nickel strip 14 μm thick with a barium-oxide layer of 20 μm is used. The warming-up period of this cathode is 0.8 second.

2) If a warming-up period of 0.5 second is required, the strip form of cathode would have to be so thin as to rule it out for applications on a commercial scale. As can be seen from fig. 3, nickel strip of about 5 μm, or tungsten strip of about 10 μm would have to be used. If the cathode is made of nickel wire, a thickness of roughly 30 μm can be taken; tungsten wire can be employed with a thickness of 50 μm.

Tungsten wire with a thickness of the order of 10 μm has long been manufactured for use in battery valves [3]. The foregoing therefore shows — somewhat surprisingly — that the objective required can be reached with an existing type of wire, which is not considered particularly thin in valve technique. For the cathode presently to be discussed the choice fell on tungsten wire with a thickness of 25 μm, which is still substantially less than the maximum permissible thickness. Experience has shown that if thicker wires are used, there are difficulties with the adhesion of the barium-oxide layer to the metal core, due to the difference between the expansion coefficients of tungsten and barium-oxide. The barium-oxide coating on the wire is 15 μm thick. The warming-up period of cathodes made from such wire is 0.35 second.

3) For a warming-up period of 0.1 second the wire or strip would have to be so thin that it would be impossible to produce such cathodes on a commercial scale. It is, however, fairly simple to reduce the warming-up period of the cathode mentioned under (2) from 0.35 second to 0.1 second by other means. This can be done by so designing the circuit that, immediately after switching on, the filament voltage is at roughly twice the normal magnitude. As soon as the cathode has reached the required temperature, the filament voltage is returned to the normal value. This can be done by a relay with its winding in the anode circuit of the valve.

Construction and assembly of the harp cathode

The current supplied by the cathode in a small transmitting valve has to be much greater than that required in a receiving valve (e.g. ten times greater). In a transmitting valve the emissive surface must therefore be larger. Since the valves concerned have to operate at high frequencies (500 Mc/s), this larger surface is necessarily associated with a very small distance between grid and cathode. Using the thin tungsten wire referred to, these requirements can easily be met if the cathode is arranged to consist of a number of parallel wires, stretched tight in the same plane. The appearance of a cathode produced in this way has given rise to the name "*harp cathode*".

This cathode is employed in various forms. The basis is always a pair of metal supports which can be mutually displaced along an electrically insulating spacer. Both supports are wound with several turns of the appropriate wire, which are welded to the supports. After winding, the wire is tensioned by pulling the two metal supports apart with the aid of a spring. This also keeps the wire tight during the changes in length caused by thermal expansion.

In the arrangement illustrated in *fig. 4a* the two supports each consist of a piece of metal tube *P* of rectangular cross-section, to which is welded a U-shaped cap *Q*. The insulating spacer consists of a mica strip *M* which is a sliding fit in the two tubes *P*. While the wire *D* is being wound on, its barium-oxide coating is scraped away from the places where the turns lie on the supports.

Another arrangement is illustrated in fig. 4*b*. Here

[3] L. Schultink and P. G. van Zanten, Thin tungsten wire for small radio valves, Philips tech. Rev. 18, 222-228, 1956/57.

the supports each consist of a metal tube *B* to which a rod *S* is fixed. The two rods slide in a ceramic tube *Kb*. The wire *D* is wound around the two tubes *B* and the turns are again tensioned by a spring.

The harp cathode thus produced is mounted in the valve in a method also commonly employed for indirectly heated cathodes, i.e. by fitting the ends into two mica discs, M_1 and M_2. *Fig. 5*, where this is illustrated, also shows how the wires are tensioned; the cathode is fixed at the lower end to a metal bracket *H*, which rests on the bottom mica disc. The required tension is exerted at the upper end by the spring *V*

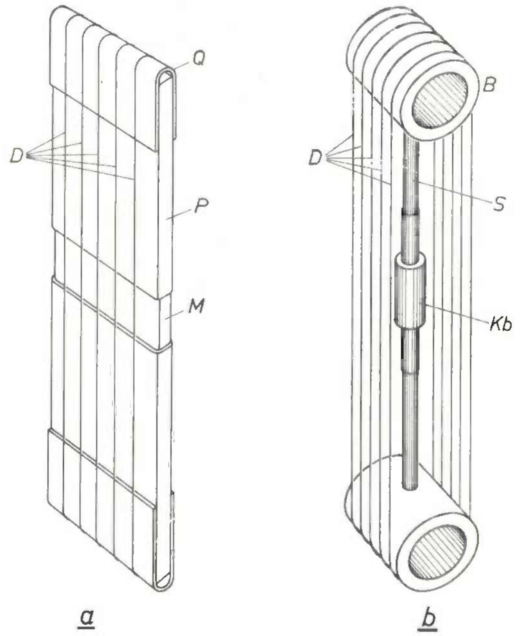


Fig. 4. Two types of harp cathode.
a) With mica strip. *P* metal tubes. *Q* caps. *M* mica strip. *D* tungsten wire.
b) With ceramic tube. *B* metal tubes. *S* metal rods. *Kb* ceramic tube. *D* tungsten wire.

The mica discs M_1 and M_2 also carry the other components of the electrode system. In *fig. 6*, which shows the construction of a YL 1190 twin tetrode, one of the anodes has been partly removed so that the method of assembling the various electrodes can be seen.

Properties of the wire used

As already mentioned, tungsten wire coated with barium-oxide has long been employed for the filaments in valves for battery receivers. Since these receivers draw their power from accumulators or dry batteries, whose terminal voltage can vary quite considerably, a fairly wide tolerance on filament voltage must be permissible without unacceptable shortening of the life of the valves. In the battery receiver valves in question it has been found that a filament voltage

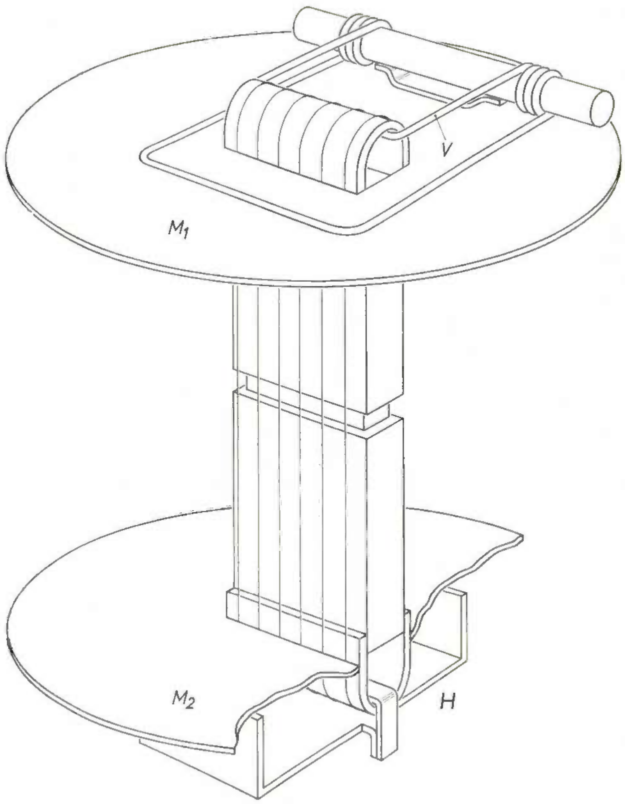


Fig. 5. Assembly of a harp cathode. M_1 and M_2 mica discs. *H* bracket. *V* spring.

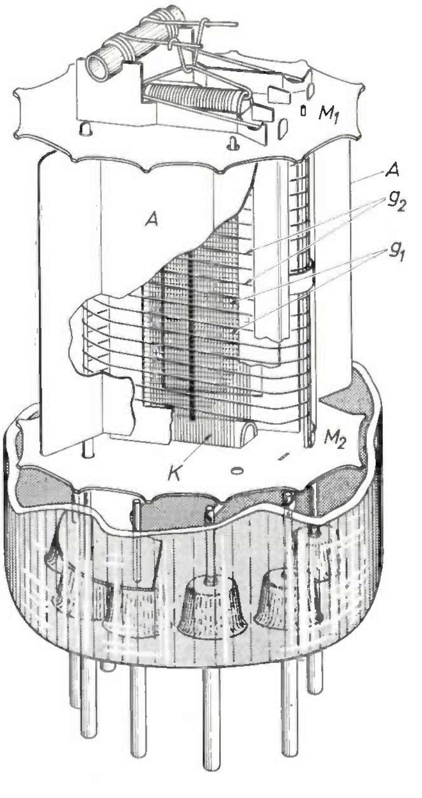


Fig. 6. View of a YL 1190 twin tetrode fitted with a harp cathode. One of the two anodes has partly been removed to show the cathode and the grids. M_1 and M_2 mica discs. *A* anodes. g_1 control grid. g_2 screen grid. *K* harp cathode.

deviation of ± 15% has no adverse consequences. In transmitting valves the required emission per cm^2 of cathode surface is greater than in receiving valves, and one may reasonably question whether under these circumstances the valve will have a sufficiently long life if its filament voltage deviates from the nominal value. Experiments have proved, however, that filament voltage variations of this order of magnitude can indeed be tolerated for transmitting valves, so that the relevant transmitters can quite safely be powered from accumulators or batteries.

An important property of tungsten, which is put to good use here, is its great tensile strength, even at high temperatures. This means that the wire can be kept under considerable tension, so that shocks will cause only very slight lateral deflection and not give rise to unduly strong microphony. Tungsten also exhibits very little sign of fatigue under varying load, and this means that the valves can be switched on and off a great number of times. At normal filament voltage, harp cathodes have withstood repeated tests in which they were switched on and off a million times. Even if the filament voltage is doubled at switching on, to reduce the warming-up period to 0.1 second as described above, the filament can still be switched on and off several hundred thousand times, which is amply sufficient for normal purposes.

A drawback of thin tungsten wire as cathode material, compared with indirectly heated nickel cathodes, ought to be mentioned at this point. Because of the nature of the core material, tungsten, and the thinness of the emissive layer, these cathodes are easily subject to emission poisoning, caused by residual gases in the valve. The thermionic emission of barium-oxide depends to a considerable extent on the presence of free barium at the cathode surface. Part of this free barium is continuously lost by reaction with residual gases in the valve, but it is replenished as a result of the reduction of the barium-oxide by substances added to the nickel. With tungsten this reaction takes place much more slowly than with nickel, and consequently the free barium lost is not so quickly replenished in a tungsten cathode. The formation of free barium might be accelerated by increasing the temperature, but this also speeds up the rate at which the cathode material evaporates, and this, in view of the thinness of the emissive layer in this case, would reduce the life of the valve.

For these reasons, transmitting valves with harp cathodes must receive a great deal of care during evacuation, to ensure that residual gases are removed.

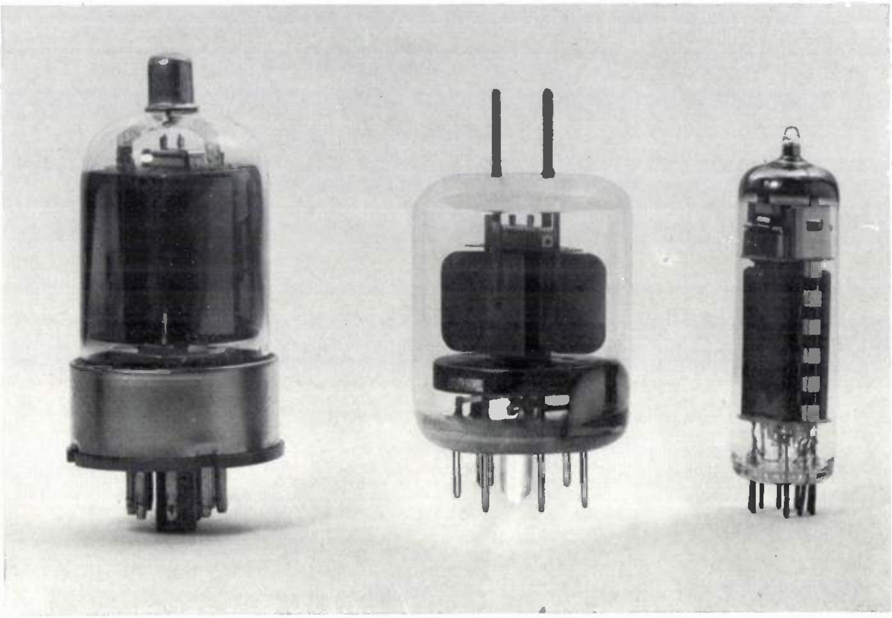


Fig. 7. Three valves fitted with harp cathodes. From left to right: a QC 05/35 tetrode (maximum frequency 175 Mc/s, power 30 W) and two twin tetrodes, YL 1020 (500 Mc/s, 18 W) and YL 1080 (200 Mc/s, 15 W).

The filament supply

The filament voltage for the various types of harp cathode is 1 to 2 V; the filament current is a few ampères. These values, if they are compared with the conventional ratings in standard valves, may seem perhaps somewhat unusual. It should be remembered, however, that a low filament voltage is advantageous in directly heated valves because the voltage between the ends of the cathodes is then lower. A relatively high voltage between various parts of the cathode can cause a non-uniform current distribution, adversely affecting the operation of the valve. This is particularly so in valves designed to handle high-frequency signals, where the driving voltage is low owing to the small distance between control grid and cathode.

The use of a filament voltage differing from that of other valves used in the same equipment presents in general no great difficulty. All mobile transmitter-receiver units are fitted with a converter for stepping up the voltage from the battery. Nowadays the converter circuit usually employs a balanced arrangement of two transistors which convert the battery voltage into a square-wave voltage [4]. By means of a transformer and a rectifier a d.c. voltage of the right magnitude is then generated. The filament voltage can be obtained quite simply by means of an extra winding on the transformer.

Valves with harp cathodes are now being produced to give a useful power of 6 W to 30 W and for operation at frequencies up to 500 Mc/s. *Fig. 7* shows three such valves — a tetrode (QC05/35) and two twin tetrodes (YL 1020 and YL 1080) — by way of example. Some data relating to these valves are mentioned in the caption.

[4] For further particulars see: T. Hehenkamp and J. J. Wilting, Transistor d.c. converters for fluorescent-lamp power supplies, Philips tech. Rev. **20**, 362-366, 1958/59.

Summary. In mobile transmitter-receiver equipment a considerable saving of energy is possible if the filaments of the transmitting valves can be switched off during the time that the transmitter is not in operation. To ensure that normal use can still be made of the equipment, the cathodes must then have a short warming-up period. The harp cathode, developed with this application in mind, consists of a number of tungsten wires connected in parallel and coated with an emissive layer of barium-oxide. The warming-up period of this cathode is only 0.35 second, and can be shortened, by means of a special circuit, to as little as 0.1 second. Two types of harp cathode arrangement are described, and some properties of the filament wire are discussed.

"PAILRED", an automatic single-crystal diffractometer

548.735

The experimental data needed for determining the arrangement of atoms in a crystalline substance are usually obtained by placing a crystal of the substance in an X-ray beam and measuring the intensity and direction of all reflected beams produced when the crystal is rotated. Depending on the complexity of the structure, the number of reflections may range from a few dozen to several tens of thousands. For processing the considerable amounts of data involved in an analysis the crystallographer generally makes use of an electronic computer [1].

Until recently the measurements referred to were usually carried out with the aid of photographic techniques. Measuring the location and intensity of every reflection spot on a photographic film is, however, tedious and time-consuming, and in general not particularly accurate.

In the Philips laboratories at Briarcliff Manor, U.S.A. [2], J. Ladell and P. G. Cath have designed and built an instrument with which the data for a crystal structure determination can be obtained with greater speed, accuracy and convenience. The instrument has been called PAILRED, which stands for Philips Automatic Indexing Linear Reciprocal-space Exploring Diffractometer. A provisional model was demonstrated in September 1963 at the Sixth Congress of the International Union of Crystallography in Rome, and since the middle of 1964 it has been in use at Philips Research Laboratories in Eindhoven (*fig. 1*). A slightly modified version has now been put on the market.

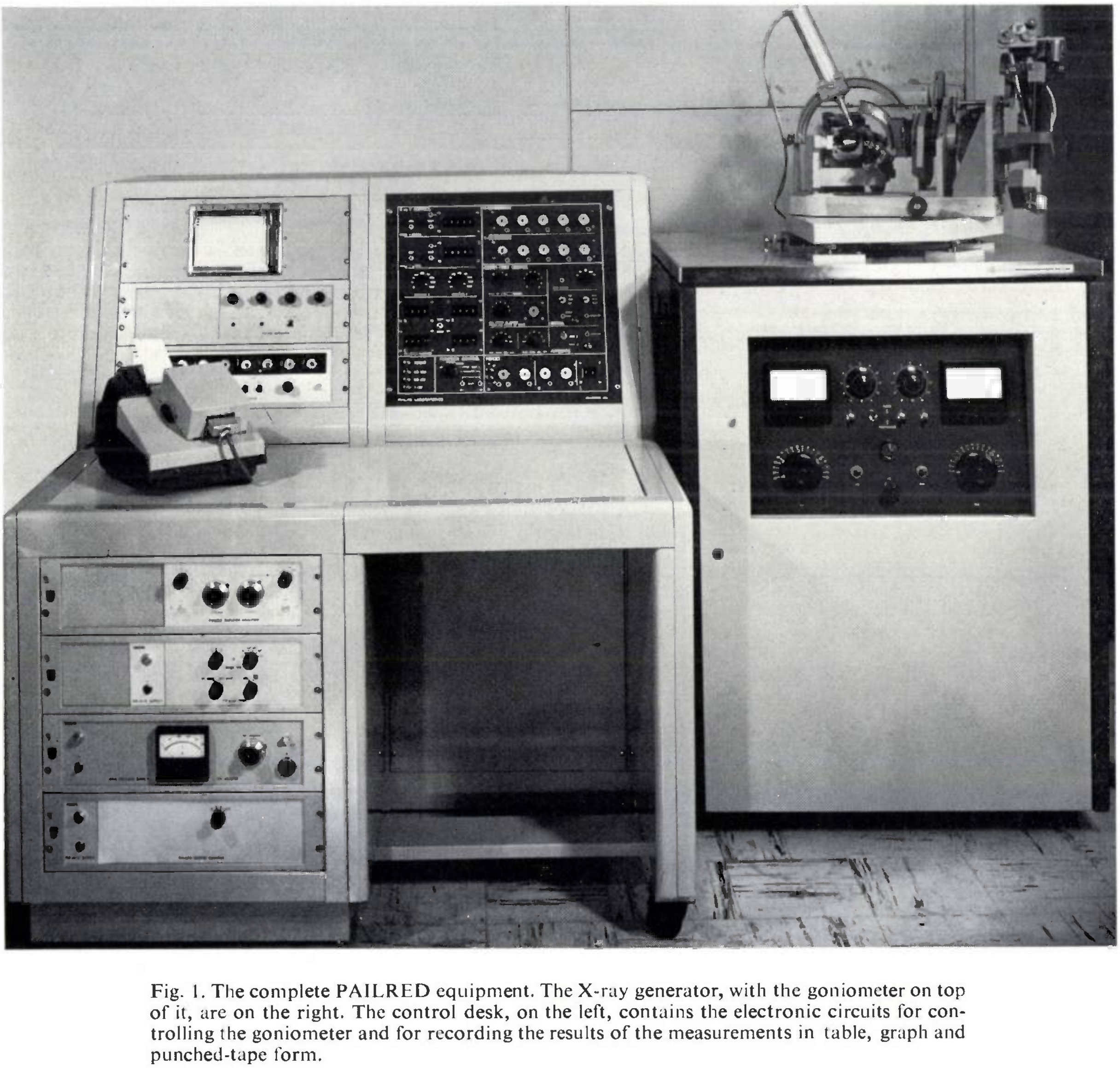


Fig. 1. The complete PAILRED equipment. The X-ray generator, with the goniometer on top of it, are on the right. The control desk, on the left, contains the electronic circuits for controlling the goniometer and for recording the results of the measurements in table, graph and punched-tape form.

This important instrument will be described here only very briefly. In due course a detailed article on it will appear in this journal.

The crystal under investigation is placed in the centre of a goniometer, on a small rotary mount; the crystal mount can be moved over a wide arc. This arrangement is known as an Eulerian cradle (*fig. 2*). Thus there are two rotational degrees of freedom which enable the investigator to select, within a wide range, any crystal direction, and make it to coincide with the horizontal axis of the instrument (ω axis). The X-ray beam incident on the crystal is made monochromatic by reflection from a flat auxiliary crystal, and this has the result that the background radiation, against which the reflection peaks stand out, is substantially reduced. The intensity measurements are made by means of a scintillation counter. The results are plotted in a graph, tabulated as decimal figures, and also punched in an eight-hole punched tape. The latter can be processed straight away by an electronic computer.

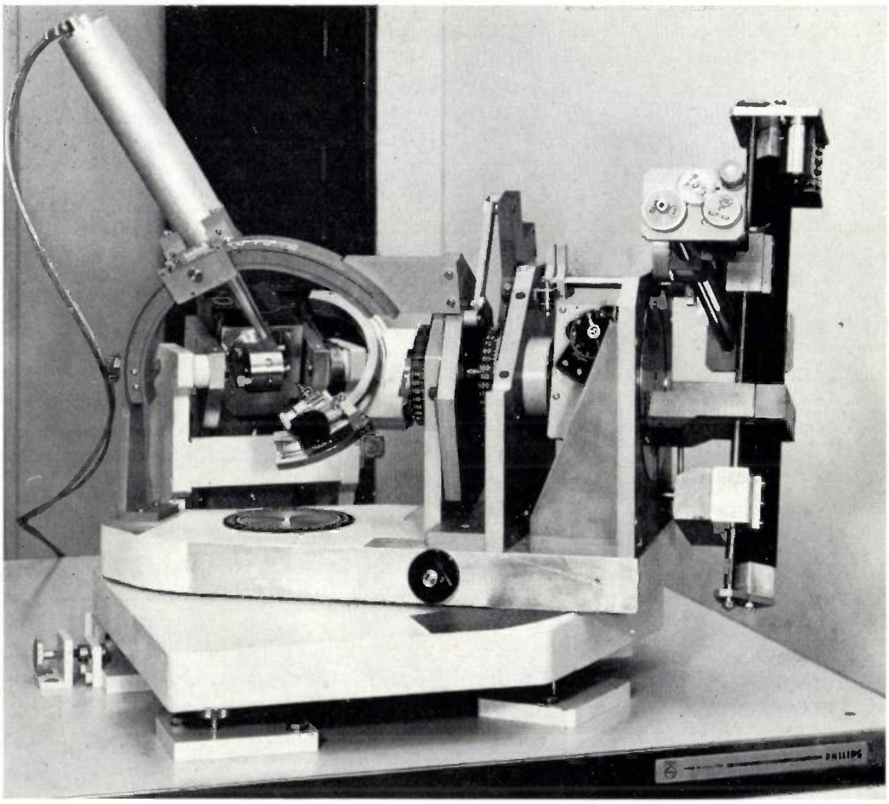


Fig. 2. The goniometer. Top left, the scintillation counter mounted on an arc and directed toward the crystal under investigation. In the centre can be seen the Eulerian cradle, which carries the crystal. The radiation source is behind it. The *X-Y* rectangular coordinate setting system is at the extreme right.

The crystal and the scintillation counter are coupled to a rectangular coordinate setting system (on the right in fig. 2), whose two drive-screws, *X* and *Y*, are each driven by a stepping motor. The mechanical design is such that a *constant* number of steps is needed in order to make the crystal and the counter change from one reflection position to the next, and this enables the measurements to be performed *automatically*. The correct (previously calculated) number of x and y steps is preset on the control panel, which can be seen in the centre of fig. 1, and behind which is located an elaborate system of logic circuits. The instrument then successively scans all reflections obtainable at the setting selected. The counter stops at each reflection position and a third stepping motor slowly rotates the crystal through a small angle around the ω axis, so that it passes through the reflection position. In this way one obtains the "integrated intensity" of each reflection. It is recorded in the three ways mentioned above, together with the background radiation, automatically measured in the neighbourhood of each reflection, and the reflection indices hkl, which give a numerical indication of the direction of the reflected X-ray beam, are also recorded.

The x-y plane is thus automatically scanned. The setting in the third direction, i.e. a rotation about the vertical axis of the instrument, has to be made by hand, but this 's not often required, and takes up little time. Typically, fifteen minutes might be spent on this once a day.

The PAILRED can work day and night without supervision. The speed of measurement depends on the accuracy required: a maximum of about 60 reflections per hour can be handled, but as a rule some 10 to 20 per hour are measured. The random error in the measured intensity of strong reflections is about 2%; that of weak reflections is determined as usual by the counting statistics. A corresponding excellent agreement has been found between calculated and measured intensities in the structure determinations so far made.

A. H. Gomes de Mesquita

[1] For the principles of crystal structure determination see for example P. B. Braun and A. J. van Bommel, Philips tech. Rev. 22, 126-138, 1960/61.

[2] Until recently this laboratory was at Irvington-on-Hudson, N.Y.

Dr. A. H. Gomes de Mesquita is a research worker at Philips Research Laboratories, Eindhoven.

Recent scientific publications

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

Philips Research Laboratories, Eindhoven, Netherlands	*E*
Mullard Research Laboratories, Redhill (Surrey), England	*M*
Laboratoires d'Electronique et de Physique Appliquée, Limeil-Brévannes (S.O.), France	*L*
Philips Zentrallaboratorium GmbH, Laboratory at Aachen, Weisshausstrasse, 51 Aachen, Germany	*A*
Philips Zentrallaboratorium GmbH, Laboratory at Hamburg, Vogt-Kölln-Strasse 30, 2 Hamburg-Stellingen, Germany	*H*
MBLE Laboratoire de Recherche, 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.

K. M. Adams: Contribution to the dynamical theory of the parametron.
Philips Res. Repts. **20**, 48-80, 1965 (No. 1). *E*

N. W. H. Addink: Anorganische massenspektrometrische Analyse.
Z. anal. Chemie **206**, 81-88, 1964 (No. 2). *E*

W. Albers and **C. Haas**: Electrical properties of binary transition metal compounds.
Physique des semiconducteurs, Comptes rendus du 7e Congrès international, Paris 1964, pp. 1261-1268. *E*

A. C. Aten and **J. H. Haanstra**: Electroluminescence in tellurium-doped cadmium sulphide.
Physics Letters **11**, 97-98, 1964 (No. 2). *E*

H. G. Beljers, **P. F. Bongers**, **R. P. van Stapele** and **H. Zijlstra**: A direct measurement of the large zero field splitting of Co^{2+} in Cs_3CoCl_5.
Physics Letters **12**, 81-82 and 360, 1964 (Nos. 2 and 4). *E*

G. Blasse: On the structure of some compounds $Li_3Me^{5+}O_4$ and some other mixed metal oxides containing lithium.
Z. anorg. allgem. Chemie **331**, 44-50, 1964 (No. 1/2). *E*

G. Blasse: The structure of some new mixed metal oxides containing lithium (II).
J. inorg. nucl. Chem. **26**, 1473-1474, 1964 (No. 8). *E*

G. Blasse and **D. J. Schipper**: Sulphospinels containing rhodium.
J. inorg. nucl. Chem. **26**, 1467-1468, 1964 (No. 8). *E*

A. J. van Bommel: Anomalous transmission in a bent germanium crystal.
Acta crystall. **17**, 956-959, 1964 (No. 8). *E*

M. Borot: Luminescence de l'arséniure de gallium sous bombardement cathodique.
J. Physique, Suppl. Phys. appl., **26**, 115A - 118A, 1965 (No. 3). *L*

A. J. Bosman and **S. van Houten**: Mechanical and dielectric relaxation in transition metal oxides.
Physique des semiconducteurs, Comptes rendus du 7e Congrès international, Paris 1964, pp. 1203-1209. *E*

G. Buchta: A reciprocal ferrite phase shifter.
Proc. IEEE **52**, 304-305, 1964 (No. 3). *H*

K. H. J. Buschow and **J. H. N. van Vucht**: On the intermediate phases in the system samarium-aluminium.
Philips Res. Repts. **20**, 15-22, 1965 (No. 1). *E*

P. J. Buysman: Het ontstaan van blazen in de diplaag van keramische gietvormen.
Klei en Keramiek **14**, 134-142, 1964 (No. 5). *E*

B. H. Clarke: Ferrimagnetic resonance linewidth in rare-earth-doped YIG at low temperatures.
J. appl. Phys. **36**, 1211-1212, 1965 (No. 3^{II}). *M*

J. A. Cundall and **A. P. King**: An investigation of the anisotropy distribution in polycrystalline nickel-iron films by intermediate and low field torque measurements.
Proc. int. Conf. on Magnetism, Nottingham 1964, pp. 847-851, publ. Inst. Phys./Phys. Soc., London. *M*

K. Deneke and **A. Rabenau**: Über die Natur der Phase In_3SbTe_2 mit Kochsalzstruktur.
Z. anorg. allgem. Chemie **333**, 201-208, 1964 (No. 4/6). *A*

J. Dieleman: Structuur van een zilvercentrum in zinksulfide.
Chem. Weekblad **60**, 632-635, 1964 (No. 45). *E*

J. Dieleman, **C. Z. van Doorn**, **S. H. de Bruin** and **J. H. Haanstra**: Reversible light-induced blackening by charge transfer in zinc sulphide single crystals.
Physique des semiconducteurs, Comptes rendus du 7e Congrès international, Paris 1964, pp. 941-944. *E*

G. Diemer: Vaste-stof-beeldversterkers (Amplificons).
Ned. T. Natuurk. **30**, 228-233, 1964 (No. 6). *E*

C. Z. van Doorn: Het M-centrum in alkalihalogeniden.
Chem. Weekblad **60**, 626-631, 1964 (No. 45). *E*

C. Ducot, J. Cayzac and **R. Astor:** La modulation d'amplitude à porteuse supprimée dans les liaisons hertziennes à grande distance.
Onde électr. **45**, 179-194, 1965 (No. 455). *L*

S. Duinker: Forschung und Entwicklung in der Industrie.
Elektronik in unserer Welt, Oct. 1964. *H*

W. Ermrich: Influence of slow-electron impact upon gases adsorbed on tungsten, investigated by means of a field electron microscope.
Philips Res. Repts. **20**, 94-105, 1965 (No. 1). *A*

B. A. Evans: A linear ratemeter for crystal pulling.
J. sci. Instr. **42**, 153-155, 1965 (No. 3). *M*

Y. Genin: Le facteur d'éclipse d'un satellite terrestre.
Rev. MBLE **8**, 32-59, 1965 (No. 1). *B*

J.-M. Goethals: Calcul de sections efficaces d'éclairement pour des surfaces de révolution et optimisation de surfaces composites.
Rev. MBLE **8**, 20-31, 1965 (No. 1). *B*

W. van Gool: The possible use of surface migration in fuel cells and heterogeneous catalysis.
Philips Res. Repts. **20**, 81-93, 1965 (No. 1). *E*

C. A. A. J. Greebe: Versterking van akoestische golven in piëzo-elektrische halfgeleiders.
Ned. T. Natuurk. **30**, 254-257, 1964 (No. 7). *E*

C. A. A. J. Greebe: Some considerations on piezoelectric plates containing charge carriers.
Philips Res. Repts. **20**, 1-14, 1965 (No. 1). *E*

E. F. de Haan and **A. G. van Doorn:** The "Plumbicon": a camera tube with a photoconductive lead oxide layer.
J. SMPTE **73**, 473-476, 1964 (No. 6^{I}). *E*

C. Haas: Optisch onderzoek van puntfouten.
Chem. Weekblad **60**, 611-619, 1964 (No. 45). *E*

G. E. G. Hardeman and **G. Gerritsen:** Suppression of nuclear dynamic polarisation by radio frequency radiation.
Physics Letters **11**, 20-21, 1964 (No. 1). *E*

F. W. Harrison, J. F. A. Thompson and **G. K. Lang:** Single-crystal magnetization data for anisotropic rare-earth iron garnets at low temperatures.
J. appl. Phys. **36**, 1014-1015, 1965 (No. 3^{II}). *M*

F. W. Harrison, J. F. A. Thompson and **K. Tweedale:** Single crystal magnetization data for rare earth iron garnets below room temperature.
Proc. int. Conf. on Magnetism, Nottingham 1964, pp. 664-665, publ. Inst. Phys./Phys. Soc., London. *M*

J. Hasker: Transverse-velocity selection affects the Langmuir equation.
Philips Res. Repts. **20**, 34-47, 1965 (No. 1). *E*

Y. Haven: Anionic type of lattice defects.
Proc. Brit. Ceramic Soc. **1**, 93-108, July 1964. *E*

F. N. Hooge: Influence of counterdoping on the distribution of Mn over substitutional and interstitial sites in Ge.
Physique des semiconducteurs, Comptes rendus du 7e Congrès international, Paris 1964, pp. 1185-1188. *E*

F. N. Hooge and **D. Polder:** Conditions for superlinear intrinsic photoconductivity.
J. Phys. Chem. Solids **25**, 977-984, 1964 (No. 9). *E*

S. van Houten and **A. J. Bosman:** Mechanical and dielectric relaxation in transition metal oxides.
Proc. 1st Buhl int. Conf. on transition metal compounds, Pittsburgh 1963, pp. 123-136, published 1964. *E*

H. J. Hubers: A general method to separate the principal stresses in photo-elastic measurements and its application in the determination of residual thermal stresses.
Glass Technol. **5**, 157-163, 1964 (No. 4).

J. Israël: Het metallografisch onderzoek.
Metalen e.a. Constr.mat. **19**, 291-297 and 350-356, 1964 (Nos. 10 and 12), and **20**, 2-7, 1965 (No. 1). *E*

G. H. Jonker: Some aspects of semiconducting barium titanate.
Solid-State Electronics **7**, 895-903, 1964 (No. 12). *E*

Th. J. van Kessel: Compatibele eenzijbandmodulatie.
T. Ned. Elektronica- en Radiogen. **29**, 31-43, 1964 (No. 1). *E*

H. A. Klasens and **H. Koelmans:** A tin oxide field-effect transistor.
Solid-State Electronics **7**, 701-702, 1964 (No. 9). *E*

J. E. Knowles: Induced uniaxial anisotropy in magnetite at room temperatures.
Proc. int. Conf. on Magnetism, Nottingham 1964, pp. 619-622, publ. Inst. Phys./Phys. Soc., London. *M*

H. de Lang: Optische aspecten van de laser.
Ned. T. Natuurk. **30**, 195-210, 1964 (No. 5). *E*

F. K. Lotgering: Ferromagnetism in spinels: $CuCr_2S_4$ and $CuCr_2Se_4$.
Solid State Comm. **2**, 55-56, 1964 (No. 2). *E*

R. Memming: On the origin of fast surface states at the germanium-electrolyte interface.
Surface Sci. **2**, 436-443, 1964. *H*

L. Merten: Piezoelektrische Potentialfelder um Stufenversetzungen beliebiger Richtung in piezoelektrischen Kristallen mit elastischer Isotropie.
Z. Naturf. **19a**, 1161-1169, 1964 (No. 10). *A*

H. Mooijweer: Parametrische versterkers; een inleiding.
Ned. T. Natuurk. **30**, 145-158, 1964 (No. 4). *E*

E. Neckenbürger, H. Severin, J. K. Vogel and **G. Winkler**: Ferrite hexagonaler Kristallstruktur mit hoher Grenzfrequenz.
Z. angew. Physik **18**, 65-68, 1964 (No. 2). *H*

E. de Niet: Parametric amplifiers used in electroacoustics, 2. Condenser microphone amplifier with semiconductor elements.
J. Audio Engng. Soc. **12**, 186-191, 1964 (No. 3). *E*

L. M. Nijland and **J. Schröder**: Lichterzeugung durch Fluor-Reaktionen in Blitzlichtlampen.
Angew. Chemie **76**, 890, 1964 (No. 21).

M. Noé: Sur le problème d'approvisionnement des centrales thermiques.
Rev. belge Stat. Rech. Opérat. **5**, No. 3, 15-24, 1965. *B*

J. Nussli: Possibilités d'utilisation des photomultiplicateurs comme détecteurs de lumière modulée et amplificateurs à large bande en détection directe et hétérodyne.
J. Physique, Suppl. Phys. appl., **26**, 113A-114A, 1965 (No. 3). *L*

D. Polder and **P. Penning**: Anomalous transmission of X-rays in an elastically deformed non-isotropic crystal.
Acta crystall. **17**, 950-955, 1964 (No. 8). *E*

H. Polnitzky: Eispunkt-Thermostat.
Z. Instrumentenk. **73**, 11-13, 1965 (No. 1). *A*

G. Prast and **G. J. Haarhuis**: Die Philips-Gaskältemaschine für sehr tiefe Temperaturen.
Kältetechnik **16**, 232-235, 1964 (No. 8). *E*

O. Reifenschweiler: Hydrogen pressure regulator with high absorption rate.
Rev. sci. Instr. **35**, 456-460, 1964 (No. 4). *E*

J. A. Rietdijk: A positive seal for pistons and axially moving rods.
Engineer **218**, 346-347, 1964 (No. 5666). *E*

J. H. van Santen: The Philips Research Laboratories at Eindhoven, the Netherlands.
Chemistry and Industry, 1964, pp. 1564-1570 (No. 37). *E*

D. A. Schreuder: De luminantietechniek in de straatverlichting.
Ingenieur **76**, E 89 - E 99, 1964 (No. 30).

E. Schwartz: Reaktanzvierpole, Breitbandanpassung, Stabilität.
Arch. elektr. Übertr. **18**, 673-678, 1964 (No. 11). *A*

E. Schwartz: Die Abhängigkeit der Verstärkung eines selektiven Tunneldioden-Verstärkers von der Aussteuerung.
Arch. elektr. Übertr. **19**, 9-12, 1965 (No. 1). *A*

P. J. Severin: The low frequency impedance of the cathode fall region.
J. Electron. Contr. **16**, 381-391, 1964 (No. 4). *E*

M. J. Sparnaay: On the electrostatic contribution to the interfacial tension of semiconductor/gas and semiconductor/electrolyte interfaces.
Surface Sci. **1**, 213-224, 1964 (No. 3). *E*

C. G. Venis: Strippen-slijpmachine voor rubber en andere materialen.
Fijntechniek **4**, 117-119, 1964 (No. 8). *E*

M. L. Verheijke: Calculated efficiencies of NaI(Tl) scintillation crystals for aqueous cylindrical sources.
Int. J. appl. Rad. Isot. **15**, 559-563, 1964 (No. 9). *E*

H. J. Vink: Fysische chemie van puntfouten in nietmetallische kristallen.
Chem. Weekblad **60**, 601-611, 1964 (No. 45). *E*

J. Volger, F. A. Staas and **A. G. van Vijfeijken**: Viscous flow of flux in a pure superconductor of the second kind.
Physics Letters **9**, 303-304, 1964 (No. 4). *E*

J. W. ter Vrugt: Optical properties of thin powder layers.
Philips Res. Repts. **20**, 23-33, 1965 (No. 1).

N. Warmoltz, D. Admiraal and **E. Bouwmeester**: Ein Mehrzweck-Massenspektrometer.
Elektronik **13**, 5-10 and 71-76, 1964 (Nos. 1 and 3). *E*

J. D. Wasscher, A. M. J. H. Seuter and **C. Haas**: Spin-disorder scattering, anomalous behaviour of the Hall coefficient and magnon-drag in antiferromagnetic MnTe.
Physique des semiconducteurs, Comptes rendus du 7e Congrès international, Paris 1964, pp. 1269-1275. *E*

C. Weber: The electron beam in a cathode-ray tube.
Proc. IEEE **52**, 996-1001, 1964 (No. 9). *E*

K. R. U. Weimer and **H. Bodt**: Electron beam-plasma interaction in a TWT.
Proc. IEEE **52**, 965-966, 1964 (No. 8). *E*

K. Weiss and **H. A. Meijer**: Die thermische Ausdehnung von AgBr und AgBr-CdBr$_2$-Mischkristallen.
Z. phys. Chemie Neue Folge **42**, 211-220, 1964 (No. 3/4). *E*

J. S. C. Wessels: ATP formation accompanying photoreduction of NADP$^+$ by ascorbate-indophenol in chloroplast fragments.
Biochim. biophys. Acta **79**, 640-642, 1964 (No. 3). *E*

H.-O. Westermann: Das Reflexionsverhalten bituminöser Straßendecken im Zusammenhang mit der Griffigkeit.
Straßen- und Tiefbau **18**, 290-291 and 293-295, 1964 (No. 3).

W. J. Witteman and **J. Haisma**: Analysis of combination tones in a short gas laser.
Phys. Rev. Letters **12**, 617-619, 1964 (No. 22). *E*

Volume 26, 1965, No. 10 pages 285-320 Published 13th May 1966

PHILIPS TECHNICAL REVIEW

VOLUME 26, 1965, No. 11/12

The exploration of the unknown

N. F. Verster

53:62

The article below, except for the introduction, gives more or less in full the text of the address given on 21st May 1965 by Dr. N. F. Verster upon his inauguration as ordinary professor in Applied Physics at the Technical University of Eindhoven. As is often the case in discourses delivered in the presence of professional associates but which are meant to be intelligible to the lay audience, the author presents his subject matter in elaborate metaphorical terms. The seemingly commonplace events he depicts, and which everyone can follow, are charged with deeper implications which only the initiated — with their own secret language — will have the satisfaction of understanding.

For more than thirty years the Technical University of Delft has had a department of "Applied Physics", in which the students are trained as" physical engineers". This faculty is also represented at the Technical University of Eindhoven.

There is no reason to doubt the importance of this branch of learning; eminent positions achieved by its graduates — positions by no means restricted to lecturing in applied physics — make this abundantly clear. Nevertheless, some uncertainty may exist as to the precise connotation of the term "applied physics". Is it a science in itself? And if so, what exactly distinguishes it from the science of physics in general? The question is of particular concern to someone who is called upon to give instruction in applied physics after having himself been trained as a physicist. The concepts "experimental physics" and "theoretical physics" — even though they create a distinction which nowadays often seems to divide a man in two — are plainer, and the aspects of the whole wide range of physics that they seek to designate are easier to recognize, than I would venture to claim for the concept "applied physics".

I shall attempt here by an indirect route to clarify the place occupied by and the nature of applied physics. To this end it will be useful to make an excursion into the whole domain of the natural sciences, which I shall liken to the discovery, mapping and development of a new world.

The landscape, rich enough in its variety, is most conspicuous for the primeval forest that covers almost the entire territory. Experimental Physics is to be recognized in the activities of those who are hacking their way through the jungle with bushknife or bulldozer. They soon realize that this form of reconnaissance alone is not very satisfactory, for they are constantly brought to a halt by the bogs or abysses which abound hereabouts. The experimenters therefore hasten to ascend to hilltops from which, through the trees, they can glimpse a small part of the country ahead of them. The indispensable compass and simple surveying instruments they carry are identified as the tools which mathematics has placed at their disposal.

Occasionally they come up against real mountains, which the theoreticians attempt to climb. The higher they climb the wider the prospects before them. From their elevated vantage points they are able to survey large tracts of country, and in order to chart them they can, indeed must, employ much more powerful and accurate telescopes and theodolites. In the handling of such instruments they display considerable virtuosity.

The new world proves to possess many natural riches, such as minerals and recreation areas, which are of great importance to humanity as a whole, or through which the princes hope to increase their power and their wealth.

Whereas earlier explorers, like Von Guericke, were content to return from their expeditions with a few curious specimens, the results of systematic exploitation are now an integral part of the lives of those

Until his appointment to the Technical University of Eindhoven, Prof. Verster was a research worker at Philips Research Laboratories, Eindhoven.

who barely know the name of the new world. Think of our present-day artificial lighting, our domestic refrigerators, or of X-ray photographs.

Only three centuries ago the whole territory was virtually unknown. This is not surprising: coming upon it from the Everyday World one sees only the dense forests which block the view and make the country seem impenetrable.

A few exploratory voyages had been made by the Ancients who, here and there, from a hilltop, had surveyed the country around them and drawn maps of it. On philosophical grounds they also drew maps of the rest of the territory, maps which were regarded as authoritative for centuries. The Alchemists, too, had ventured forth into the unknown land, being spurred on, as were certainly their patrons, by an old legend that somewhere in the forest an inexhaustible goldmine was to be found.

Gradually, other parts of the territory began to be penetrated. The maps handed down were no longer adequate; the Age of the Great Voyages of Discovery had dawned.

One by one the first hills and peaks were found, from which it was possible to map a surprisingly large area.

In the 17th century Newton, from the base camp which Kepler and Galileo had established on Mount Mechanics, climbed to a peak which was later named after him. Others, like Lagrange and Hamilton, later found observation points on this same mountain where they could set up more sophisticated and more powerful telescopes. Nowadays, grammar-school pupils, carefully guided by their teachers, go by coach and cable-car directly to the observation tower on the Newton peak, so that they can get to know the surrounding landscape. They are then required, to the point of tedium, to measure up a number of standard landmarks with primitive theodolites. It is questionable whether they are able to recognize any of these landmarks when they come across them in the land itself.

Some forty years ago an expedition climbed Mount Quantum Mechanics. It still seems a miracle that by taking bearings from the most unlikely positions, in a totally unknown range with such a strange formation, they were able to find the mountain itself, let alone find the way to the top in such a short time. The view surpassed the wildest expectations. Not only did it extend over the entire domain of Physics as it was then known, but it included some of the principal landmarks of Chemistry. Since the days of Alchemy the Chemists had, it is true, explored and opened up a great deal of territory, but complete triangulations had not yet succeeded. It now proved possible to remedy this situation once and for all, and what is more to link up the triangulations so as to produce one vast coordinate system for both Chemistry and Physics. Among all those who, heavily laden with surveying instruments, are now toiling up the sides of this mountain, you are sure to find at least as many chemists as physicists.

As more and more of the territory was explored it became clear that all provinces bordered upon one another often most unexpectedly, and in a manner that cannot be represented topologically in our landscape. I think, however, it would be going too far to compare physics to a landscape in a space of four or more dimensions, although it would better express the limitlessness of the individual provinces.

In this respect the profession of physics sometimes resembles the making of flaky pastry: innumerable unpredictable contacts are found between points that were far apart when the slice was just an ordinary slab of dough.

Because of these complex interrelationships it happens occasionally that an expedition reaches a summit that affords a view of the beacons put by other expeditions which have become bogged down, or which are laboriously blazing a trail through virgin forest adjoining a busy thoroughfare.

Accurate map-making and better observation stations are becoming increasingly necessary, and the problem of liaison between the many expeditions is growing increasingly acute. More and more publications appear with descriptions of routes, methods and techniques of overcoming obstacles. Many photographs, too, are taken from the summits newly attained and, great is the joy of those who recognize in them beacons they themselves have erected.

Round about the turn of the century exploration on a grand scale was begun in the Province of Electromagnetism. The nature of the landscape made it possible to advance quickly; wherever the soil was turned, new riches came to light.

The isolationism that often stemmed from the size and wealth of this Province is now being gradually overcome by the growing volume of frontier traffic with the mother country.

In addition to valuable raw materials the Province was found to possess recreation areas too. Now the capital really began to flow in. Broad highways were laid in order to make the fields of Radio, and later of Television, accessible to the seekers of entertainment. With the laying of the Transistor Route through the Domain of Solid State Physics, peace and quiet have been banished from the recreation areas forever. Many expeditions now depend on the roads through the Region of Electronic Engineering to get their equipment through to their own working territory. The development of this Region also led to the discovery of minerals which could be employed in the manufacture of surveying and measuring instruments of unprecedented power and precision, called computers.

The computer is increasingly changing the art of map-making. Although earlier explorers were able to acquire a broad view of the landscape from the various hilltops and peaks, topographical details were often concealed by the Mist of Mathematical Drudgery. Now, however, the new instruments are being used to take pictures through the mist. All the same, since the magnification is greater, the field of view is proportionally smaller, and therefore the use of computers for taking properly directed pictures of such details presupposes thorough previous study by the old methods. Proceeding in this way, and provided viewpoints have been established and the local situation is not too unfortunate, much of the time can be saved that used to be spent on tedious reconnaissance of the ground itself.

It is not to be wondered at, now that Computers are steadily getting bigger and more powerful, that a separate branch of engineering is growing up around them. It is all rather reminiscent of the building of the Tower of Babel, as described in Genesis:

> "And now nothing will be restrained from them, which they have imagined to do.
> Go to, let us go down, and there confound their language, that they may not understand one another's speech."

Meanwhile there has been further penetration into the unknown country, first into the Region of Nuclear Physics, and now into the hinterland of the Elementary Particles. The latter in particular is for the major part not perceptible from the summits scaled hitherto. The first reconnaissances were undertaken by explorers who travelled thither along the narrow pathways of Natural Radioactivity and Cosmic Radiation. It was soon realized that better roads were needed to gain proper access to this area.

The country to be traversed lies completely within the purview of mountains discovered long ago, but the nature of the ground and the great distances involved left no doubt that the laying of roads here, certainly up to the hinterland of the Elementary Particles, was going to be a major feat of engineering.

The first results were achieved with a Cockroft-Walton generator. This brought the pathfinders as far as the 1 to 2 mega-electron volt area, most of which has now been mapped.

The next step forward came with the invention of the cyclotron by Lawrence in 1929 [1]. A fleeting inspection from above disclosed the route to be followed. It was a route that ran between the precipitous sides of the Phase Deviation and Vertical Instability Ravines, where the path became ever steeper and narrower. No-one could get further than about 20 MeV. When a start was made on laying the roads which the fleeting inspection had indicated, the whole territory proved to be much more treacherous than might have been supposed from the rather scantily detailed maps. Nevertheless, some groups, who were not easily discouraged by setbacks, learned by experience to clear the obstacles in their path. The building of cyclotrons became a romantic adventure. This situation looked like continuing indefinitely, for it was difficult by making systematic observations from good vantage points to produce maps detailed enough to be able to compete with the experience of seasoned guides. Moreover, the routes followed had never been established by accurate measurements in the field, so that the bogs and quicksands which the guides spoke of were not to be identified with the results of systematic mapping.

In 1938 an astronomer named Thomas [2] scrutinized the territory once again from the observation stations familiar to him. He indicated an alternative route, later to be known as the synchronous or AVF Cyclotron Route, which would bypass the perilous terrain between the two ravines we have mentioned and lead up to heights of 100 MeV or more. To really penetrate thus far it was necessary, however, after thoroughly studying the layout of the land, to mark out the trail with extreme care, and to verify by scrupulous measurements that there was no departure from it. In spite of its obvious advantages, nobody at the time was willing to venture upon the project.

After the war, the synchrocyclotron became known thanks to the discovery of Phase Stability by Veksler [3] and McMillan [4]. This constituted a method of avoiding the abyss called the Phase Error, making it possible to climb to 600 MeV and even more and so to penetrate a short distance into the hinterland of the Elementary Particles. As the pole pieces of the biggest cyclotron were six metres in diameter, how far it was possible to go now depended partly on whether lathes were now available big enough to machine them.

The experience gained in the building of classical

[1] E. O. Lawrence and N. F. Edlefsen, Science 72, 376, 1930.
[2] L. H. Thomas, Phys. Rev. 54, 580, 1938.
[3] V. I. Veksler, Doklady Akad. Nauk S.S.S.R. 43, 329, 1944.
[4] E. M. McMillan, Phys. Rev. 68, 143, 1945.

cyclotrons was turned to fruitful use, and as the path was strewn with fewer pitfalls a relatively simple theoretical analysis was sufficient. The synchrotron, likewise based on the principle of phase stability, was the next step, and in this way 10 000 MeV was reached. It took 36 000 tons of steel to make the magnet for this synchrotron, so that expense now became the major barrier to progress.

About 15 years ago the procedure began to be modified. Gradually it became the practice to draw up beforehand detailed maps of the country to be traversed. Although the advent of computers did much to bring this about, better analyses by well-tried methods proved to be the main equipment required, provided the surveyors took the trouble to climb a little higher.

Improved maps acted as a stimulus to more accurate measurements. Such was the success achieved that the scales sometimes looked like tipping the other way. The routes marked out have meanwhile become more and more ingenious, and even before roads are completed along previously planned paths they are hailed as "classic". Let me give you three examples of this new procedure.

First, the system of *extraction* in synchrocyclotrons.

For the classic cyclotrons methods had been devised by which, albeit with drastic means, the fast particles could be extracted from the accelerator and conducted to the experimental area. These methods were not applicable with the much larger synchrocyclotrons. In 1951 Tuck and Teng [5] proposed a new method for the big Chicago synchrocyclotron, but primitive analysis restricted their view of the territory and the project failed. After the theoretician Le Couteur [6], looking out from a higher plain from which he could perceive the hazards more clearly, had traced out a safe route, Crewe and his associates [7] brought his results successfully into practice with the Liverpool cyclotron.

Le Couteur was not content with merely drawing a safe route along the lines proposed by Tuck and Teng; he mapped a somewhat wider region and was therefore able later to postulate a shorter route [8]. Owing to the haze of rather more difficult equations that hung over the landscape a computer had to be used to mark out the exact route. After a searching analysis, mainly following the directions indicated by Le Couteur, a few detailed pictures taken with a computer sufficed to carry the method into effect at Orsay [9] and elsewhere. The detailed maps thus made of the area were found to be surprisingly accurate.

Further reconnaissance led to the discovery of a somewhat more refined method, which was subsequently used with success with the synchrocyclotron at Göttingen [10]. Since synchrocyclotrons are scarcely ever built nowadays, the development I have described is now a chapter of history.

As my second example I shall take the building of the giant *synchrotrons* at Geneva and at Brookhaven.

The "Alternating Gradient Principle", discovered by Christophilos [11] and by Courant, Livingston and Snyder [12], made it possible to cut down drastically on the costs of building large synchrotrons, bringing a 30 000 MeV machine within the realms of possibility.

For three reasons it was essential here to map out the territory in the minutest detail beforehand. In the first place, the very size of the project allowed no risks to be taken. Secondly, the Alternating Gradient route was no easy one, and a preliminary survey showed that the terrain was beset with deep rifts and swamps. The marking of the route therefore called for extreme care. And thirdly, as the project was so expensive, a great deal of optimization work was needed. The realization of these machines has been extremely successful.

One of the obstacles in the theoretical analysis was the Chasm of Non-linear Resonances. It could be seen that the chasm was there, but just how wide and deep it was proved to be difficult to ascertain. Attempts to find this out had, it is true, already been made, but the results were not noticed in the flood of publications. It was discovered, however, that Moser [13], like Thomas an astronomer, had recently been working at Göttingen on a similar problem, encountered when investigating the origin of the rings around Saturn. He found that these same non-linear resonances, here due to the moons circling the planet, provided an explanation of the ring structure.

To reach this conclusion he had chosen on the summit of Mount Mechanics an observation point which, as we have seen, had previously been found by Hamilton and others. The view from here is the same as from the Newton peak, but the situation allows the use of far more powerful mathematical instruments. It is indeed an excellent situation from which to study the movements of the planets. Considerable knowledge on this subject is now available, however, so that as a rule the only persons to be found there were a few investigators engrossed in fundamental problems or more advanced students on an excursion under the careful tutelage of a professor. You will not be surprised to learn that nowadays the designers of accelerators are constant visitors to this part of the summit. It turns out moreover that Le Couteur's observation point for studying Cyclotron Deflections lies close to this summit.

My third example is the *Isochronous* or *AVF* (Azimuthally Varying Field) cyclotron. Some ten years ago the enormous advances made in accelerator design led investigators to focus attention on the proposal published by Thomas a good fifteen years previously. The first synchronous cyclotron was completed in 1958 by Heyn [14] at Delft.

The theory is rather more complex than that of the alternating gradient synchrotron, but the size of the project is many times smaller. Generally speaking, the building of these machines was preceded by the most painstaking studies, again employing the theory of classical mechanics as formulated by Hamilton and others. Although the results of that theory seem rather abstract, it so happens that in their application to cyclotrons a straightforward interpretation is possible.

The employment of the electronic computer in the development of these machines has proved most fruitful, especially in connection with elaborate and exact measurements of the magnetic field. Accurate mapping from above thus goes hand in hand with precise measurements on the ground. I know for a fact that one project involved more than 50 000 magnetic field measurements. Measurements on such a scale are automated as far as possible, and the processing of the mass of figures obtained is also left to the computer[15].

In the picture I have given you of the activities in the profession of physics we recognize Fundamental or Pure Physics in the explorer's efforts to penetrate into unknown territory while constantly endeavouring to improve and extend the main triangulation. The explorers do not scrutinize every square yard but try to find as quickly as possible new vantage points from which they can discern the route to greater heights.

When they discover natural riches in the country they have opened up, or at least deduce their existence, the explorers first attempt to gather as much information as possible. If the results are encouraging, work then starts in earnest, for the path by which the explorers travelled is seldom suited to full-scale exploitation. Canals are needed, railways and well-surfaced roads which, because of the complex interdependence of the individual provinces, have to be cut through many and various regions that had once been explored but have since, owing to lack of time or interest, remained virtually untrodden. The thorough groundwork now required may at times assume the form and character of fundamental research.

In exploiting the land every use is made of communications previously established for one purpose or another. Conversely, it is more the rule than the exception that newly laid roads also prove useful to others, not least to fundamental research workers.

Exploiting and bringing forth new natural resources, surveying the approaches, marking out the routes and then laying the necessary roads and communications, with all that this involves, all these constitute *applied* physics. It also comprises, of course, the corresponding activity associated with fundamental research, as we have just seen in the case of cyclotrons, for example.

The training of the applied physicist has much in common with that of the engineer.

There is much that is appropriate here in Simon Stevin's instruction to the first Dutch school for "engineers" in 1600 [16]. I quote: "When they have gained enough experience in making measurements on paper, and, by understanding small matters, know what to do in great ones, they shall proceed to direct measurements on the land, whereby it shall be explained to them how they may use for land measurements other instruments than the rule, compass and set-square used for drawing on paper, but which serve the same purpose." "Thereafter they shall learn to draw on paper the contours of the lands they have thus measured, and conversely to mark out on the land, by the placing of rods and staves, the same which they have drawn on paper."

An engineer's education is not only a matter of "land measurement", in other words of learning theory but also of doing practical work: "And having reached thus far, they may in the summer go into the country or to places where fortifications are being built, and seeing there that which is real, even take part in the pursuit of the work".

With this rather elaborately sustained metaphor I have attempted to place applied physics in its proper context within the ambit of physics as a whole. I hope that I have given you some inkling of how fascinating it all can be. For my own part I find it particularly so in cases where theoretical analysis is able to

[5] J. L. Tuck and L. C. Teng, Phys. Rev. **81**, 305, 1951.

[6] K. J. Le Couteur, Proc. Phys. Soc., London B **64**, 1073, 1951.

[7] A. V. Crewe and K. J. Le Couteur, Rev. sci. Instr. **26**, 725, 1955.

[8] K. J. Le Couteur and S. Lipton, Phil. Mag. **46**, 1265, 1955.

[9] G. T. de Kruiff and N. F. Verster, The Orsay 160 MeV synchrocyclotron with beam-extraction system, Philips tech. Rev. **23** 381-400, 1961/62.

[10] G. T. de Kruiff and N. F. Verster, CERN report 63-19 (Proc. Intern. Conf. on sector-focused cyclotrons, Geneva 1963), p. 80.

[11] N. Christophilos, U.S. patent 2736799.

[12] E. D. Courant, M. S. Livingston and H. S. Snyder, Phys. Rev. **88**, 1190, 1952.

[13] J. Moser, Nachr. Akad. Wiss. Göttingen IIa, 1955, p. 87 (No. 6).

[14] F. A. Heyn and Khoe Kong Tat, Rev. sci. Instr. **29**, 662,1958.

[15] N. F. Verster and H. L. Hagedoorn, Investigation of the magnetic field of an isochronous cyclotron, Philips tech. Rev. **24**, 106-120, 1962/63.

[16] See P. C. Molhuysen, Bronnen tot de geschiedenis der Leidsche Universiteit, 's-Gravenhage 1913, p. 389 ff.

go arm in arm with technical accomplishment. At first sight the theory seems a maze of meaningless formulae, and results turned out by a computer are just sheets full of figures. But as you juggle with them more and more they begin to acquire form and meaning until a point is reached, often after manipulations by apparently abstruse mathematical methods, where the way through the maze becomes clear. It is then that you perceive the significance and understand the interplay of all kinds of factors, which you try to fit into a harmonious design. In applied physics the greatest advantage of good theoretical insight is often that it enables you to predict the likely effects of imperfections in the realization of a project, and on the basis of your estimates to formulate the accuracy required. It gives you a great feeling of satisfaction when your predictions are later confirmed by experiment.

What, then, should be the character of education in this science? The highly specialized branches, as for example the design and building of machines like cyclotrons, are advancing rapidly, but much is just as rapidly becoming outdated or ceasing to be of interest. No one can predict what branch of physics will come into greatest prominence in the next thirty years. Owing to the complex interdependence of all the various domains, as sketched in the picture I have given you, a slight shift of viewpoint may easily open up vistas of entirely different problems. That is why, as I see it, the chief ingredients of our programme should continue to be a fundamental knowledge of the main branches of physics and of their connections with one another, a stiff dose of mathematics and of practical work in one field or another.

To conclude, let me read to you the words with which a former professor of applied physics, Bosscha, began his lectures in 1873 at the Polytechnische School at Delft [17]: "We must not forget that, in order to make effective use of the forces of nature, something more is needed than a knowledge of certain practical rules, recipes that have merely to be followed, and of certain methods and appliances that have merely to be imitated... A fundamental knowledge of the natural forces themselves, and a proper understanding of their laws and operation, these are the prerequisites for acquiring skill as an engineer."

[17] Commemorative publication issued by the Koninklijke Akademie and the Polytechnische School 1842-1905, Delft 1906, p. 244

Summary. Principal contents of the address delivered by the author upon his inauguration as ordinary professor at the Technical University of Eindhoven. Using extensive verbal imagery, in which the profession of the natural sciences is likened to the exploration and development of a New World, the speaker makes it clear in what respects "applied physics" — the subject he will teach — is to be distinguished from experimental and theoretical physics. Among the provinces metaphorically explored, special prominence is given to that of cyclotrons (a province in which the author has helped to blaze a trail in recent years).

Stereophonic radio broadcasting

I. Systems and circuits

N. van Hurck, F. L. H. M. Stumpers and M. Weeda

534.76:621.396

Stereophonic records and gramophones for playing them have been on the market for a good many years. "Stereo" is now breaking into sound radio, and in some countries there are already regular broadcasts. This article will deal with a number of questions relating to the radio transmission of stereophonic sound. In Part I the authors discuss various approaches to the problem, giving more detailed treatment to one particular system that is already being widely employed (FCC system). In Part II they deal with the susceptibility of stereo radio receivers to interference.

For a long time music reproduction via a single channel has been considered satisfactory, but in recent years it has been realized that a great deal can be lost through the lack of "body" or "depth" in the reproduced acoustic image. This has led to the development of gramophones and tape recorders in which two separate electrical channels handle signals originating from two microphones or groups of microphones placed at different angles. Naturally enough, it was asked whether or not stereo programmes could be transmitted by radio broadcasting stations. The question had all the more force in that stereophonic reproduction would offer clear advantages in radio plays, a specific product of sound broadcasting.

Any system or standard for transmitting stereophonic sound must in the first place be "compatible". The same consideration applies when broadcasting authorities are contemplating the introduction of single-sideband modulation [1] or colour television. In everyday parlance, it means that whatever stereo (or colour television) signals are broadcast, they must be suitable for reception and satisfactory reproduction by the millions of ordinary radio sets or ordinary monochrome TV receivers already in existence. Now, one obvious way of broadcasting stereo programmes, which was in fact tried in the early days, is to use two separate transmitters to radiate the signals from the right-hand and left-hand microphones. But this method is ruled out at once by the compatibility requirement: an ordinary radio set would only pick up the signals from one microphone, and neither of the two microphones in the studio is favourably placed to give satisfactory overall reproduction.

The question was, then, could stereophonic signals be handled by a single transmission channel in such a way that mono as well as stereo receivers would be able to provide good-quality reproduction at an adequate sound level?

This question has led to world-wide research. Since the stereo broadcasts envisaged would normally be of a high technical standard, most attention was given to frequency modulation; it is only in the VHF/UHF range that the necessary bandwidth is available. A number of systems suitable for this range of frequencies were studied by an international working party set up by the European Broadcasting Union, and one system was selected for further consideration by the Comité Consultatif International des Radiocommunications, an international advisory body that comes under the International Telecommunications Union. The CCIR was, however, unable to reach a decision at its 1963 meeting. Some countries, the Netherlands amongst them, backed a proposal to adopt a system that had been approved by an American governmental agency concerned with radio and television, the Federal Communications Commission; this system is already being employed in the United States by about a hundred radio stations. Representatives of other nations maintained that the time was not ripe for a final choice, without however putting forward definite technical objections

N. van Hurck, Dr. F. L. H. M. Stumpers and Ir. M. Weeda are research workers at Philips Research Laboratories, Eindhoven. In 1962, 1963 and 1965 Dr. Stumpers took part in the CCIR and EBU discussions on the subject of stereophonic broadcasting.

[1] See T. J. van Kessel, "Compatible" single-sideband modulation, Philips tech. Rev. 25, 311-319, 1963/64 (No. 11/12).

to the system proposed. The advantages of universal acceptance of the same system are clear to all; yet individual countries are sometimes reluctant to abandon a system that their own investigators have worked on or, alternatively, they may hope that further research will provide better solutions to certain problems. In such circumstances it is not the practice of the CCIR to decide by a majority vote, and on this occasion the decision was simply put off until the next general assembly, which is due to be held in 1966 [2]. It is however unlikely that the various countries will wait till then to introduce stereo broadcasting. Stereo programmes are already being regularly broadcast in the Netherlands, using the system just referred to, and the radio industry has developed receivers suitable for that system. It is to be hoped that postponement of the CCIR decision will not lead to the introduction of differing national systems, as has happened in the case of television.

In what follows we shall start by dealing with the various possible ways of transmitting stereophonic signals, and after this general introduction we shall discuss the system accepted by the FCC, which we can briefly call the "FCC system" [3]. In Part II of this article we give the results of investigations into the susceptibility to interference of signals transmitted under the FCC system, as picked up by stereo receivers specially designed for the system and by ordinary mono receivers.

Stereophonic signal transmission system

The fact that two audio-frequency channels are enough to provide a sound image satisfactory for most purposes has been confirmed by a whole series of investigations. Research on the subject has been done in these laboratories since 1939, by K. de Boer amongst other workers, and later by N.V. Franssen [4]. Subsequently, investigators began to question whether two complete transmission channels were really necessary in order to give the reproduced sound image "body" or "depth". This question was particularly investigated in Britain, and Percival [5] and his co-workers were able to show that an impressive three dimensional effect could be had from a normal a.f. channel plus a second channel with a bandwidth of only 100 c/s. However, demonstrations before the European Broadcasting Union failed to convince the experts from the broadcasting organizations that this system met the high standard of quality they wanted. The choice was thus narrowed down to systems that permit the transmission of two full-scale a.f. signals.

To make do with a single transmitter, all systems have to employ a *subcarrier* whose frequency lies above the highest audible sound frequency, and which is modulated by one of the two a.f. signals. Together with the complementary information constituted by the second a.f. signal, the modulated subcarrier is impressed in turn on a main carrier that is then radiated by the signal transmitter.

For the sake of compatibility it is the *sum* of the outputs from the right-hand and left-hand microphones that is chosen for direct modulation on to the main carrier. Mono receivers detect only this direct modulation, and since the sum is a kind of average of the left-hand and right-hand microphone outputs, quite satisfactory reproduction of the original sound is obtained in this way. The other a.f. signal, for modulating the subcarrier, is obtained by taking the *difference* between the right-hand and left-hand microphone outputs. Having recovered these *sum* and *difference signals* from the incoming waveform, the stereo receiver must derive from them the original right-hand and left-hand signals for its two loudspeakers. The required circuits are not complicated, and the method has the further merit of giving the left and right components equal treatment in the receiver, with the result that the quality of reproduction is the same for both. (This would not be so if the left-hand information was modulated direct on to the main carrier and the right-hand information was modulated on to the subcarrier, or vice versa.)

Subcarriers are commonly employed in telephony. They can be either amplitude-modulated or frequency-modulated, and many further variations are possible (for instance, either the carrier itself or one of its sidebands can be suppressed). In short, there is a variety of systems. Some work on the subject has been done in our own laboratories [6], and incidental use will be made of some of the results in the review that now follows.

The majority of investigators have preferred to amplitude-modulate the subcarrier, for two reasons. Firstly, it simplifies the detection arrangements. Secondly, in certain systems the amplitude-modulated subcarrier leaves more room for the sum signal sweep when both are frequency-modulated on to the main carrier. Accordingly, we shall confine ourselves here to systems in which the subcarrier is amplitude-modulated.

We shall use the symbols $L(t)$ and $R(t)$ to stand for the outputs from the left-hand and right-hand microphones, and introduce $M(t)$ and $S(t)$ to denote the sum and difference signals respectively:

$$M(t) = L(t) + R(t),$$
$$S(t) = L(t) - R(t).$$

In this method, then, the waveform modulating the radiated carrier consists of $M(t)$ plus the subcarrier, which has been amplitude-modulated with $S(t)$. Thus we can express the transmitted waveform as:

$$F(t) = M(t) + A(t) = M(t) + a\left\{1 + \frac{S(t)}{S_{\mathrm{m}}}\right\} \sin \omega_{\mathrm{h}} t, \quad (1)$$

where $A(t)$ is the modulated subcarrier, a is its amplitude, ω_{h} its angular frequency, and S_{m} is a constant representing depth of modulation. Obviously, overmodulation must not occur, so S_{m} is at the same time the maximum permissible value for $S(t)$.

An interesting case arises when a is equal to S_{m}. In this eventuality equation (1) becomes:

$$F(t) = M(t) + \{a + S(t)\} \sin \omega_{\mathrm{h}} t. \quad \ldots \quad (2)$$

A good approximation to the upper and lower envelopes of this function can be arrived at by first allotting $\sin \omega_{\mathrm{h}} t$ the value $+1$ and then the value -1. The two equations thus found are:

$$F_1(t) = a + M(t) + S(t) = a + 2L(t), \quad \ldots \quad (3a)$$

$$F_2(t) = -a + M(t) - S(t) = -a + 2R(t). \quad . \quad (3b)$$

Thus the a.c. components of the envelope curves are equal to twice the left-hand and twice the right-hand signal components. By making sure that both $L(t)$ and $R(t)$ remain smaller than $\frac{1}{2}a$ we can ensure that $F_1(t)$ is always positive and $F_2(t)$ is always negative, and it will then be possible to recover the left-hand and right-hand components from the incoming waveform with the aid of two amplitude detectors of opposite polarity.

Fig. 1 will give some idea of the way this kind of transmitted waveform is built up. The case illustrated is a simplified one: the right-hand component is assumed to be zero, the left-hand component, $L(t) = b \sin pt$, is a pure sine wave (fig. 1*a*). $M(t)$ is the same as $L(t)$, and so is $S(t)$. Fig. 1*b* shows the modulated subcarrier, which is described by:

$$A(t) = (a + b \sin pt) \sin \omega_{\mathrm{h}} t. \quad \ldots \quad (4)$$

The transmitted waveform appears in fig. 1*c*. The equations of the two enveloping curves are:

$$F_1(t) = a + 2b \sin pt,$$

$$F_2(t) = -a.$$

The frequency spectrum of this complete signal is given in fig. 1*d*. As may be inferred from equation (4), both sidebands have an amplitude of $\frac{1}{2}b$.

A signal of the type shown in fig. 1*c* modulates the

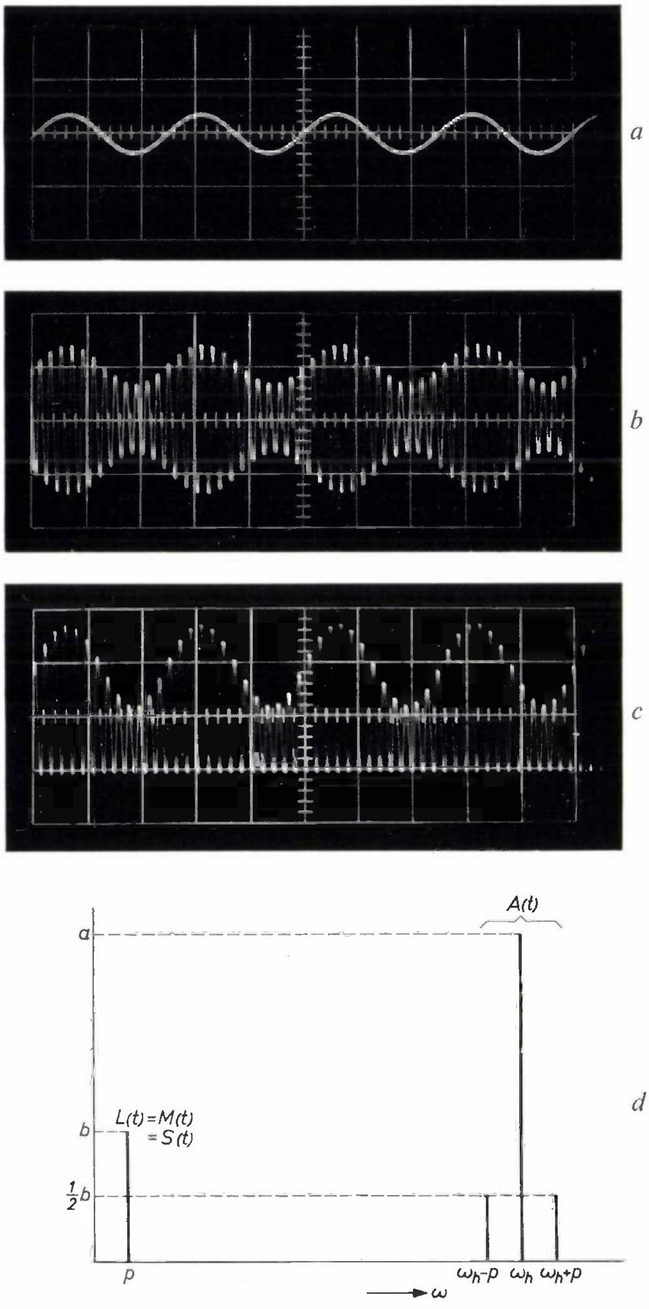


Fig. 1. One type of stereophonic waveform: in the case illustrated here only a left-hand component $L(t)$ is being transmitted.
a) Left-hand signal $L(t)$ which, in the case dealt with here, is the same as sum signal $M(t)$ and difference signal $S(t)$.
b) Modulated subcarrier $A(t)$.
c) $M(t) + A(t)$, the complete signal.
d) The frequency spectrum of the complete signal.

[2] In its Vienna meeting of Study Group X, CCIR has accepted a draft recommendation approving both the FCC system, and another FM-AM system with partly suppressed carrier, developed in the USSR.

[3] It is alternatively known as the "Zenith-General Electric system", after the companies which carried out the early investigations. The name "pilot tone system" is also occasionally used.

[4] See for example K. de Boer, The formation of stereophonic images, Philips tech. Rev. **8**, 51-56, 1946, or N. V. Franssen, Some considerations on the mechanism of directional hearing, thesis Delft, 1960.

[5] W. S. Percival, A compressed-bandwidth stereophonic system for radio transmission, Proc. Instn. Electr. Engrs. **106 B**, suppl. No. 14, 234, 1959.

[6] See for example F. L. H. M. Stumpers and R. Schutte, Stereophonische Übertragung von Rundfunksendungen mit FM-modulierten Signalen und AM-moduliertem Hilfsträger, Elektron. Rdsch. **13**, 445-446, 1959.

r.f. carrier radiated by the transmitter. As in FM the amplitude of the modulating signal determines the frequency deviation, the envelope of the transmitted waveform must be kept within certain limits. Owing to the presence of the subcarrier, only part of the allotted frequency sweep can be utilized for the sum signal. This means that a listener who does not have a stereo receiver will pick up a signal with a smaller frequency deviation than that of a non-stereophonic broadcast. In the special case to which fig. 1 relates, the maximum amplitude of the envelope $F_1(t)$ is equal to $a + 2b$ (corresponding to the maximum frequency deviation). Now, if we want to keep $F_1(t)$ positive at all times, so that direct detection can be employed, we must ensure that b never attains a value in excess of $\frac{1}{2}a$. This imposes a maximum value of $4b$ on $F_1(t)$, and implies that the frequency deviation caused by signal $M(t) = b \sin pt$ must be restricted to a *quarter* of the maximum allotted frequency deviation.

The situation turns out to be a little less unfavourable when we consider another special case, where the left-hand and right-hand components are exactly the same. Then $L(t) = R(t) = b \sin pt$ (*fig. 2a*). The sum signal $M(t)$ now has an amplitude of $2b$ (fig. 2*b*) and the difference signal $S(t)$ is zero; that is to say, the subcarrier remains unmodulated (fig. 2*c*). The final waveform carrying the stereo information has the shape shown in fig. 2*d*, which reveals that the frequency deviation caused by $M(t)$ can now extend over as much as *half* of the allotted sweep. Fig. 2*e* shows the resulting frequency spectrum.

In practice, of course, the simple cases represented in figs. 1 and 2 rarely arise, nevertheless, the fact remains that the deviation due to the sum signal has to be much smaller than the deviation allowed for an FM channel. Consequently the listener not possessing a stereo receiver loses such of the benefit, in terms of freedom from interference, afforded by the use of frequency modulation. This loss can to some extent be remedied by reducing the amplitude of the subcarrier, and so making a bigger share of the allotted frequency sweep for the sum signal. But this will not improve matters for the *stereo* set-owner, as the subcarrier and hence also the difference signal will then have a smaller signal-to-noise ratio. His position is already made worse by the fact that, because of the higher frequency on which the subcarrier is transmitted, the difference signal has a poorer signal-to-noise ratio than the sum signal. (We shall be coming back to this point in Part II.) It is not wise, therefore, to weaken the subcarrier more than strictly necessary.

A system has also been tried out in which the amplitude of the subcarrier is controlled by the difference signal in such a way that the degree of modulation is

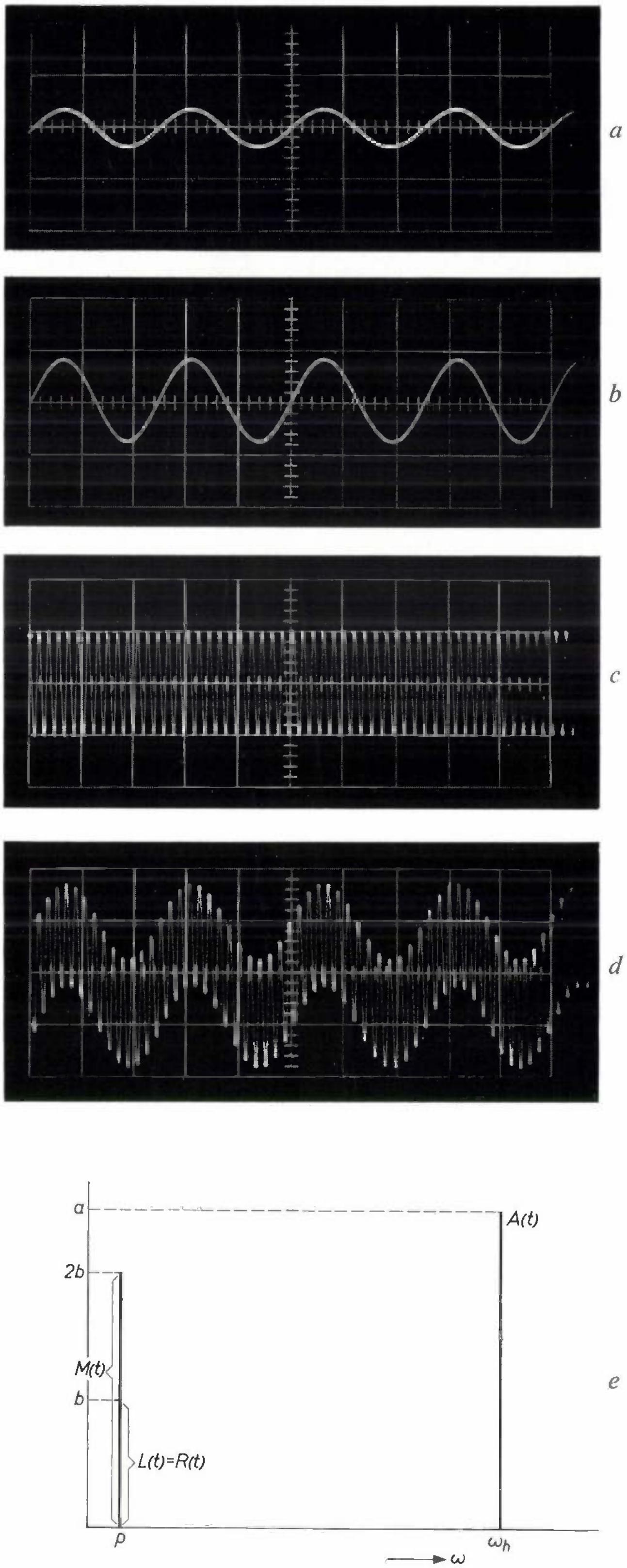


Fig. 2. Another stereophonic waveform for the same system as for fig. 1; here the left-hand and right-hand signals are identical.
a) Left- and right-hand signal $L(t) = R(t)$.
b) Sum signal $M(t)$.
c) Subcarrier $A(t)$.
d) The complete signal $M(t) + A(t)$.
e) The frequency spectrum of the complete signal.

always close to 100%. Among the advantages are reduced susceptibility to interference from other transmitters and, conversely, reduced liability of the stereo signals under consideration to interfere with those from other transmitters. But the system is not free from drawbacks; one of the most important is that detection of a deeply-modulated waveform may easily involve excessive distorsion.

Suppressed subcarrier system

Subsequently there was an American proposal to suppress the subcarrier altogether, so that only its two sidebands modulate the main carrier. But the subcarrier frequency is an indispensable element in the detection of the difference signal, and this means that it has to be generated locally, in the receiver. As an aid to this, the system provides for the transmission of a signal whose frequency is half that of the suppressed subcarrier. Let us take the symbol $P(t)$ to denote this "*pilot signal*", as we shall call it. We shall refer to the sidebands of the suppressed subcarrier as the "*stereo sub-signal*", with symbol $H(t)$. A frequency of 38 kc/s has been chosen for the subcarrier in this system, which is the FCC system mentioned in the introduction to this article; accordingly, the pilot signal has a frequency of 19 kc/s. The waveform impressed on the carrier, which we shall call the "*multiplex signal*" $F(t)$, thus consists of three parts:

a) the sum signal $M(t)$,
b) the stereo sub-signal $H(t)$, and
c) the pilot signal $P(t)$.

These are connected by the relation:

$$F(t) = M(t) + H(t) + P(t)$$

$$= M(t) + a\,\frac{S(t)}{S_\mathrm{m}}\,\sin \omega_\mathrm{h}t + p \sin \tfrac{1}{2}\omega_\mathrm{h}t, \quad . \quad (5)$$

where a, S_m and ω_h have the same meaning as in equation (1) and p is the amplitude of the pilot signal. If we put a equal to S_m, as before, then the multiplex signal will be given by:

$$F(t) = M(t) + S(t) \sin \omega_\mathrm{h}t + p \sin \tfrac{1}{2}\omega_\mathrm{h}t. \quad . \quad . \quad (6)$$

Fig. 3 shows the frequency spectrum of the multiplex signal. The sum signal $M(t)$ has a spectrum extending approximately from 50 c/s to 15 kc/s, while that of the stereo sub-signal $H(t)$ extends from $38 - 15 = 23$ kc/s to $38 + 15 = 53$ kc/s, leaving a blank about 100 c/s wide at 38 kc/s. The pilot signal $P(t)$ fits into the middle of the unoccupied interval, between $M(t)$ and $H(t)$. This makes it possible for the receiver to recover the pilot signal from the multiplex signal by means of a simple filter which in most cases need only consist of a simple circuit tuned to 19 kc/s. If the system employed a pilot signal having the same frequency as the subcarrier (this would mean attenuation instead of complete suppression of the subcarrier) a selective filter with a very sharp cut-off would be necessary in the receiver to separate this signal from the other components, as will be clear from inspection of fig. 3.

The amplitude of the pilot signal is so chosen that it is responsible, when modulated on to the carrier, for a frequency deviation extending over 10% of the allotted frequency sweep. An important feature of the FCC system is that both the sum signal and the stereo sub-signal can be given an amplitude corresponding to a 90% deviation. Thus the listener with a mono receiver has at his disposal a frequency deviation only 10% less than that used for a monophonic transmission occupying the same bandwidth. The stereo sub-signal can be included "at no extra cost" in terms of frequency deviation. The reason for this perhaps rather surprising fact will be clear when it is remembered that when $L(t)$ and $R(t)$ are equal, the difference signal $S(t)$, and hence also rhe stereo sub-signal $H(t)$, are zero. There can therefore be no objection to letting $M(t)$ have the full 90% of the allotted frequency sweep. (The reader is reminded that this reasoning does not apply to a system in which the subcarrier is not suppressed because this latter is always present, even when $S(t) = 0$.)

If the amplitude of one of the microphone outputs, say $R(t)$, diminishes, then $M(t)$ will become smaller and the frequency sweep thus vacated can be occupied by the stereo sub-signal $H(t)$. If $R(t) = 0$, then $L(t) = M(t) = S(t)$, and 45% of the allotted sweep can be utilized for the sum signal and another 45% for the stereo sub-signal, so that the two signals are together responsible for a 90% frequency deviation, as before. *Figs. 4a* and *c* show the shapes of the multiplex signal in the two cases just dealt with; the pilot signal has been omitted, and the left-hand and right-hand components are assumed to be sine waves.

Let us take another extreme case, that of $R(t) = —L(t)$, in which the right-hand and left-hand components are identical but *opposite in phase*. $M(t)$ is now zero and 90% of the sweep can be occupied by the

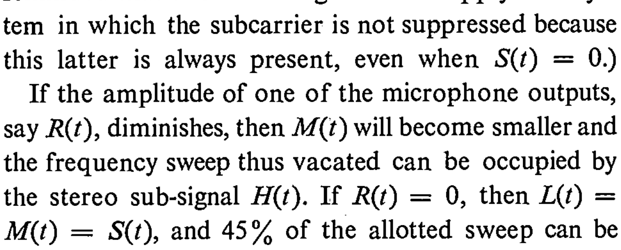


Fig. 3. Spectrum of the multiplex signal obtained with the FCC system. The diagram shows the frequency-ranges in which the various components of the transmitted signal lie, and the amplitudes it is permissible for them to attain, expressed as percentages of the maximum multiplex signal level.

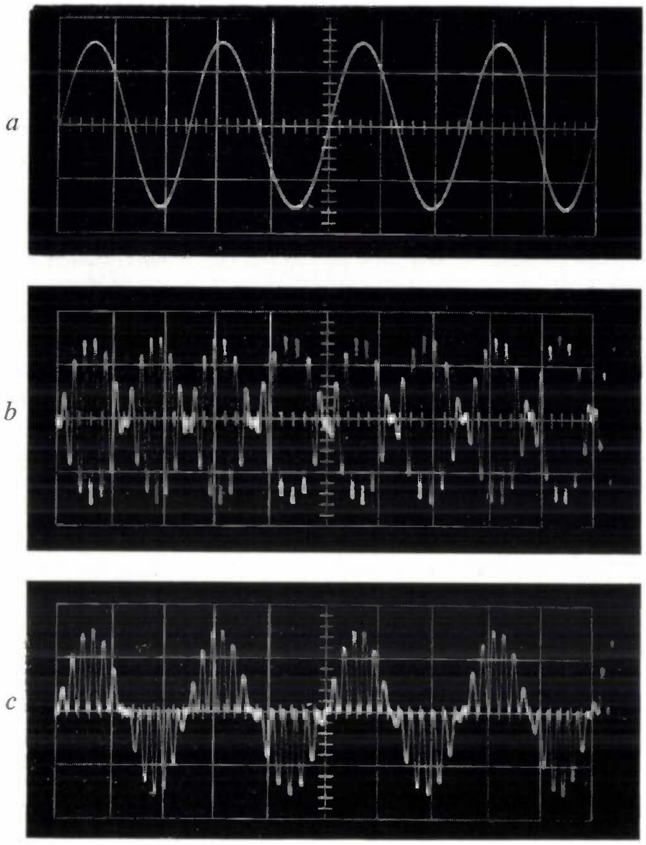


Fig. 4. Multiplex signal minus pilot signal, obtained with the FCC system. The frequency of the audio signal is 3000 c/s.
a) Left- and right-hand signals equal and in phase.
b) Left- and right-hand signals equal and of opposite phase.
c) No right-hand signal.

stereo sub-signal (fig. 4*b*). It is permissible for each of the two sidebands composing this latter to attain a value corresponding to 45 % of the maximum frequency deviation. The maxima explained above have been taken into account in fig. 3: the height of $P(t)$ corresponds to its constant amplitude, and $M(t)$ and the two sidebands composing $H(t)$ have their maximum values — all as measured against the vertical scale of relative amplitude — marked off in percentages of the maximum multiplex signal amplitude.

Fig. 5 casts further light on the nature of the FCC multiplex signal, showing the shape of its various *components* in the simple case already treated in fig. 1, in which only a left-hand signal $L(t) = b \sin pt$ is present (fig. 5*a*). As in fig. 1, the sum signal $M(t)$ and the difference signal $S(t)$ are identical with $L(t)$. Fig. 5*b* shows the stereo sub-signal; the relevant equation is:

$$H(t) = b \sin pt \sin \omega_{\text{h}}t. \quad \dots\dots \quad (7)$$

(This should be compared with equation (4) which describes a stereo sub-signal with unsuppressed subcarrier.) In fig. 5*c*, corresponding to fig. 4*c*, the sum of $M(t)$ and $H(t)$ has been plotted. Adding pilot signal

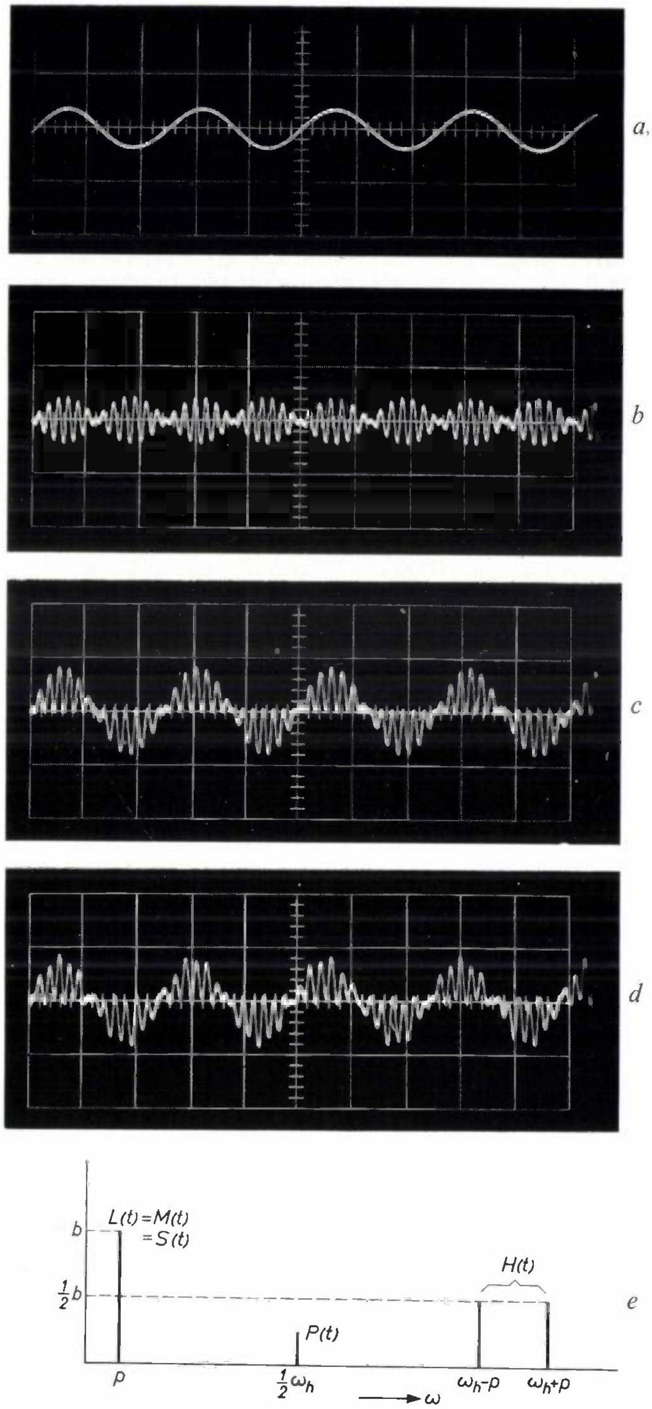


Fig. 5. Composition of a multiplex signal obtained with the FCC system and containing only a left-hand component.
a) Left-hand signal, sum signal, and difference signal $L(t) = M(t) = S(t)$.
b) Stereo sub-signal $H(t)$.
c) $M(t) + H(t)$, the multiplex signal minus the pilot signal.
d) $M(t) + H(t) + P(t)$, the complete multiplex signal.
e) Spectrum of the multiplex signal.

$P(t)$ produces the complete multiplex signal $F(t)$ in fig. 5*d*. Fig. 5*e* gives the frequency spectrum of $F(t)$.

The equation for the FCC multiplex minus pilot signal, as derived from equation 6, is:

$$F_0(t) = M(t) + S(t) \sin \omega_{\text{h}}t. \quad \dots\dots \quad (8)$$

The curves "enveloping" this multiplex signal can be found, as before, by giving successive values of $+1$ and -1 to $\sin \omega_h t$. We thus obtain:

$$F_{01}(t) = M(t) + S(t) = 2L(t), \quad \ldots \quad (9a)$$

$$F_{02}(t) = M(t) - S(t) = 2R(t). \quad \ldots \quad (9b)$$

We see that provided $a = S_m$, the enveloping curves once again have the same shape as the left-hand and right-hand signal components. But owing to the absence of a subcarrier (compare eqs. (3a) and (3b)) the envelopes cross the zero line. A received signal of this kind could not therefore be detected directly using two diodes. Use can however be made, both in the transmitter and in the receiver, of this particular feature of the envelope; we shall come back to the point when discussing circuits for the system.

A very clear picture of the multiplex signal transmitted under the FCC system can be had by considering a case in which the left-hand and right-hand components have different frequencies. In *fig. 6* they have been assumed to have frequencies of 1000 and 3000 c/s; these two sine waves appear in fig. 6*a* and *b*. Fig. 6*c* shows the multiplex signal minus pilot signal. Addition of the pilot signal results in the complete multiplex signal, shown in fig. 6*d*.

The left-hand and right-hand components can also be discerned in the envelope of the complete multiplex signal, though their presence is not quite so evident as in the trace of the multiplex signal minus pilot signal. The reason for this is as follows. When $\sin \omega_h t = 1$, and again when $\sin \omega_h t = -1$, the value of $\sin \frac{1}{2}\omega_h t$ is alternately $+\frac{1}{2}\sqrt{2}$ and $-\frac{1}{2}\sqrt{2}$, with the result that successive peaks of the complete signal are slightly raised and slightly depressed. (To see this more clearly fig. 5*d* should be compared with fig. 5*c*, and fig. 6*d* with fig. 6*c*.) At their different levels, both the raised and depressed peaks and troughs trace out the patterns of $L(t)$ and $R(t)$. (One could put it another way, and say that the curves enveloping the complete multiplex signal contain pilot signal $P(t)$ as well as the left and right components.)

The transmitter

The designer of a stereophonic radio transmitter using the FCC system can take either of two principles as his point of departure. The *first method* entails building up the multiplex waveform in exact accordance with the theoretical explanation given above. The block circuit of a transmitter of this kind can be found in *fig. 7*.

The outputs of the two microphones marked *Mi* are each fed via an amplifier *A* to a circuit usually referred to as "matrix", marked *Ma* in the diagram, which forms from $L(t)$ and $R(t)$ the sum signal $M(t)$ and the

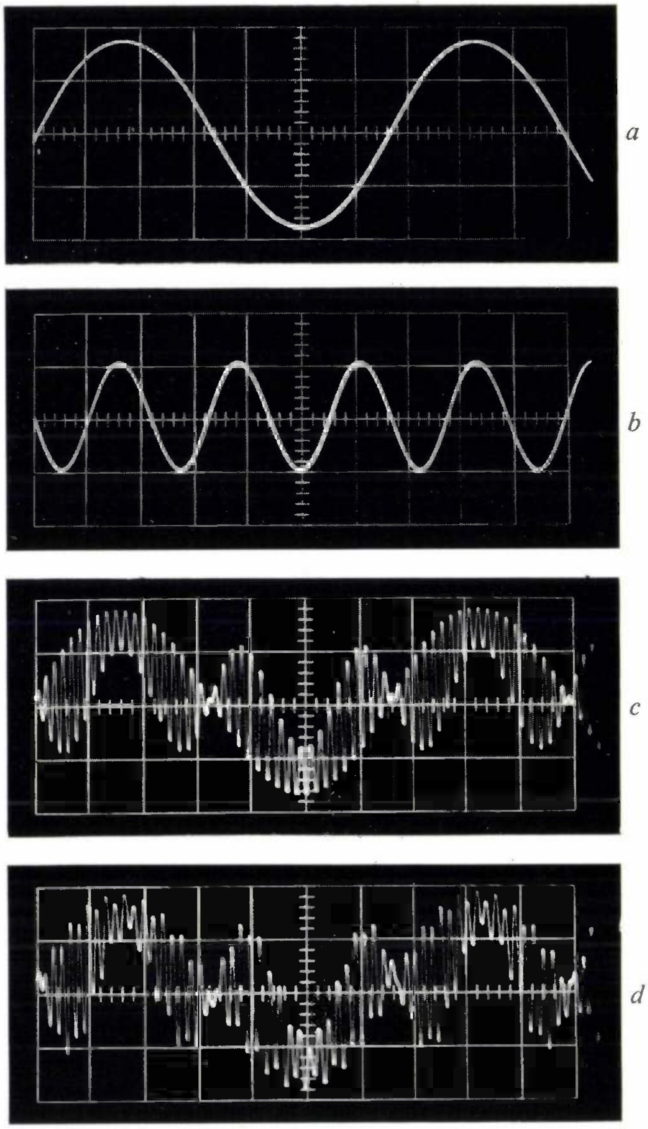


Fig. 6. Multiplex signal containing left-hand and right-hand signals differing in frequency and amplitude.
a) Left-hand signal (frequency 1000 c/s).
b) Right-hand signal (frequency 3000 c/s).
c) Multiplex signal minus pilot signal.
d) Complete multiplex signal.

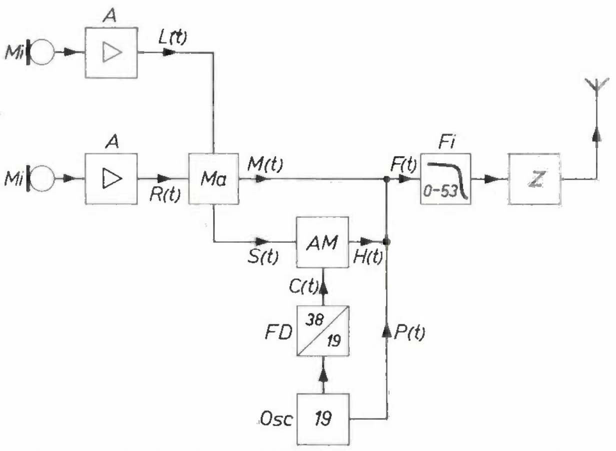


Fig. 7. Block circuit of an FCC system transmitter for stereophonic programmes. *Mi*: microphones. *A*: amplifiers. *Ma*: "Matrix". *Osc*: oscillator operating at 19 kc/s. *FD*: frequency doubler. *AM*: amplitude modulator. *Z*: transmitter. *Fi*: low-pass filter.

difference signal $S(t)$. The latter, together with subcarrier $C(t)$, is passed on to the amplitude modulator *AM*. The subcarrier has been obtained with the aid of the frequency doubler *FD* from the output of an oscillator *Osc* operating at 19 kc/s. *AM* is a balanced modulator whose output signal contains the sidebands of the subcarrier but not the subcarrier itself. In the case under consideration these sidebands constitute the stereo sub-signal $H(t)$. $H(t)$ is combined in the right proportions with the sum signal $M(t)$ and the pilot signal $P(t)$, derived from oscillator *Osc*, to form the multiplex signal $F(t)$, which reaches the modulator in *Z*, the FM transmitter, by way of a low-pass filter *Fi*. This filter removes all signals with frequencies higher than 53 kc/s — necessary because the signal delivered by the modulator stage *AM* contains higher harmonics of the subcarrier, which must not of course appear in the multiplex signal.

In stereophonic radio broadcasting, as in normal FM transmitter practice, the a.f. signals are subjected to a certain amount of pre-emphasis, that is to say, the higher audio frequencies are given rather more amplification. Pre-emphasis can be applied either to the left and right components $L(t)$ and $R(t)$ or to the sum and difference signals $M(t)$ and $S(t)$. In either case it is effected by inserting *RC* networks in the appropriate signal paths.

The circuit marked *Ma* in fig. 7 can be designed in various ways. A simple matrix circuit comprising four resistors $R_1 \dots R_4$, all of the same value, is given in *fig. 8* as an example. Clearly, if components $L(t)$ and $R(t)$ are applied to terminal pairs a - b and c - d respectively, a voltage $\frac{1}{2}M(t)$ will appear across resistors R_1 and R_4 and a voltage $\frac{1}{2}S(t)$ will appear across R_2 and R_3.

In a transmitter based on the *second method*, use is made of the fact that provided $a = S_m$ — see equations (5) and (6) — the envelope of the multiplex signal minus pilot signal will reproduce the shapes of $L(t)$ and $R(t)$. The block circuit in *fig. 9* represents one possible approach for the designer. As before, the outputs of microphones *Mi* are handled by separate amplifiers *A*. *Sw* denotes a fast electronic switch (shown in the diagram as a mechanical two-way switch, for the sake of simplicity) which links point *r* to points *p* and *q* alternately, working at a frequency of 38 kc/s. Thus the right and left components present at *p* and *q* are each sampled in turn, and a 38 kc/s square-wave voltage whose envelope follows $L(t)$ and $R(t)$ appears at point *r*. As will be clear from the foregoing, this signal contains both $M(t)$ and $H(t)$. The addition of pilot signal $P(t)$ results in the multiplex signal $F(t)$, which reaches the transmitter stages *Z* via low-pass filter *Fi*. The pilot signal is generated in the oscillator *Osc*. Subcarrier $C(t)$, which controls the electronic switch *Sw*, is derived from the same oscillator by way of the frequency-doubler *FD*.

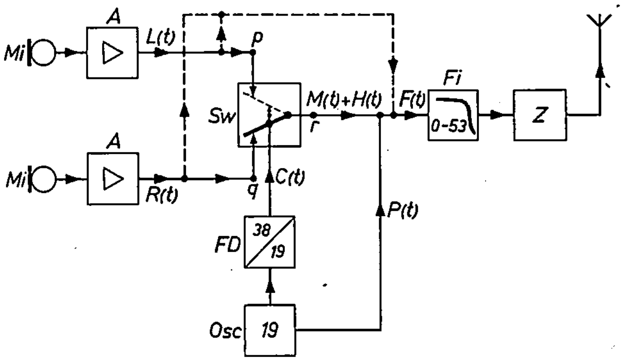


Fig. 9. Another block circuit for an FCC system stereophonic transmitter. *Sw*, shown here as a mechanical two-way switch, is in reality an electronic switching device which takes the place of the matrix and amplitude modulator appearing in fig. 7.

The square-wave voltage arising at point *r* contains harmonics of the subcarrier that must not be allowed to appear in the multiplex signal. These are removed by the filter *Fi*. As a result of this however, the output voltage from *Fi* no longer has an envelope that correctly reproduces $L(t)$ and $R(t)$. Closer investigation of this error shows that it can be corrected by injecting a weak sum signal into the output from the electronic switch. In fig. 9 the path of this sum signal is shown schematically as a broken line.

There are various methods for electronic sampling of the microphone outputs at 38 kc/s. One suitable circuit may be seen in *fig. 10*. Transistors Tr_1 and Tr_2 act as switches which are intermittently opened and closed by gating in counter-phase with a 38 kc/s voltage applied to their bases. The result is that points *p* and *q* are earthed alternately, and that the desired square-wave voltage with an envelope reproducing $L(t)$ and $R(t)$ is obtained at point *r*.

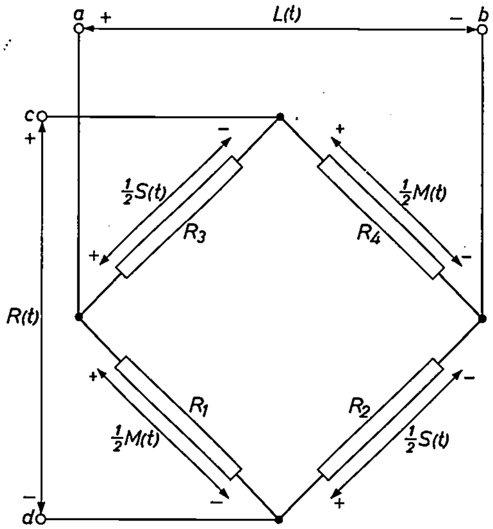


Fig. 8. Matrix circuit.

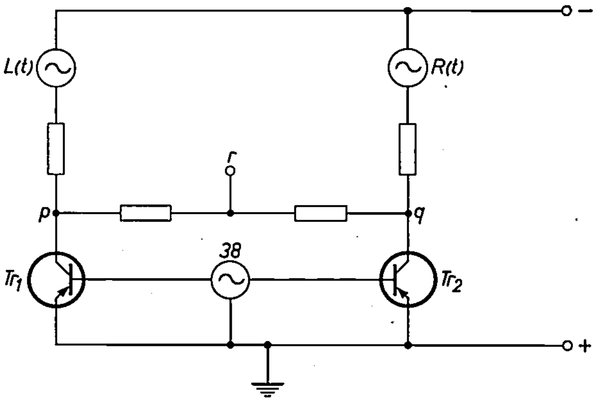


Fig. 10. Circuit capable of performing the two-way switching function required of *Sw* in fig. 9.

The receiver

The radio-frequency and intermediate-frequency sections of a receiver equipped for stereophony are the same as those of a normal FM receiver. The discriminator is rather different. Because the multiplex signal covers a spectrum 53 kc/s wide, the frequency discriminator must have a response curve (a curve showing output voltage as a function of modulation frequency at constant frequency deviation) which is flat up to 53 kc/s. The discriminator in a normal receiver only needs to have a flat response up to 15 kc/s. This requirement can be satisfied by conventional design methods and therefore we shall not go into detail here.

A circuit known as an *"adaptor"* or *"decoder"* is used to extract the left and right components from the multiplex signal available at the output of the discriminator. Subsequent amplification up to loudspeaker level must take place in two amplifiers which may or may not form an intrinsic part of the receiver.

An adaptor must satisfy certain requirements if good stereo reproduction is to be obtained; the following are the most important.

1) As in all receivers and amplifiers, there must be very little distortion; obviously, this applies equally to the left-hand and to the right-hand signals.
2) Cross-talk must be kept to a minimum; in other words, there must be efficient separation of the left and right signals.
3) Susceptibility to any form of interference must be reduced to a minimum.

In addition to these requirements a number of desiderata can be listed — features which, though not essential to a satisfactory technical performance, may considerably facilitate the operation of the receiver. Some of these desirable features are listed.

4) The user should be able to see from a visual indicator that he has picked up a station transmitting a stereophonic programme.
5) On retuning to another station, the receiver should switch automatically from stereo to mono operation or vice versa, as appropriate.
6) The volume of the loudspeaker output should be about the same for mono and stereo reception.

Various principles can be taken to form a basis for the design of an adaptor circuit. The most obvious method is suggested by the composition of the multiplex signal and consists in breaking down the multiplex signal into its constituents by means of filters. Sum signal $M(t)$, stereo sub-signal $H(t)$ and pilot signal $P(t)$ having been recovered in this manner, $P(t)$ can be used to restore the sub-carrier and the subcarrier and stereo sub-signal can be added together and fed to a detector which yields the difference signal $S(t)$. Finally, the left and right components can be obtained from $M(t)$ and $S(t)$ by addition and subtraction.

The block circuit of an adaptor working on this principle may be found in *fig. 11.* The multiplex signal $F(t)$ is split by filters Fi_1, Fi_2 and Fi_3 into pilot signal $P(t)$, stereo sub-signal $H(t)$ and sum signal $M(t)$. The pilot wave undergoes frequency-doubling in *FD*, yielding the subcarrier $C(t)$ which, together with the stereo sub-signal, is fed into the amplitude detector *AD*. The difference signal $S(t)$ is available at the output of *AD*. A matrix *Ma* similar to that discussed when the transmitter was dealt with (fig. 8) can be employed for adding and subtracting $M(t)$ and $S(t)$.

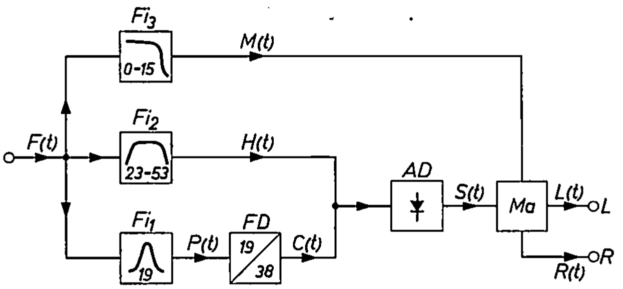


Fig. 11. Block circuit of an adaptor for extracting the left-hand and right-hand components from the multiplex signal. Filter Fi_1 passes the pilot signal, Fi_2 the stereo sub-signal and Fi_3 is a low-pass filter for the sum signal. *FD*: frequency doubler. *AD*: diode detector. *Ma*: matrix. *L* and *R* are terminals for left-hand and right-hand signals.

In another kind of adaptor use is made of the fact that the envelope reproduces the shapes of $L(t)$ and $R(t)$ provided $a = S_m$ (see equations (5) and 6)). It has already been pointed out that direct detection using two diodes is not feasible in the absence of a subcarrier. To get around this difficulty a local subcarrier is generated with the aid of the pilot signal and added to the multiplex signal (from which the pilot signal may or may not have been extracted). The block circuit of an adaptor working on this principle is given in *fig. 12.* Filter Fi_2 is shown in broken outline, indicating that suppression of the pilot wave is optional [7]. The pilot frequency is only a little higher than the higher audio frequencies, and may give rise to some slight distortion of these at the detector; there is therefore something to be said for inserting a filter to suppress $P(t)$.

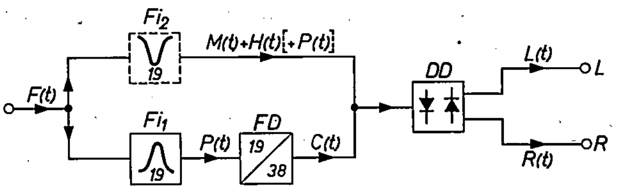


Fig. 12. Block circuit of an adaptor working on the principle of "envelope detection". Filter Fi_1 passes the pilot signal; Fi_2 suppresses the pilot signal. DD: double diode detector.

The use of two diodes to detect the envelope is not the only way of retrieving the left and right components from $C(t)$ and $M(t) + H(t)$. Instead of being combined with of the multiplex signal, the subcarrier $C(t)$ can be used to effect two-way switching of $M(t) + H(t)$ at a frequency of 38 kc/s. For the sake of simplicity the switching function is again represented in *fig. 13* by an ordinary two-way switch *Sw*. The result of the switching is that the left and right components appear at terminals p and q, as may be inferred from the explanation, given above, of the sampling process at the transmitter. The reversal of this process constitutes, in fact, a kind of detection which is often called "switch detection".

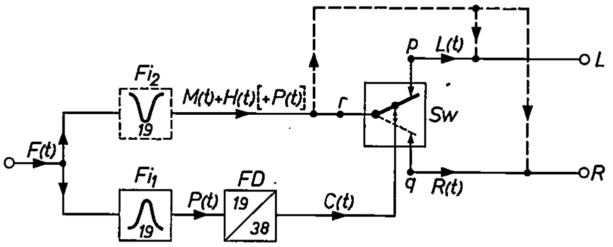


Fig. 13. Block circuit of an adaptor using switch detection. Filter Fi_1 passes the pilot signal, Fi_2 suppresses the pilot signal. *Sw* represents the switch detector, which in practice is an electronic switch.

Closer study of this method of detection reveals that the voltages it yields are not an exact copy of the incoming modulation envelope. The deviation can be compensated by subtracting an $M(t)$ signal of low amplitud from the voltages delivered by the switch detector; in other words, by adding a weakened version of $M(t)$ in opposite phase. In fig. 13 the relevant signal paths are shown as broken lines. (In practice $M(t) + H(t)$ is subtracted, but the correction is not affected by the presence of the stereo sub-signal $H(t)$, whose frequency range is well outside that of the audio frequencies to which the ear responds.) All in all, the correction arrangements are the counterpart of those incorporated in a transmitter of the type shown in fig. 9.

If desired, the pilot signal can be extracted from the multiplex signal before this is injected at point r in the switch detector.

There is another way of utilizing the principle of switch detection. Instead of presenting the switch detector with the complete multiplex signal, or the multiplex minus pilot signal, as in fig. 13, the stereo sub-signal can be isolated and subjected to the switching process. Since $H(t)$ is equal to $S(t) \sin \omega_{\rm h}t$, its envelope reproduces the two functions $S(t)$ and $-S(t)$ — see for example fig. 4b — and two-way switching at 38 kc/s will therefore yield two difference signals of opposite polarity. The relevant block circuit may be found in *fig. 14*. As in fig. 11, the multiplex signal $F(t)$ is split into its three constituents $P(t)$, $H(t)$ and $M(t)$. Frequency-doubler *FD* converts pilot signal $P(t)$ into a subcarrier $C(t)$ which controls the "switching action" of *Sw*, so that point r is linked first to point p and then to point q. The difference signal is now available in opposite phase at p and q. By adding the sum signal $M(t)$ the left and right components $L(t)$ and $R(t)$ are obtained.

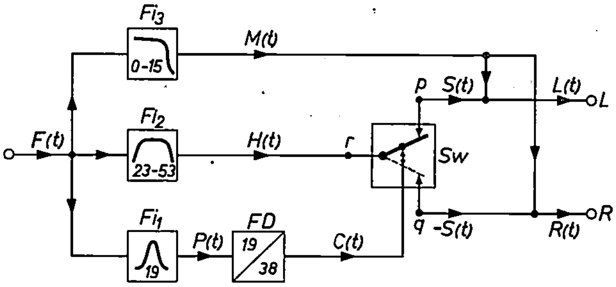


Fig. 14. Block circuit of an adaptor in which switch detection is applied only to the stereo sub-signal. Filter Fi_1 passes the pilot signal, Fi_2 the stereo sub-signal and Fi_3 the sum signal. *Sw* represents the switch detector, which in practice is an electronic switch.

Figs. 13 and 14 have much in common with figs. 12 and 11 respectively, the only difference being in the detection method (switch detectors being used in place of envelope detectors in the two designs just discussed). It can be mentioned, without going into details, that the advantage of a switch over a diode detector is its reduced susceptibility to certain forms of interference.

De-emphasis

Since pre-emphasis is invariably applied in FM transmitters the a.f. signals recovered by the receiver must undergo a certain amount of "de-emphasis". For this purpose, both the left and right components can be passed through an *RC* network before leaving the adaptor at output terminals *L* and *R*. In the circuits

[7] Complete or partial suppression of the pilot signal in the residual signal is automatically effected in some circuits by means of the filter Fi_1 passing the pilot frequency.

appearing in figs. 12 and 13, this is the only possible method. In circuits which generate sum and difference signals (figs. 11 and 14) an alternative method is to arrange for these signals to undergo de-emphasis, by inserting an RC network in the lines marked $M(t)$ and $S(t)$. If this is done it must be borne in mind that the higher frequencies in both the sum and difference signals have already suffered a certain amount of attenuation in the filters which extract these signals from the multiplex (they are marked Fi_3 and Fi_2 in both fig. 11 and fig. 14). This fact can however be turned to good account by arranging for *full* de-emphasis to take place in these filters. An RC network of the appropriate time constant for de-emphasis (50 μs) must then be used for Fi_3, and for Fi_2 use will be made of an LC circuit tuned to 38 kc/s, with a Q factor such that both sides of its resonance curve fall off in the same way as the bandpass characteristic of a network producing the desired attenuation of higher audio frequencies.

A big advantage of introducing a filter Fi_2 to separate $M(t)$ and $H(t)$ is that it prevents any interference at frequencies roughly equal to those of the subcarrier harmonics from reaching the amplitude detector (AD in fig. 11) or the switch detector (Sw in fig. 14). Components having these frequencies may be present in noise or other interfering signals, but they may also be generated internally, due to curvature in the characteristic of the frequency discriminator of the receiver. In the absence of preventive measures incoming signals of approximately 76 kc/s, 114 kc/s etc. would generate a.f. voltages in these detectors. One way of obviating this would be to precede the detector by a filter that would only pass signals within the desired frequency range of 0 to 53 kc/s. But a fairly complicated filter would be required to produce high attenuation of signals with frequencies outside this range without involving any amplitude or phase distortion of signals within it. A much simpler course is to let this attenuation take place in filter Fi_2. This then has to be rather selective, but this can be put to good use in de-emphasis.

Fuller description of one type of adaptor

To conclude this article we shall give a brief account of an adaptor equipped with semiconductor elements; the circuit diagram appears in *fig. 15.* The principle underlying the circuit is that of fig. 14. The multiplex signal enters at terminal K_m. The filters marked Fi_1 and

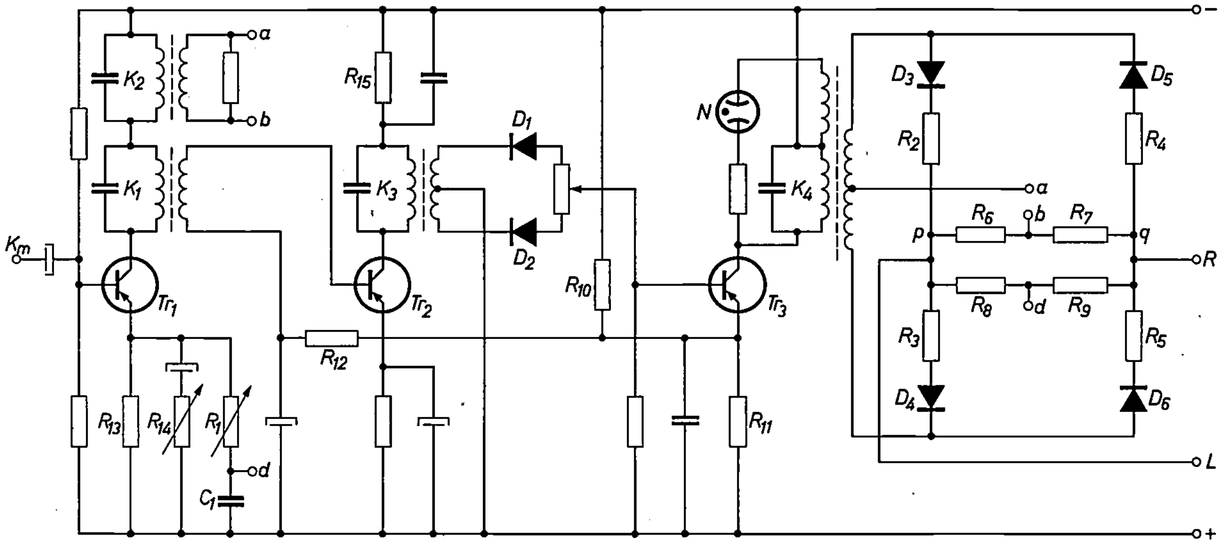


Fig. 15. Circuit diagram of adaptor working on the principle of fig. 14. K_m: input terminal. K_1 and K_3: tuned circuits whose resonant frequency is 19 kc/s. K_2 and K_4: tuned circuits whose resonant frequency is 38 kc/s. N: miniature neon indicator.

Derived from K_1 by inductive coupling, the pilot signal is applied to transistor Tr_2 and amplified. LC circuit K_3, which is also tuned to 19 kc/s, is inductively coupled to a full-wave rectifier formed by diodes D_1 and D_2. This yields (in addition to the direct voltage discussed in the text) the 38 kc/s subcarrier which, having undergone amplification in transistor Tr_3, is fed into oscillatory circuit K_4 tuned to that frequency. The subcarrier is transferred by inductive coupling to the part of the circuit consisting of diodes $D_3 \ldots D_6$ and a number of resistors, which is responsible for the intermittent switching of the stereo sub-signal. Diode pairs D_3-D_4 and D_5-D_6 are arranged with opposite polarity; the result is that the current paths through $R_2 + R_3$ and $R_4 + R_5$ are opened and closed intermittently at the frequency of the subcarrier. The stereo subsignal derived from K_2 is present between points a and b in this part of the diagram (they are connected to the corresponding points on the left) and so the periodic switching referred to causes a current to flow through R_6 and R_7 alternately. Hence voltage shapes which copy the envelope of the stereo sub-signal, and so can be represented by $S(t)$ and $-S(t)$, are available between points p and b and between points q and b respectively. Two resistors of the same value, R_8 and R_9, have been inserted between p and q. Consequently voltages $S(t)$ and $-S(t)$ are also available between p and d, and q and d. The sum of $M(t)$ and $S(t)$, and that of $M(t)$ and $-S(t)$, are obtained in virtue of the fact that sum signal $M(t)$ is fed in between point d and earth, having been derived from the emitter voltage of transistor Tr_1 and passed through the filter and de-emphasis network R_1-C_1. This therefore has the result that the two signal voltages $M(t) + S(t) = 2L(t)$ and $M(t) - S(t) = 2R(t)$ are available between point p and earth and between point q and earth respectively. Points p and q are accordingly connected to the output terminals L and R.

Fi_2 in fig. 14 are represented here by the LC circuits K_1 and K_2 in the collector lead of transistor Tr_1. As has already been explained, these circuits are tuned to 19 kc/s and 38 kc/s respectively. The sum signal is derived from the emitter loop; the low-pass filter marked Fi_3 in fig. 14 is formed here by a resistor R_1 and a capacitor C_1. The frequency doubler FD is made up of two diodes D_1 and D_2, and the switch detector Sw is constituted by diodes $D_3 \ldots D_6$ and a number of resistors.

Potential divider R_{10}-R_{11} serves to bias the emitter of Tr_3 and, via R_{12}, the base of Tr_2 with negative voltages such that the gain of the latter transistor is small, and no collector current flows in Tr_3 if no pilot signal is present. The absence of a pilot signal means that the receiver is tuned to a *non-stereophonic* broadcast. In these circumstances the stereo side of the adaptor circuit is not operative, and the a.f. signal from the emitter of Tr_1 is passed via R_1, R_8 and R_9 to the output terminals, the two outputs being in phase. If a *stereophonic* transmission is picked up, its pilot wave will generate in the full-wave rectifier D_1-D_2 a negative voltage which is small to start with, but enough to cause a flow of emitter and collector current in Tr_3. The resulting rise in the direct voltage across R_{11} causes Tr_2 to provide more gain, and this in its turn drives the base of Tr_3 more negative. The effect is cumulative, and continues up to the point where the gain of Tr_2 and Tr_3 is as large as it can be under the conditions prevailing. (Resistor R_{15} in the collector circuit of Tr_2 is sufficiently large that this transistor then limits.) The adaptor is now adjusted to stereo operation.

So long as the pilot signal amplitude is below a certain low level, transistor Tr_3 will remain non-conducting and the receiver will be adjusted to mono reception. There is in fact a threshold for the changeover to stereo operation. This ensures that the adaptor will not switch over to stereo because of noise or other interference at a frequency of about 19 kc/s. (If it did, the noise level of the incoming mono signal might well be raised.)

Once Tr_2 and Tr_3 have finally switched to the "on" state the pilot signal amplitude has to drop well below the threshold value before the circuit changes back to mono operation, as will be clear from *fig. 16*, in which the subcarrier V_c present at K_4 has been plotted as a function of V_p, the pilot signal entering the adaptor. The threshold value here is 100 mV, but the incoming signal must drop to 50 mV before the adaptor can change over from stereo to mono operation. The purpose of this hysteresis is to prevent unwanted switching back to mono operation due to small changes of signal strength when a weak stereo signal is being received, and vice versa.

The inductor in tuned circuit K_4 has a few extra turns to supply power for a miniature neon indicator N, which lights up when a subcarrier is present, giving visual indication that a stereophonic transmission has been picked up.

As has been explained above, de-emphasis can be achieved in this circuit by appropriate design of tuned circuit K_2 and low-pass filter R_1-C_1. This means that the band-pass curve of K_2 must have such a shape that the presence of a weak sum signal between points a and b, in addition to the stereo sub-signal, cannot be avoided. Now, if this weak sum signal were able to leak through to the adaptor output terminals it would cause crosstalk at high audio frequencies, so upsetting the stereophonic sound image. The disturbing effect would be all the greater as the signal in question has not yet undergone de-emphasis. However, it is impossible in the circuit of fig. 15 for the sum signal reaching the switch detector via K_2 to make any contribution to the adaptor output voltages. It is true that the sum signal is present between points p and b and between points q and b, but the voltages at p and q with respect to b are equal and *in phase*. They cannot therefore be responsible for any flow of current through R_8 and R_9, or for any voltage between p and d or q and d; nor, therefore, can they contribute to the signal between the output terminals and earth.

The circuit could have been simplified by omitting resistors R_8 and R_9 and connecting de-emphasizing filter R_1-C_1 direct to point b. This would however have led to the possibility of the difficulties mentioned above: cross-talk would have been liable to occur, with an adverse effect on stereo reproduction quality.

An unwanted signal, mainly consisting of high audio frequencies that have not undergone de-emphasis, likewise arises between a and b when a mono transmission is being received. In this case also, the voltages between p and b and between a and b have the same phase, so that the unwanted signal cannot contribute anything to the adaptor output voltages.

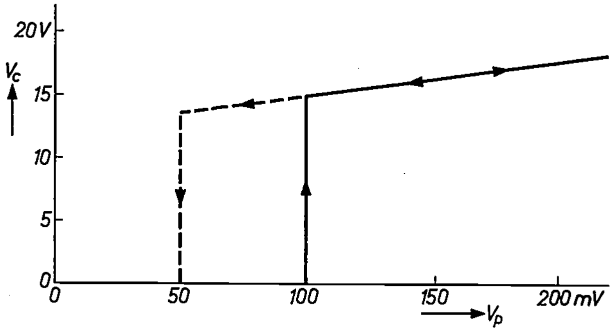


Fig. 16. The subcarrier V_c present in K_4 as a function of that of the pilot signal V_p entering the input terminal of the adaptor. The full line relates to increasing signal-strength, the broken line to decreasing signal-strength.

Reference to the circuit diagram reveals that the emitter circuit of Tr_1 contains, in addition to the resistor R_{13} responsible for d.c. biasing, a variable resistor R_{14} (which is in series with a high value capacitor). This is used to adjust the gain provided by Tr_1, and so to raise or lower the level of the stereo sub-signal, as supplied to the switch detector, to a value such that the left-hand and right-hand components of the difference signal delivered by the detector have the same amplitudes as those of the sum signal supplied to point d. A lack of balance in these components makes it impossible to retrieve the left-hand and right-hand information by adding and subtracting $M(t)$ and $S(t)$. The condition here that must be satisfied is $a = S_m$, as laid down in the introduction to this article. Failure to fulfil this condition may give rise to cross-talk. Crosstalk may also occur, particularly at higher frequencies, if the sum and difference signals have not undergone the same amount of de-emphasis (for this too involves an incorrect proportion between the components of the sum signal and the difference signal). The necessary adjustment has been provided for by making R_1 a variable resistor. When the adaptor is being pre-adjusted the setting of R_1 is so chosen that cross-talk at higher frequencies is minimized; on the other hand, in the adjustment of R_{14} attention is paid to the lower frequencies (unaffected by de-emphasis).

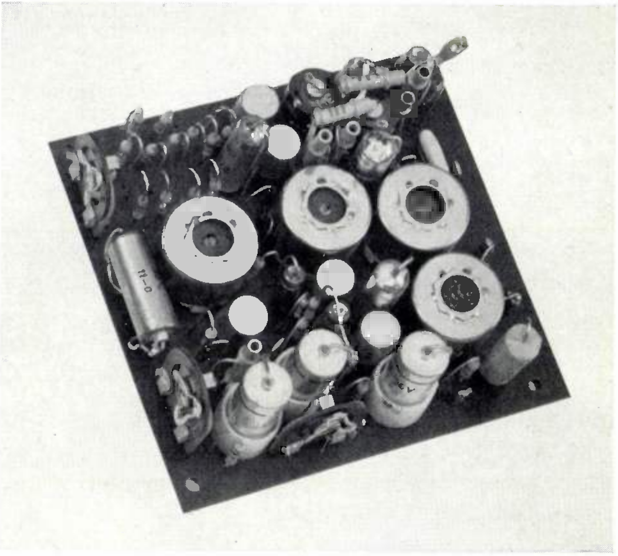


Fig. 17. An adaptor following the circuit diagram of fig. 16, using the printed wiring technique.

A practical circuit based on fig. 15 can be made in a very compact form. *Fig. 17* shows a version built on a printed-wiring board measuring 8 × 8.5 cm.

Summary. A stereophonic radio broadcasting system should be "compatible", that is to say, it must be possible for the signals transmitted under the system to be satisfactorily handled by a receiver that is *not* equipped for stereophony. Instead of transmitting the left-hand and right-hand microphone outputs separately the practice is to transform them into a sum and a difference signal. Under the FCC system, which is already being employed in several countries, including the U.S.A., the difference signal is impressed on a 38 kc/s subcarrier by amplitude modulation. The subcarrier itself is suppressed; its sidebands, together with the sum signal, are frequency-modulated on to the main carrier. This means that the subcarrier has to be restored on the receiving side. The recovery process is aided by transmitting a pilot signal whose frequency is 19 kc/s. Under the FCC system, receivers not equipped for stereophony receive the sum signal, which has a maximum frequency deviation covering 90% of the frequency sweep available for a non-stereophonic broadcast. The "multiplex signal" reaching the transmitting station can be produced in two different ways. A balanced modulator can be employed for modulating the subcarrier or alternatively, the left-hand and right, hand signals can be sampled in turn by switching at a frequency of 38 kc/s. There are likewise various techniques available for recovering the left-hand and right-hand signals from the multiplex signal picked up by the receiver. In some systems a "switch detector" is used, certain components of the incoming signal being subjected to a two-way switching process. The article concludes with an account of an adaptor circuit working on this latter principle.

A silicon carbide mortar

W. F. Knippenberg, G. Verspui and J. Visser

542.222

If a solid substance has to be chemically analysed, it is often first necessary to grind it to powder in a mortar — either so that it will dissolve quickly in the case of a wet analysis, or in order to make certain spectral analyses possible.

It makes a difference here whether one wishes to analyse relatively large amounts, or only traces. For trace analysis, the chance of the introduction of impurities from the mortar is often a reason for not grinding the substance; and if it is necessary to do so the mortar must be chosen with very great care for the reason just mentioned.

The analytical chemist normally has a choice between mortars of porcelain, agate and various metal carbides for the grinding of solid substances [1]. In general, mortars of these kinds are not sufficiently hard and resistant to abrasion for the detection of trace elements. The recently available mortars of aluminium oxide and boron carbide — which can now be sintered to almost the theoretical density — are a great improvement in this respect, and have become practically indispensible for trace analysis [2].

However, for the grinding of very hard materials, even these mortars are not really good enough: too much material is still removed from the mortar to allow certain trace analyses to be carried out with success.

It the course of investigations in this laboratory on silicon carbide, long known as a hard material (carborundum), we have succeeded in making a mortar of this material. This mortar was originally mainly of use for our own investigations [3]: it is an ideal situation to be able to grind a substance in a mortar made of the same (pure) material — a principle one would like to apply more often, were it not that for most substances it remains impossible owing to the associated technical difficulties.

It was later found that this SiC mortar also has very favourable properties for general analytical purposes, apart from the investigation on SiC. Silicon carbide is not only hard and resistant to abrasion, but also chemically resistant, and we have moreover succeeded in obtaining it in a very pure form. For wet analyses, the chemical resistance is the most important, since it means that the little SiC which is worn off will not interfere with the analysis. For the spectral analyses, it is rather the high purity of the SiC material which makes the mortar useful. Only the silicon (and in exceptional cases also the carbon) then occurs as an impurity in these analyses. (Since the substance under investigation in spectral analyses is normally evaporated between two carbon electrodes, the carbon of the mortar is seldom a factor of importance; it only makes its presence felt when other materials are used for the electrodes.) It is also worthy of mention that the high chemical resistance of silicon carbide allows a mortar made of this material to be very thoroughly cleaned. Contamination from substances left over from a previous analysis can thus generally be avoided.

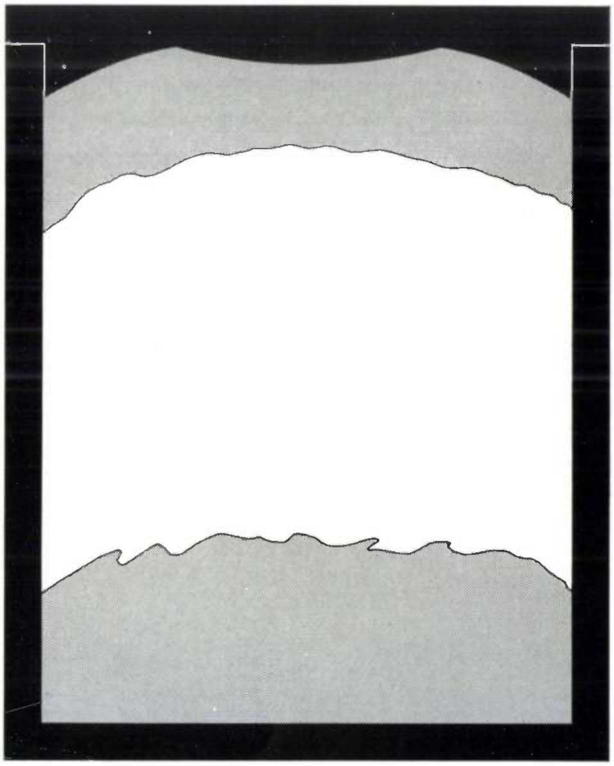


Fig. 1. Making a silicon carbide mortar in a carbon cylinder. Silicon carbide crystallizes out from silicon and carbon-containing vapour on the lid, which has the form of the negative of the desired mortar.

It will be clear, then, that this mortar forms a very welcome addition to the usual assortment of mortars in an analytical laboratory; together with the mortar of pure boron carbide, which will still be needed for

Dr. W. F. Knippenberg, G. Verspui and J. Visser are research workers at Philips Research Laboratories, Eindhoven.

the determination of silicon, and the aluminium oxide mortar in case one also wishes to determine carbon, it makes trace analysis of all elements possible.

The mortar is made of polycrystalline silicon carbide of high purity (the impurities amounting at the most to a few hundred-thousandths of a percent). This material is obtained by the thermal decomposition of methyl dichlorosilane in the presence of hydrogen [4]. The silicon carbide thus obtained is heated in a carbon crucible at 2600°C in an inert gas atmosphere. At this temperature, the silicon carbide decomposes into carbon (solid and vapour) and silicon (vapour).

The crucible is covered with a carbon lid in the form of a negative of the desired mortar. The temperature of this lid is maintained between 2400 and 2500 °C. At this temperature silicon carbide is deposited on the lid, while recrystallization processes ensure a high density of the deposited layer. Relatively large single crystals occur in this layer, about 0.1 cm thick and with an area of about 1 cm^2. *Fig. 1* shows a section through the carbon lid with the silicon carbide deposited on it.

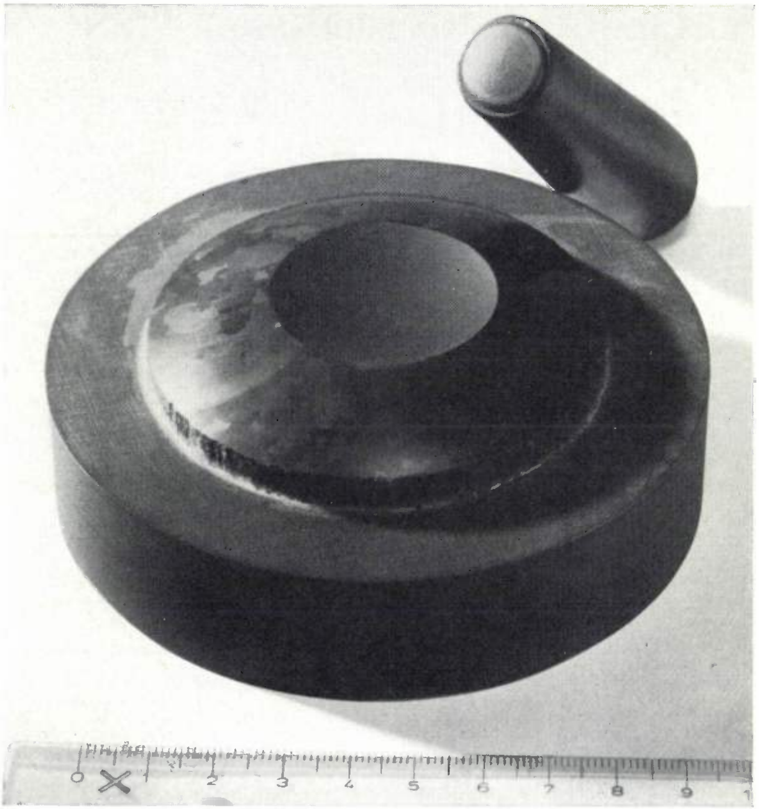


Fig. 2. A mortar of silicon carbide in a plastic base. The relatively large single crystals of which the mortar is built up can be seen at the left-hand edge. The pestle is also of silicon carbide (with a plastic handle).

When the layer has reached the desired thickness, the carbon lid is burnt and the mortar finished with diamond powder. *Fig. 2* shows a SiC mortar cemented into a plastic base.

The pestle for use with the mortar can be made in a similar way. A certain amount of silicon carbide is "sublimed" on to a flat carbon plate as described above, and the pestle is ground out of a rough lump of the material obtained, again with the aid of diamond powder.

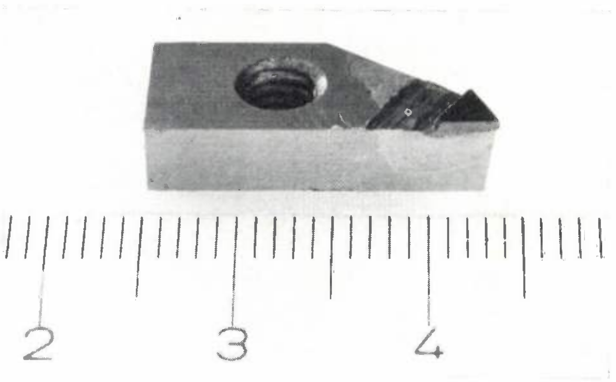


Fig. 3. A silicon carbide cutting tool ground from a single crystal (the triangle at the tip). The chisel is welded on to a molybdenum rod with the aid of a gold-tantalum alloy at a temperature of 1300 °C.

The pure "sublimed" silicon carbide is also used as a material for cutting tools. Such tools are needed for the turning or milling of materials which have to be analyzed in shaving form, and which must not be contaminated with the usual materials of which chisels are made. The SiC tools, like the pestles, are ground to the desired shape. The starting material can be a polycrystalline fragment, or if so desired one of the component single crystals, the single crystals being large enough for this purpose. *Fig. 3* shows a tool where the SiC head is welded on to a molybdenum rod with the aid of a gold-tantalum alloy.

[1] See e.g. O. G. Koch and G. A. Koch-Dedic, Handbuch der Spurenanalyse, Springer, Berlin 1964.

[2] See e.g. N. W. H. Addink, Chem. Weekblad **56**, 622, 1960 and A. Claassen, Chem. Weekblad **58**, 33, 1962.

[3] W. F. Knippenberg, Philips Res. Repts. **18**, 251, 1963.

[4] W. F. Knippenberg, Philips Res. Repts. **18**, 205, 1963.

Surface-wave transmission lines for microwave frequencies

I. The various types of transmission line
II. Applications of the dielectric line

621.372.826

In microwave techniques, the millimetre and submillimetre wavebands are coming to acquire a practical importance comparable to that at present occupied by the centimetre waveband. Components and measuring techniques for these extremely high frequencies are often a result of the marriage of known principles from transmission lines and optics.

A special problem is the low-loss transmission of millimetre and submillimetre waves. There are various surface-wave transmission lines that can be used for this purpose. Since the transmission of the electromagnetic energy chiefly takes place in the space surrounding the line, not only the attenuation of these lines but also the radial extent of the field has to be taken into account.

The dielectric line, on the basis of theoretical and practical results so far obtained, seems to be the most appropriate surface-wave line for transmitting electromagnetic energy in the 1 mm wavelength range.

I. The various types of transmission line

H. Severin

In comparing various forms of transmission line for electromagnetic waves with their practical applicability in mind, two classes of problem must be taken into account. The first relates to the transmission properties of the straight infinitely long line. Maxwell's equations have to be solved for the special boundary-values, giving the field configuration, the phase velocity and the attenuation of the wave as a function of guide dimensions and frequency.

The second class of problem is of a more practical nature. It concerns, for example, the possibilities of exciting a certain mode, of matching various lines, of constructing components and measuring devices, such as attenuators, directional couplers, slotted lines, etc.; and also, for surface-wave lines, the method of mounting and the avoidance of radiation losses at bends. The theoretical and the practical problems are certainly of equal importance in deciding which type of transmission line is the most favourable for a particular application. Nevertheless the theoretical problem can be said to be of primary interest, because one cannot attempt to treat the more practical questions until the transmission properties are known. Part I of this paper deals exclusively with the first class of problem, giving a survey of the various types of transmission line for surface waves. Part II follows with a survey of the more practical questions for the dielectric line, which the theoretical work showed to be the most promising.

By way of introduction a short summary is given of the properties of the conventional transmission lines for high frequencies. For decimetre and centimetre waves coaxial cables and waveguides of rectangular cross-section are chiefly employed at present [1]. The attenuation of these transmission lines increases with the square root of the frequency because the skin depth decreases with frequency [2], so that there are already considerable losses at centimetre wavelengths. Since, on the other hand, the attenuation decreases with increasing surface it can — to a certain extent — be reduced by enlarging the cross-section of the transmission line. This is one of the reasons for the preferred application of waveguides for centimetre waves. At a frequency of 10 Gc/s — corresponding to a wavelength of 3 cm — the attenuation of a standard copper coaxial line without dielectric (outer conductor diameter 9.53 mm, inner conductor diameter 2.65 mm) is 226 dB/km, whereas the attenuation of the standard copper rectangular waveguide for the 10 Gc/s frequency-band (inside dimensions 22.9 mm × 10.2 mm) is 96 dB/km. At a wavelength of 5 mm, the waveguide for the 60 Gc/s frequency-band (inside dimensions 3.76 mm × 1.88 mm) has an attenuation of 1300 dB/km.

In view of these attenuation values, the coaxial line and conventional waveguides are not suitable for longer distance transmission, because, with the present state of amplifier technique, the attenuation for sline transmission must be no greater than 3.5 dB/km.

Prof. Dr. H. Severin, formerly a research worker at the Hamburg laboratory of Philips Zentrallaboratorium GmbH, is now Professor of High Frequency Techniques at the University of Bochum.

Even for short connections, e.g. antenna feeders, the attenuation is considerable, especially at high frequencies.

The idea of increasing the conducting surface takes us from the coaxial cable to the "Clogston-type" or "laminated cable" in which the inner and outer conductors consist of many concentric metal layers insulated from each other [3]. If the thickness of the individual layers is small compared with the skin depth at the highest frequency to be transmitted, such a cable has a constant attenuation over a wide frequency range. In addition, the attenuation is much smaller than that of a coaxial line of the same size. The laminated cable will perhaps find application as a broad-band cable in the frequency range from 0.1 to 10 Mc/s, for example in carrier-frequency telephony or for television programme transmission, as soon as the difficulties of economic production are overcome.

For waveguides, increasing the conducting surface to reduce the losses leads to larger waveguide cross-sections. Apart from being expensive, this gives rise to difficulties because of the simultaneous excitation of unwanted higher-order modes, whose number grows with increasing cross-section. The H_{01}-mode in circular waveguide proves to be superior to all other waveguide modes because for equal attenuation it requires the smallest guide cross-section, and is thus accompanied by the smallest possible number of unwanted modes. While for waveguide modes at high frequencies the attenuation due to ohmic losses usually increases with the square root of the frequency, the H_{01}-mode shows a completely different behaviour: for this mode the attenuation approaches zero with increasing frequency. The ohmic losses arise exclusively from circular wall currents due to the axial component of the magnetic field. Since the magnitude of the axial field component decreases with increasing frequency for all waveguide modes, the wall currents for the H_{01}-mode and hence the ohmic losses decrease with frequency.

An immediate consequence of this behaviour is the fact that in principle the attenuation at any frequency can be made as small as desired if the diameter of the waveguide is chosen large enough [4]. For example, at a carrier frequency of 50 Gc/s, corresponding to a wavelength of 6 mm, an attenuation of about 1.25 dB/km (13% reduction in amplitude), which is typical for microwave links, is obtained by using a copper waveguide of 5 cm diameter.

The following example illustrates the superiority of the H_{01}-mode for circular waveguide compared with the dominant H_{11}-mode. For the required attenuation of 1.25 dB/km the diameters required for the two modes at a frequency of 10 Gc/s are 11.2 cm and 43 cm respectively, and the number of possible modes is 40 for the H_{01} mode and more than 300 for the H_{11}-mode.

The low attenuation of the surface wave transmission lines treated below is due to the fact that the electro-magnetic field extends appreciably into the surrounding space. In order to judge the applicability of these lines it is therefore not sufficient to know their attenuation: the extent of the electro-magnetic field around the line must also be known. A line supporting a wave whose field extends considerably into the surrounding space is unsuitable for practical purposes, even if its attenuation is extremely low.

The single-wire transmission line (Sommerfeld line)

The problem of propagating electromagnetic waves along a *single* wire ("single-wire transmission line") was solved by Sommerfeld in 1899. Heinrich Hertz had previously treated this problem without success. We know today that his experimental investigation failed because of the large extent of the electro-magnetic field around the wire. Because of the large field extent the surroundings, such as the laboratory walls, disturbed the field configuration. Theoretically Hertz idealized the problem too much: he took the wire to be infinitely thin and to have an infinite conductivity σ. Sommerfeld's solution shows that in the limiting case $\sigma \to \infty$ a surface wave along the wire cannot exist.

For single-wire transmission line one can make the same distinction as for waveguides between H- and E-modes, in which either the magnetic field H or the electric field E respectively has a component along the direction of the wire. Sommerfeld's solution refers to the axially symmetrical E-mode, whose field is independent of φ and consists of the components E_r, E_z and H_φ if the wire coincides with the z-axis of a system of cylindrical coordinates r, φ, z (*fig. 1*). For all other E- and H-modes the major part of the field is inside the wire, as Hondros later showed. The attenuation of these

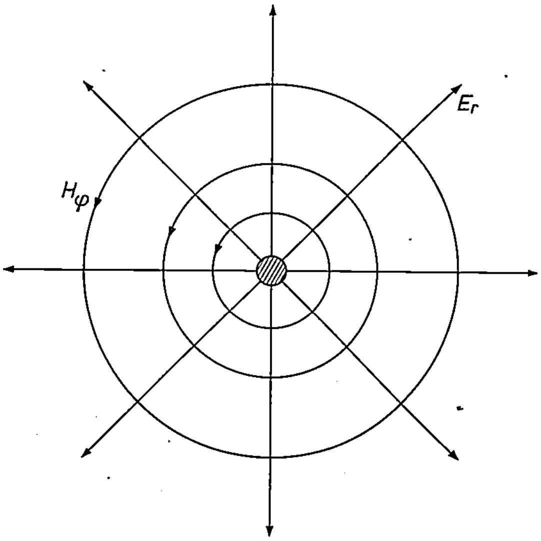


Fig. 1. Configuration of the transverse components E_r and H_φ of the electromagnetic field of the Sommerfeld wave. The third component E_z of this axially symmetrical E-mode is perpendicular to the plane of the figure (r, φ and z are cylindrical coordinates).

[1] W. Opechowski, Electromagnetic waves in waveguides, I and II, Philips tech. Rev. **10**, 13-25 and 46-54, 1948/49.

[2] The "skin depth" or "thickness of the equivalent conducting layer" is proportional to $1/\sqrt{f}$ (f = frequency). For copper the skin depth at f = 1000 Mc/s is 2×10^{-3} mm.

[3] A. M. Clogston, Reduction of skin effect losses by the use of laminated conductors, Bell Syst. tech. J. **30**, 491-529, 1951. E. F. Vaage, Transmission properties of laminated Clogston type conductors, Bell Syst. tech. J. **32**, 695-713, 1953.

[4] S. E. Miller and A. C. Beck, Low-loss waveguide transmission, Proc. IRE **41**, 348-358, 1953.

modes is extremely large and therefore they cannot be observed. Thus in practice the Sommerfeld wave is the only surface wave which can be excited on the single-wire transmission line.

Because of the cylindrical symmetry E_r, E_z and H_φ represent solutions of Bessel's differential equation. Without going into details of the theory, we shall briefly review a few essentials to give a better understanding of the following. The Bessel differential equation is linear and of second order, and thus has two independent linear solutions. The parameter p appearing in the differential equation and determining the order of the corresponding Bessel function can here for physical reasons be only zero or an integer. The two independent linear solutions are called Bessel functions of the first and the second kind, $\mathrm{J}_p(u)$ and $\mathrm{N}_p(u)$. ($\mathrm{N}_p(u)$ is also known as a Neumann function.) For real argument u these functions are rather similar to damped sine or cosine oscillations; for $u = 0$ the Bessel functions of the second kind become infinite.

In order to describe the electromagnetic field of the cylindrical transmission line, non-oscillatory functions are required which fall off sufficiently quickly in a certain distance from the wire. These requirements are satisfied by the Bessel functions of the third kind $\mathrm{H}_p^{(1)}(u)$ and $\mathrm{H}_p^{(2)}(u)$ ($\mathrm{H}_p(u)$ is also known as a Hankel function), which result from a linear combination of $\mathrm{J}_p(u)$ and $\mathrm{jN}_p(u)$. For imaginary argument u the asymptotic behaviour of these functions

$$\mathrm{H}_p^{(1,2)}(u) = \rightarrow \sqrt{\frac{2}{\pi u}}\, \mathrm{e}^{\pm \mathrm{j}\left(u - p\frac{\pi}{2} - \frac{\pi}{4}\right)} \quad . \; . \quad (1)$$

shows the resemblance to the decreasing exponential function.

A direct measure of the field concentration around a cylindrical surface-wave transmission line is that radius of the cylindrical cross-section within which a certain fraction of the total energy is transmitted. For the Sommerfeld line the power N_z transmitted across a cross-section of radius R is:

$$N_z(R) = \pi \,\mathrm{Re}\left(\int_a^R E_r H_\varphi{}^* r \,\mathrm{d}r\right), \quad . \; . \quad (2)$$

where a is the radius of the wire. The radial distribution of E_r and H_φ (and therefore of course also of the conjugate complex quantity $H_\varphi{}^*$) is described by the first Hankel function. In order to express the argument u by the radial coordinate r we put $u = hr$, the coefficient h resulting from the solution of the wave equation for the boundary-value problem in question. Because of the asymptotic behaviour of $\mathrm{H}_1(hr)$ it is clear that finite conductivity of the wire material is essential for the existence of the Sommerfeld wave (see eq. 1): this wave cannot exist for an ideal conductor, as when h is real the integral in eq. (2) does not converge for $R \rightarrow \infty$, which means that an infinitely large energy flow would be necessary for the propagation of the wave. For finite conductivity, on the other hand, h has an imaginary part which, for large values of r, causes an exponential decrease of the electromagnetic field, so that finite total energy is ensured.

Fig. 2 shows how, for a copper wire, field concentration and attenuation of the Sommerfeld wave depend on the frequency f and the radius a of the wire [5]. For a given frequency the losses decrease with increasing wire radius whereas the field extent increases. For a wavelength of 3 cm the attenuation α of a copper wire of 5 mm radius amounts to 12 dB/km, and the half-power radius $r_{3\mathrm{dB}}$ is 7 cm.

The radius $r_{10\mathrm{dB}}$ or $r_{20\mathrm{dB}}$ of the circular cross-section, i.e. the radius within which 90% or 99% of the power is transmitted, is of more practical interest. However the calculation of these radii for the range of a and f values in fig. 2 requires the tedious numerical evaluation of Hankel functions with complex argument, for which no appropriate tables are available. This difficulty can be avoided by introducing another parameter characterizing the field extension. Putting

$$h = h' + \mathrm{j}h'' = h' + \mathrm{j}\frac{1}{l}, \quad . \; . \; . \quad (3)$$

the real quantity l, which has the dimension of length, can serve as a measure of the field extent, For, since

$$\mathrm{e}^{\mathrm{j}hr} = \mathrm{e}^{\mathrm{j}h'r}\, \mathrm{e}^{-r/l}, \quad . \; . \; . \; . \; . \quad (4)$$

the field amplitude decreases to $1/\mathrm{e} = 0.37$ of its value

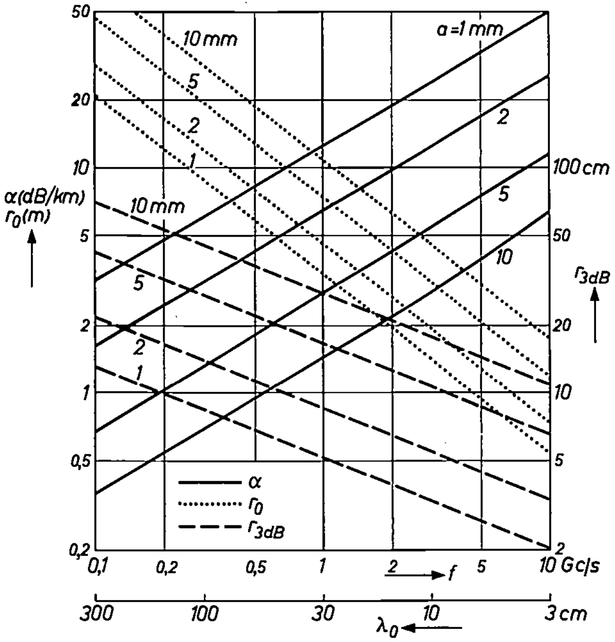


Fig. 2. Characteristic radius r_0, half-power radius r_{3dB} and attenuation α of the Sommerfeld wave for copper wires of various radii a (1 mm, 2 mm, 5 mm and 10 mm) as a function of the frequency f.

if radial distance increases by l, as long as r is such that eq. (1) is valid. Over the range of values of wire radius and frequency used in fig. 2, $|ha| < 0.1$, and as calculation shows, more than 90% of the energy is transmitted across the circular cross-section of radius $r_0 = l$. r_0 is called the "characteristic radius". The comparison of the curves in fig. 2 shows that for the chosen wire radii the characteristic radius, which gives a more realistic indication of the field concentration, is 20 to 150 times larger than the half-power radius $r_{3\text{dB}}$.

The characteristic radius has significance only if it is greater than the wire radius a. If one is interested in the properties of the Sommerfeld line for the millimetre waveband [6] and values of a greater than the wavelength are permitted, the range of values of ha cannot be restricted, as now we may have $h''a = a/l > 1$. Here one can introduce another distance characterizing the field extent, this distance being measured not from the axis but from the surface of the wire. The calculation shows that at least 87% of the energy is transmitted across the circular cross-section of radius $a + l = a + x_0$. The distance x_0 measured from the surface of the line is called the "characteristic distance". As long as $a \ll r_0$, x_0 is nearly equal to r_0. In the limiting case $a/\lambda \rightarrow \infty$, i.e. for the plane surface-wave transmission line, x_0 is that distance from the surface of the transmission line for which the field strength has fallen to 1/e of the value at the surface [7]. The curves of characteristic distance x_0 and attenuation α as functions of the frequency f and the wire radius a are, for large values of a/x_0, completely analogous to the curves of fig. 2 for small values of a/x_0. Again, for a given frequency, the losses can be reduced by using larger wire radii. This, however, increases the extent of the electromagnetic field. If a maximum attenuation of 3.5 dB/km is allowed, the Sommerfeld line is not very practical for use at wavelengths below 5 cm because the wire diameter becomes too large (see *fig. 3*). The upper frequency limit is determined by the field extent. At a wavelength of 15 cm the characteristic distance has reached a value of 5 m.

This type of transmission line does not seem particularly attractive as a stronger field concentration around the wire can only be achieved by reducing the conductivity. This is an unsatisfactory solution as the ohmic losses then increase. The basic difficulty with the Sommerfeld line is due to the fact that a finite conductivity is necessary for the existence of the surface wave. For conventional transmission lines, on the contrary, ohmic losses in the conductor are completely undesirable and superfluous. Attempts have therefore been made to modify the single-wire line in such a way that the conductivity is no longer a determining factor for the existence of surface waves.

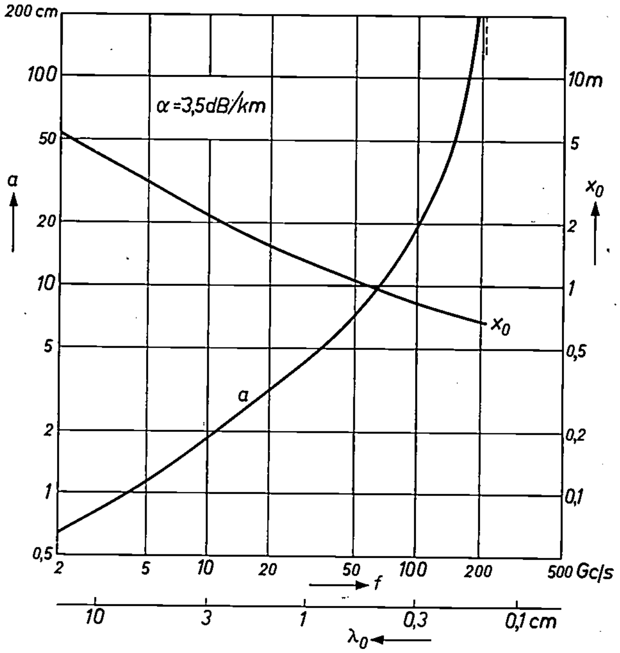


Fig. 3. Radius a of a Sommerfeld line, of copper, with attenuation of 3.5 dB/km and characteristic distance x_0, as a function of frequency f. For $a \rightarrow \infty$, x_0 approaches the value for the *plane* surface-wave transmission line with an attenuation of 3.5 dB/km. The corresponding frequency f is 202.2 Gc/s.

The dielectric-coated wire (Harms-Goubau line)

The radial distribution and the asymptotic behaviour of the field of the single-wire line (eq. (1) with $u = hr$) indicate how an exponential decrease of the field can be achieved for large values of r : h must have an imaginary component. If the propagation in the positive z-direction is described in the usual way by $\exp(\mathrm{j}\omega t - \gamma z)$ (γ = propagation constant), it follows from the wave equation that

$$h = \sqrt{k^2 + \gamma^2}, \quad . \quad . \quad . \quad . \quad . \quad (5)$$

where $k = \omega/c = 2\pi/\lambda_0$, $\omega = 2\pi f$ is the angular frequency of the signal, c is the velocity of light in free space, and λ_0 is the free-space wavelength. For the Sommerfeld line, because of the attenuation α, the propagation constant γ is complex: $\gamma = \alpha + \mathrm{j}\beta$. It would, however, be more desirable if γ were imaginary ($\alpha = 0$), so that

$$-\mathrm{j}\gamma = \beta = \frac{\omega}{v} = \frac{2\pi}{\lambda}, \quad . \quad . \quad . \quad . \quad (6)$$

and if at the same time $\beta^2 > k^2$, i.e. the phase velocity

[5] G. Goubau, Surface waves and their application to transmission lines, J. appl. Phys. **21**, 1119-1128, 1950.
O. Zinke, Kabel- und Funkweg im Mikrowellenbereich, Nachrichtentechn. Z. **10**, 425-430, 1957.
H. Kaden, Fortschritte in der Theorie der Drahtwellen, Archiv elektr. Übertr. **5**, 399-414, 1951.

[6] H. Severin, Sommerfeld- und Harms-Goubau-Wellenleiter im Bereich der Zentimeter- und Millimeterwellen, Archiv elektr. Übertr. **14**, 155-162, 1960.

[7] H. Kaden, Dielektrische und metallische Wellenleiter, Archiv elektr. Übertr. 6, 319-332, 1952.

$v < c$ (λ = wavelength on the line, $\lambda < \lambda_0$). h would then become imaginary and a surface wave can exist without the attenuation being a necessary condition.

One possibility of reducing the phase velocity and therefore the field extent of the wave is given by coating the wire with a thin dielectric layer. The theory of such a transmission line was developed by Harms in 1907 following Sommerfeld's paper. The possibility of an application in the microwave range was first discussed in 1950, by Goubau [5], who showed in particular how surface waves can be excited with good efficiency on single wires.

The mathematical treatment of this problem shows that the axially symmetrical E-mode is again the only mode with low attenuation as long as the greater part of the electro-magnetic field extends into the surrounding space. The thin dielectric coating brings about the desired concentration of the field around the transmission line by means of the retardation of the wave front in the dielectric, even if the conductivity of the wire is infinite. For the Harms-Goubau line ohmic losses are undesirable and not necessary for the existence of the surface wave .This physical fact permits an essential simplification of the calculation, that could not be made with the Sommerfeld line: the field extent can be given for the idealized assumption that there are no losses. The attenuation of the Harms-Goubau wave is then calculated separately, assuming as is usual for low-loss lines that the field distribution is to a first-order approximation the same as in the loss-free case.

For the problem thus idealized the field extent of the Harms-Goubau wave can be given as a function of the line data and frequency. Since Hankel functions with imaginary argument are tabulated, the radius of the cross-section transmitting a stated percentage of the energy may be calculated. Exactly as with the Sommerfeld line, for large $|h|a'$ (a' = radius of the Harms-Goubau line) at least 87% of the energy, and for small $|h|a'$ about 95% of the energy is transmitted within a distance x_0 from the surface of the line as fig. 4 shows. Even with the above simplifying assumptions it is again not possible to express $|h| = 1/x_0$ explicitly as a function of the transmission line data. A transcendental equation is obtained, and graphical evaluation shows that the field concentration round the line increases as the wire radius decreases, and increases as the dielectric constant of the coating, its thickness, and the frequency increase. *Figs. 5* and *6* show for example that at 10 Gc/s (λ = 3 cm) with a wire radius a of 1 mm, the field extent can be reduced by a factor of about 10 by applying a 0.05 mm dielectric coating, but that the attenuation due to ohmic losses is increased only by a factor of 1.6. The additional dielectric losses are only a small fraction of the ohmic losses (*fig. 7*). In practice by suitably choosing the transmission line data, any desired field concentration can be obtained. However, at the same time the attenuation increases so that a compromise between field extension and attenuation always has to be made. The result is that the Harms-Goubau line can be applied for metre and decimetre waves down to a smallest wavelength of about 8 cm. For longer wavelengths it is also possible to use a magnetic coating instead of a dielectric one (Kaden [5]).

In a first application the Harms-Goubau line has been used for feeding transmitting aerials [8].

Metallic conductors with periodic structure

For the existence of a surface wave, longitudinal field components are necessary. In the Sommerfeld line they are provided for by the finite conductivity of the wire; in the Harms-Goubau-type line by the retardation of the wave front in the dielectric layer.

Another way of obtaining longitudinal field components is to use a helical wire. Such helical lines have already become familiar through their application in travelling wave tubes. In this application its function is to reduce the phase velocity of the wave so that interaction can take place between the axial component of the electromagnetic field and the electron beam. For this application helices of small pitch (helix angle 5° — 8°) are used. For transmission lines however large pitch (helix angle 70° — 80°) should be chosen to keep ohmic losses low.

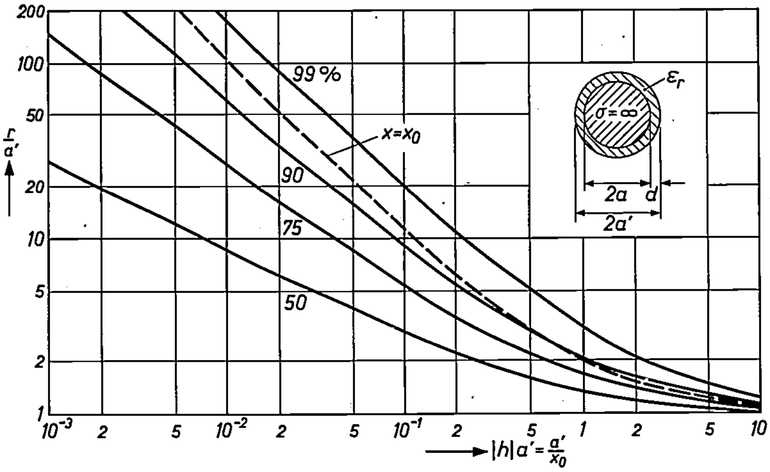


Fig. 4. Radii of the cross-sections across which 50%, 75%, 90% and 99% of the energy of the Harms-Goubau wave are transmitted, as a function of the frequency and the line data.

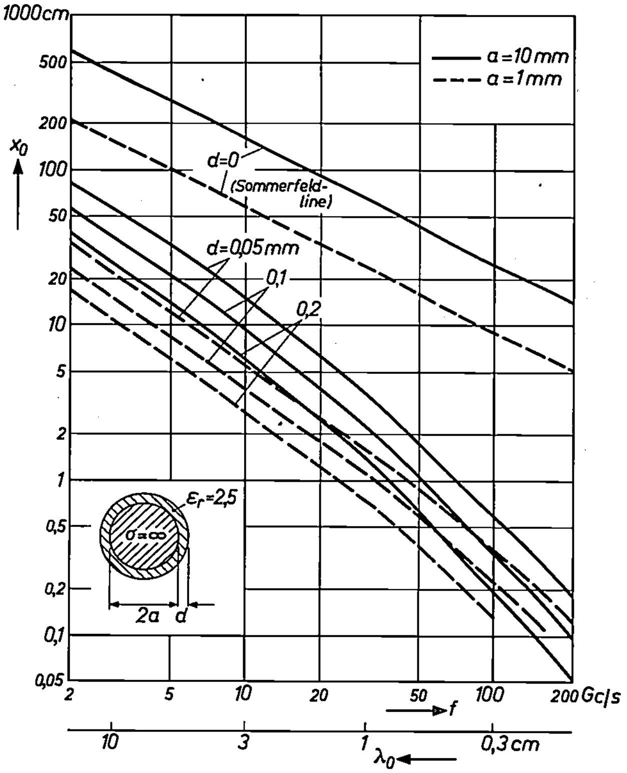


Fig. 5. Characteristic distances x_0 of the Harms-Goubau and the Sommerfeld lines as a function of the frequency and the line data.

Pitch D, helix angle ψ and radius ϱ of the helix are connected by the relation:

$$\cot\psi = \frac{2\pi\varrho}{D}. \quad \ldots\ldots\ldots \quad (7)$$

In a first approximation [9] useful in many cases, the helix is replaced in the calculation by an extremely thin cylindrical tube whose wall has very high conductivity in the direction determined by the pitch angle ψ and is non-conducting in the perpendicular direction ($\psi + \pi/2$). The resulting electromagnetic field has all six components and can be interpreted as superposition of an H- and an E-mode with equal propagation constants. As with the Sommerfeld and the Harms-Goubau lines, the relationship between the propagation constant and the radius, the pitch angle of the helix, and the frequency is by no means simple. For a large pitch, which is of interest here, the helical line has the same transmission properties as the Harms-Goubau line if between the data for the two lines the relation

$$\left(1 - \frac{1}{\varepsilon}\right)\frac{d}{a} = \frac{1}{2}\cot^2\psi. \quad \ldots \quad (8)$$

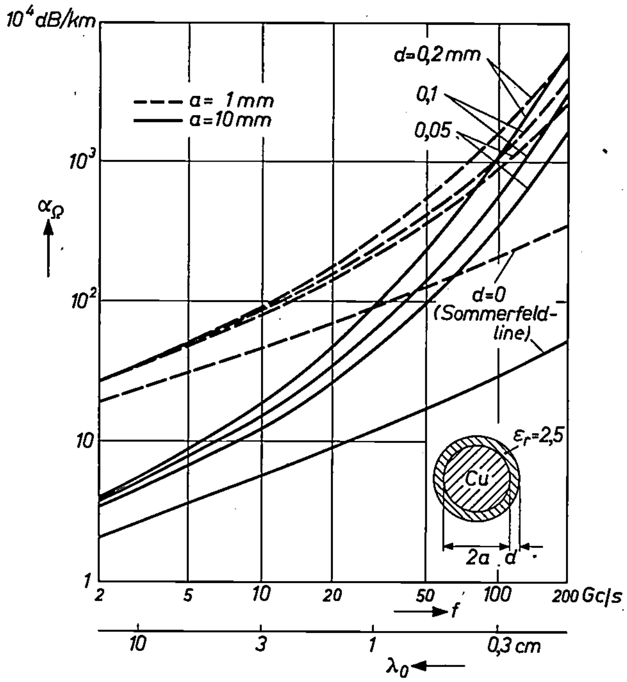


Fig. 6. Attenuation α_Ω due to ohmic losses for the Harms-Goubau and the Sommerfeld lines as a function of the frequency and the line data.

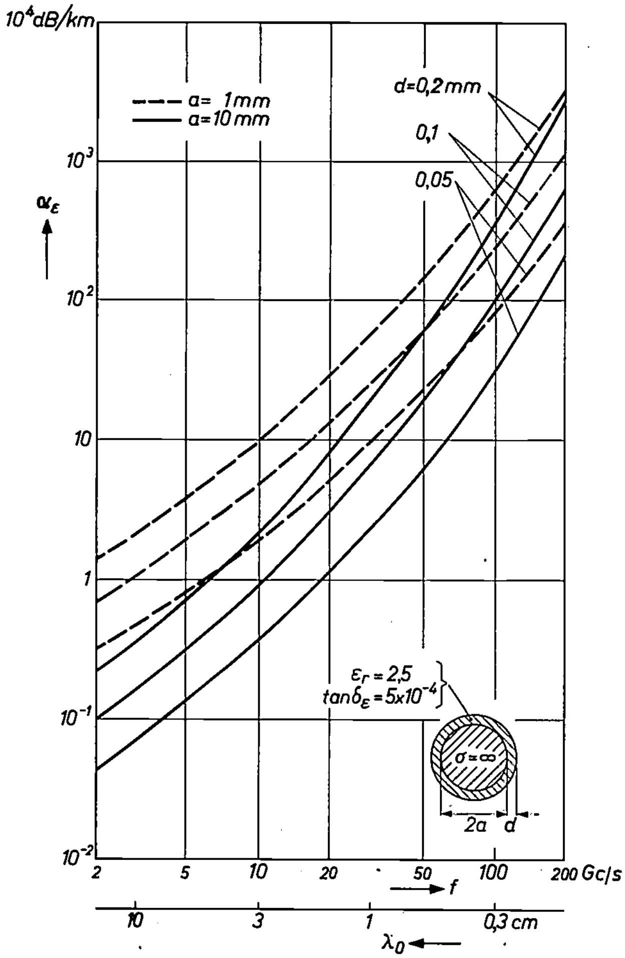


Fig. 7. Attenuation α_ε due to dielectric losses for the Harms-Goubau line as a function of the frequency and the line data.

[8] F. R. Huber, Speisung von Sendeantennen mit Hilfe von Goubau-Leitungen, Nachrichtentechn. Fachber. 23, 114-125, 1961.

[9] H. Kaden, Eine allgemeine Theorie des Wendelleiters, Archiv elektr. Übertr. 5, 534-538, 1951.

holds. This means that the curves of figs. 4, 5 and 6 evaluated for the Harms-Goubau line apply also to the helical line.

Longitudinal field components and a reduction of phase velocity can also be obtained by a suitable "roughening" of the surface of the perfectly conducting wire. Goubau indicates in his paper [5] that instead of a dielectric coating, any periodic microstructure in the longitudinal direction, such as a screw thread, will also give rise to a greater field concentration. On deepening the grooves of the screw thread the structure eventually becomes that of a disc-loaded line. The periodic structure of the surface which to some extent may be interpreted as a kind of artificial dielectric shows, however, an essential difference when compared with the plain Harms-Goubau line: besides the dominant mode, the axially symmetrical E-mode already described, the periodic structure always has higher-order modes as well, the so-called "space harmonics" of the dominant mode [10]. Their relative amplitudes depend on depth, width and distance of the grooves. In order to keep the amplitudes of the space harmonics within the theoretical limits, a high degree of uniformity of the periodic structure is necessary. Moreover, surface-wave lines with periodic structure exhibit band-pass filter characteristics, i.e. they do not transmit over the whole frequency range but only in certain frequency-bands.

The dielectric line

In connection with Sommerfeld's paper dealing with waves on a metal wire Hondros and Debye, in 1910, investigated the propagation of electiromagnetic waves along circular cylinders of homogeneous dielectric. The experimental proof of the existence of such waves was given by Zahn in 1915. These basic studies, as well as more recent work [11], have shown that for the dielectric line as in the hollow metallic waveguide an infinite number of modes is possible.

All the modes, with the exception of the dominant mode, have a cut-off frequency, below which they cannot exist. Apart from the dominant mode, wave propagation is possible only if the diameter of the dielectric line is approximately equal to or larger than half the wavelength in the dielectric. The situation in which either the electric or the magnetic field has no components other than transverse ones only arises for waves of axially symmetric field. For such waves the field configurations over the cross-section will be similar to those of the corresponding modes in the metallic circular waveguide. For the H_{0n}-modes the electric field lines are concentric circles; however, the surface of the dielectric cylinder does not have to be a nodal surface of the electric field. For the E_{0n}-modes the transverse component of the electric field is radial.

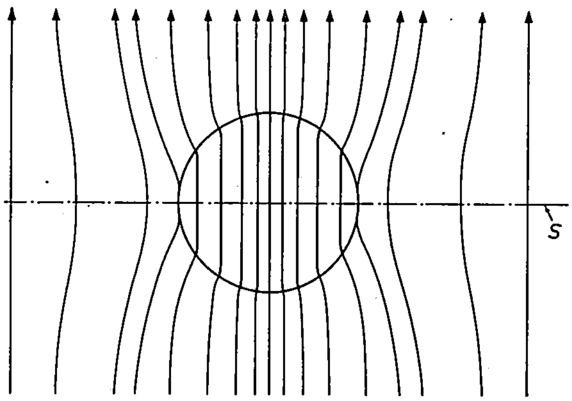


Fig. 8. Transverse component of the electric field E for the HE_{11}-mode (dominant mode) on a dielectric line.

For the non-axially symmetric waves the boundary conditions can no longer be satisfied by an H- or an E-mode alone. Modes with the first index $m \neq 1$, are, depending on their axial field components, called HE_{mn} modes if the field configuration is similar to that of an H-mode, and EH_{mn}-modes if it resembles an E-mode.

The HE_{11}-mode is the only mode with no cut-off frequency. This is the dominant mode of the dielectric line. *Fig. 8* shows the configuration of the transverse component of the electric field. Due to the fact that there

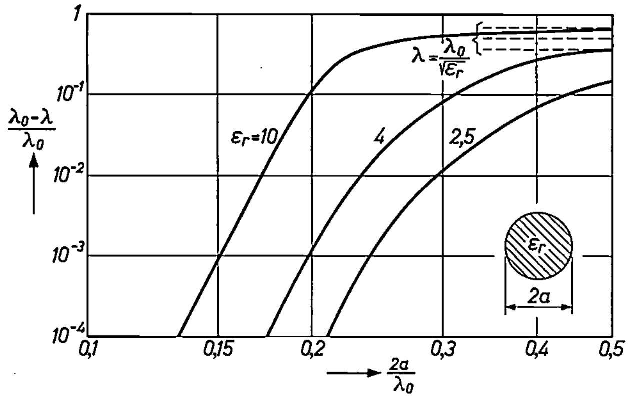


Fig. 9. Deviation of the wavelength λ for the HE_{11}-mode on a dielectric line from the wavelength in vacuo λ_0, given as a function of the ratio of the line diameter $2a$ to the wavelength in vacuo λ_0.

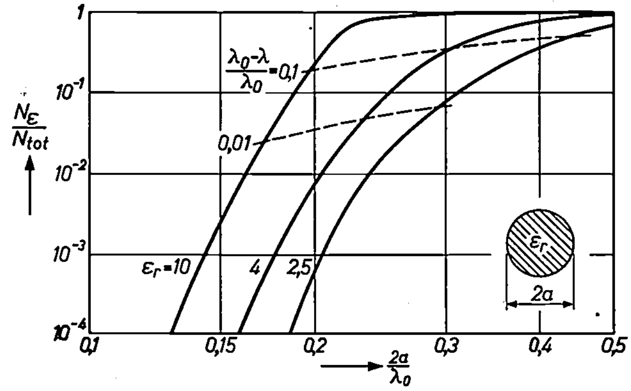


Fig. 10. Ratio of the energy N_ε transmitted in the dielectric to the total energy $N_{\rm tot}$ of the HE_{11}-mode on a dielectric line as a function of the ratio of the line diameter $2a$ to the wavelength in vacuo λ_0.

is no cut-off frequency, the line can be dimensioned in such a way that only a small percentage of energy is transmitted in the dielectric. This means that the attenuation can be kept very low, but the radial field extent is again large. An appreciable reduction of phase velocity and a corresponding concentration of the field is obtained only above a certain ratio of the diameter of the dielectric "string" to wavelength (*figs. 9* and *10*). For high values of the dielectric constant ε_r there is only a narrow band of frequencies in which the velocity remains substantially the same as that of light in vacuo. Only then is the field moderately concentrated. Therefore, the field of the HE_{11}-mode is either almost completely inside or almost completely outside the dielectric string. For transmission line application the strong frequency dependence of phase velocity, field extent and attenuation is undesirable. A small value of the dielectric constant should therefore be chosen in order to remain within the range of low dispersion. For ε_r values of 2.5 and 10, line wavelength and wavelength in vacuo differ from each other by 1% when $2a/\lambda_0 = 0.3$ and 0.17 respectively. When the deviation of the phase velocity from that of the velocity of light in vacuo is so small, only a small percentage of the energy is transmitted in the dielectric. According to fig. 10 the percentages are 7% and 2.5% for $\varepsilon_r = 2.5$ and 10 respectively.

The weaker frequency dependence of field extent and attenuation for small values of ε_r is also confirmed by *fig. 11*, which shows these quantities as a function of the ratio of dielectric string diameter to wavelength in vacuo for several values of ε_r. For each dielectric constant there is a minimum field extent, which cannot be reduced by further increase of the string diameter. For low-loss transmission however, it is the range of small $2a/\lambda_0$ which is of interest, for which the greater part of the field lies outside the string. In this case, contrary to that with the Sommerfeld and the Harms-Goubau lines, the attenuation decreases as the ratio of the string diameter to the wavelength in vacuo is decreased. The fact that the transmission depends so strongly on diameter and frequency necessitates close tolerances for

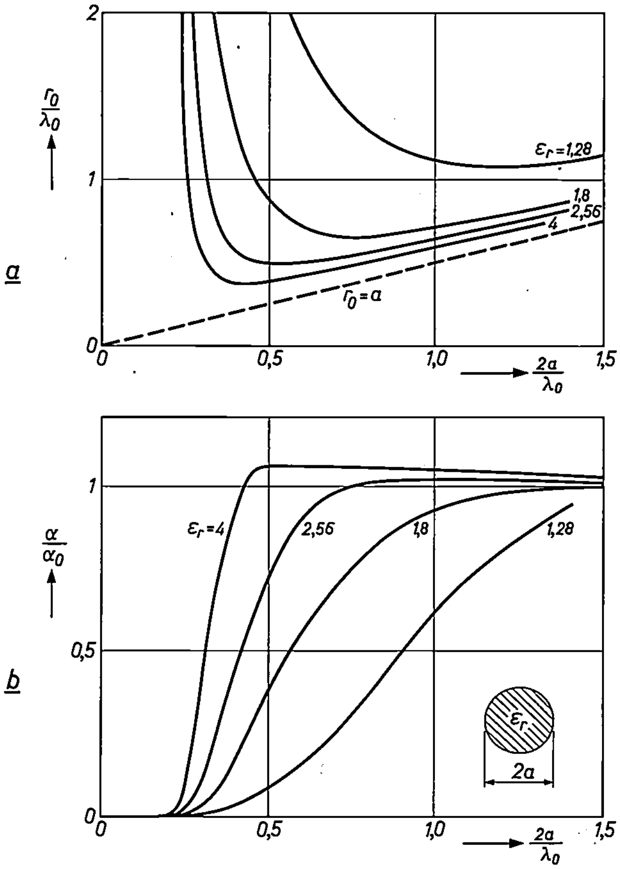


Fig. 11. *a*) Characteristic radius r_0 and *b*) attenuation α of the HE_{11}-mode of a dielectric line for various values of the dielectric constant ε_r, as a function of the ratio of the line diameter $2a$ to the wavelength in vacuo. $\alpha_0 = (\pi/\lambda_0)\sqrt{\varepsilon_r}\tan\delta_\varepsilon$ is the attenuation of a plane wave in the infinite dielectric.

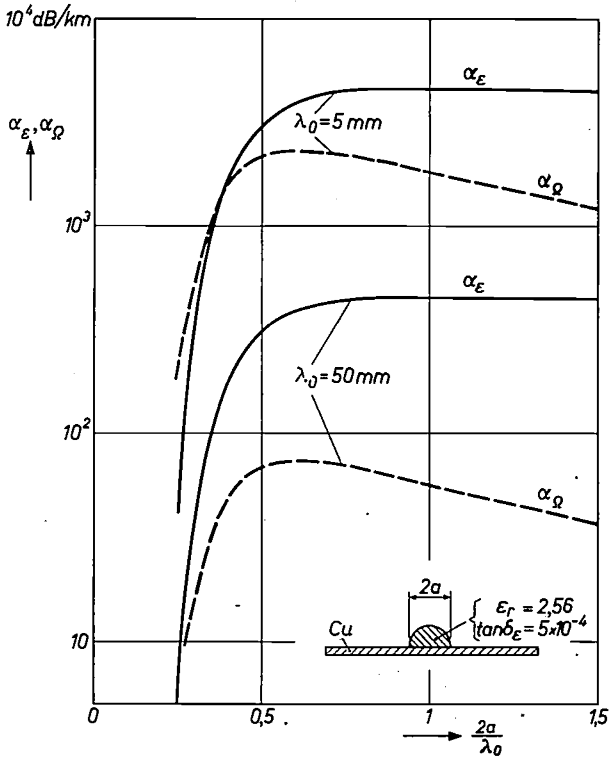


Fig. 12. Attenuation α_ε due to dielectric losses and attenuation α_Ω due to ohmic losses for the dielectric image line, for two wavelengths, as a function of the diameter.

[10] L. Brillouin, Wave guide for slow waves, J. appl. Phys. **19**, 1023-1041, 1948.
L. M. Field, Some slow-wave structures for travelling wave tubes, Proc. IRE **37**, 34-40, 1949.
W. Rotman, A study of single-surface corrugated guides, Proc. IRE **39**, 952-959, 1951.

[11] C. H. Chandler, An investigation of dielectric rod as waveguide, J. appl. Phys. **20**, 1188-1192, 1949.
W. M. Elsasser, Attenuation in a dielectric circular rod, J. appl. Phys. **20**, 1193-1196, 1949.
P. Mallach, Untersuchungen an dielektrischen Wellenleitern in Stab- und Rohrform, Fernmeldetechn. Z. **8**, 8-13, 1955.

[12] D. D. King, Properties of dielectric image lines, IRE Trans. MTT-3, 75-81, 1955.
S. P. Schlesinger and D. D. King, Dielectric image lines, IRE Trans. MTT-6, 291-299, 1958.

the dimensions and homogeneity of the dielectric line, and permits the transmission of only narrow frequency bands.

The problem of supporting surface-wave guides leads to another type of dielectric line, the so-called dielectric image line [12]. Since the electric field of the HE_{11}-mode at a plane of symmetry S is everywhere perpendicular to this plane (fig. 8), a perfectly conducting sheet in this plane would not disturb the field distribution. In practicy of course the finite conductivity of the reflector will add ohmic losses to the dielectric losses. *Fig. 12* shows that the attenuation due to the metal sheet is smaller than that due to the dielectric down to wavelengths of a few millimetres. However, this no longer holds if for small values of $2a/\lambda_0$ only a small percentage of the energy is transmitted inside the dielectric so that the dielectric losses are very small.

Experience has shown that the dielectric image line is applicable down to a wavelength of about 3 mm. For even shorter wavelengths the dielectric line is to be preferred.

Summary. Surface-wave transmission lines may be advantageous for the transmission of high-frequency signals. The most important types of line are the single metal wire (the Sommerfeld line), the metal wire with dielectric coating (the Harms-Goubau line) and the dielectric line. Their applicability depends on the attenuation and the radial extent of the electromagnetic field. The properties of the various types of lines treated in this paper lead to the conclusion that for wavelengths down to about 8 cm the Harms-Goubau line is the most suitable. For longer wavelengths a magnetic instead of dielectric coating can also be used. Between 15 and 5 cm wavelength the Sommerfeld line, whose attenuation decreases with increasing field extent is also applicable. For still shorter wavelengths the dielectric line and the dielectric image line are more suitable, the latter down to about 3 mm wavelength. Below this wavelength the dielectric line is up to now the only suitable surface-wave transmission line.

II. Applications of the dielectric line

G. Schulten

Further considerations for the dielectric line

Part I of this article explained the circumstances in which electromagnetic surface waves of cylindrical form can exist. It was shown that a cylindrical filament is always necessary for this purpose, but that it need not be of metal. The longitudinal component E_z of the electric field, without which a cylindrical surface wave is not possible, can be obtained just as well with a filament or "string" made from a dielectric material. It was also shown in part I that in this case it is possible to have a mode of the electromagnetic field which has no cut-off frequency, and which can therefore appear at arbitrarily low frequencies. The possible applications of the dielectric line are to be found, however, where it is difficult to employ coaxial conductors and waveguides, i.e. in the millimetre-wave range. The reason for this is that, on a line propagating surface waves, low attenuation is only possible if the field extent in the radial direction is fairly large — a distance of at least several wavelengths. Dielectric lines are therefore ruled out for lower frequency applications.

The thickness of the dielectric "string" can be suitably chosen so that only the dominant mode, known as the HE_{11} mode, may be set up, and not the other modes, i.e. the modes with cut-off frequencies. Let d be the diameter of the dielectric string, λ_0 the free space wavelength and ε_r the relative dielectric constant of the material, then the HE_{11} mode alone is possible if:

$$\frac{d}{\lambda_0}\sqrt{\varepsilon_r - 1} \leqq 0.7655 . \quad . \quad . \quad . \quad (1)$$

At a wavelength of 4 mm the diameter of the polyethylene dielectric ($\varepsilon_r = 2.4$) must be smaller than 2.59 mm in order to satisfy this condition. In general, however, the string will have to be very much thinner than this, say about 1 mm, to achieve lower attenuation. The best dielectric materials at present available (polyethylene, polytetrafluorethylene, polystyrene) have a loss angle of the order of tan $\delta = 10^{-4}$. This means that a wave propagating completely inside the material would, at a wavelength of 4 mm, be attenuated at the rate of about 1 dB per metre. This still fairly considerable attenuation can be appreciable diminished if the electromagnetic energy is transmitted largely outside the material. This is accomplished by using a small diameter for the string.

If the power transmitted inside the dielectric is N_ε and the total power is N, the attenuation is given by:

$$a = \frac{\pi}{\lambda_0}\sqrt{\varepsilon_r}\tan\delta\,\frac{N_\varepsilon}{N}\ \text{(nepers/metre)}, \quad . \quad (2)$$

where λ_0 is again the free space wavelength.

Fig. 1 shows in graphical form the dependence of N_ε/N on the ratio of the string diameter to the wavelength in air (d/λ_0), assuming $\varepsilon_r = 2.5$. It is clearly

Dipl.-Phys. G. Schulten is a research worker at the Hamburg Laboratory of Philips Zentrallaboratorium GmbH.

the dimensions and homogeneity of the dielectric line, and permits the transmission of only narrow frequency bands.

The problem of supporting surface-wave guides leads to another type of dielectric line, the so-called dielectric image line [12]. Since the electric field of the HE_{11}-mode at a plane of symmetry S is everywhere perpendicular to this plane (fig. 8), a perfectly conducting sheet in this plane would not disturb the field distribution. In practicy of course the finite conductivity of the reflector will add ohmic losses to the dielectric losses. *Fig. 12* shows that the attenuation due to the metal sheet is smaller than that due to the dielectric down to wavelengths of a few millimetres. However, this no longer holds if for small values of $2a/\lambda_0$ only a small percentage of the energy is transmitted inside the dielectric so that the dielectric losses are very small.

Experience has shown that the dielectric image line is applicable down to a wavelength of about 3 mm. For even shorter wavelengths the dielectric line is to be preferred.

Summary. Surface-wave transmission lines may be advantageous for the transmission of high-frequency signals. The most important types of line are the single metal wire (the Sommerfeld line), the metal wire with dielectric coating (the Harms-Goubau line) and the dielectric line. Their applicability depends on the attenuation and the radial extent of the electromagnetic field. The properties of the various types of lines treated in this paper lead to the conclusion that for wavelengths down to about 8 cm the Harms-Goubau line is the most suitable. For longer wavelengths a magnetic instead of dielectric coating can also be used. Between 15 and 5 cm wavelength the Sommerfeld line, whose attenuation decreases with increasing field extent is also applicable. For still shorter wavelengths the dielectric line and the dielectric image line are more suitable, the latter down to about 3 mm wavelength. Below this wavelength the dielectric line is up to now the only suitable surface-wave transmission line.

II. Applications of the dielectric line

G. Schulten

Further considerations for the dielectric line

Part I of this article explained the circumstances in which electromagnetic surface waves of cylindrical form can exist. It was shown that a cylindrical filament is always necessary for this purpose, but that it need not be of metal. The longitudinal component E_z of the electric field, without which a cylindrical surface wave is not possible, can be obtained just as well with a filament or "string" made from a dielectric material. It was also shown in part I that in this case it is possible to have a mode of the electromagnetic field which has no cut-off frequency, and which can therefore appear at arbitrarily low frequencies. The possible applications of the dielectric line are to be found, however, where it is difficult to employ coaxial conductors and waveguides, i.e. in the millimetre-wave range. The reason for this is that, on a line propagating surface waves, low attenuation is only possible if the field extent in the radial direction is fairly large — a distance of at least several wavelengths. Dielectric lines are therefore ruled out for lower frequency applications.

The thickness of the dielectric "string" can be suitably chosen so that only the dominant mode, known as the HE_{11} mode, may be set up, and not the other modes, i.e. the modes with cut-off frequencies. Let d be the diameter of the dielectric string, λ_0 the free space wavelength and ε_r the relative dielectric constant of the material, then the HE_{11} mode alone is possible if:

$$\frac{d}{\lambda_0}\sqrt{\varepsilon_r - 1} \leqq 0.7655 . \quad . \quad . \quad (1)$$

At a wavelength of 4 mm the diameter of the polyethylene dielectric ($\varepsilon_r = 2.4$) must be smaller than 2.59 mm in order to satisfy this condition. In general, however, the string will have to be very much thinner than this, say about 1 mm, to achieve lower attenuation. The best dielectric materials at present available (polyethylene, polytetrafluorethylene, polystyrene) have a loss angle of the order of tan $\delta = 10^{-4}$. This means that a wave propagating completely inside the material would, at a wavelength of 4 mm, be attenuated at the rate of about 1 dB per metre. This still fairly considerable attenuation can be appreciable diminished if the electromagnetic energy is transmitted largely outside the material. This is accomplished by using a small diameter for the string.

If the power transmitted inside the dielectric is N_ε and the total power is N, the attenuation is given by:

$$a = \frac{\pi}{\lambda_0}\sqrt{\varepsilon_r}\tan\delta\,\frac{N_\varepsilon}{N}\ \text{(nepers/metre)}, \quad . \quad (2)$$

where λ_0 is again the free space wavelength.

Fig. 1 shows in graphical form the dependence of N_ε/N on the ratio of the string diameter to the wavelength in air (d/λ_0), assuming $\varepsilon_r = 2.5$. It is clearly

Dipl.-Phys. G. Schulten is a research worker at the Hamburg Laboratory of Philips Zentrallaboratorium GmbH.

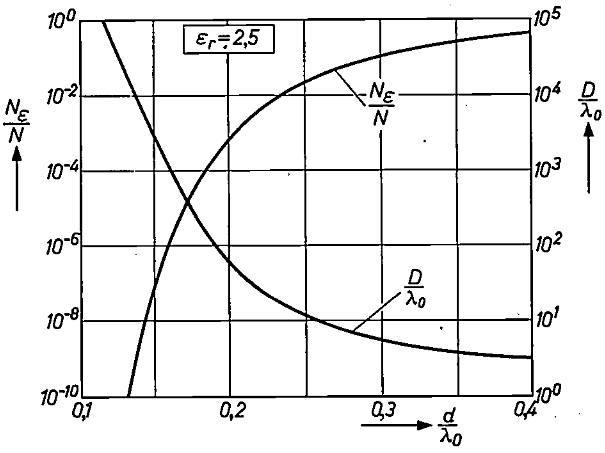


Fig. 1. Field energy distribution for the HE_{11} mode on a dielectric filament or "string" (circular cross-section); dielectric constant of the material $\varepsilon_r = 2.5$. The graph is a plot of the diameter D of the cylindrical volume within which 99% of the power is transmitted, and the fraction N_ε/N of the total power transmitted inside the string, both as a function of the ratio of the string diameter d and the wavelength λ. At small string diameters only a very small fraction of the power is transmitted inside the string. The attenuation will then be low, but the radial extent of the field is quite large. At large string diameters the field concentration is favourable but the attenuation is high.

seen that the energy is transmitted almost entirely outside the string if the diameter of the string is very small. This has the disadvantage, however, that the electromagnetic field then extends much farther into free space. To demonstrate this, fig. 1 gives as well the curve of the effective field diameter D, defined as the diameter of the cylinder within which 99% of the power is transmitted. This diameter D is seen to increase rapidly as the string diameter d decreases. For a wavelength of 4 mm and polyethylene string 1 mm thick the result is rather more clearly illustrated in *fig. 2*. This figure also shows the diameter of the cylinder within which 99.9% of the energy is transmitted.

Fig. 3 gives a perspective drawing of the field configuration of the HE_{11} mode. The full lines represent the electric field E, and two magnetic field lines H are represented by dashed lines. There are two planes of symmetry, which pass through the axis of the line. It is particularly important to note that the electric field, which always crosses the horizontal plane of symmetry at right angles lies almost entirely in vertical planes. In much the same way as for a waveguide, one can speak here of a direction of polarization. The energy distribution however, has almost complete circular symmetry. These properties are of importance in connection with the dielectric image line, to be discussed shortly, and with the excitation of the wave.

The electromagnetic field can be calculated accurately only for a string of circular cross-section. Some of the principles underlying the calculation were given in part I of this article. Experience has shown, however,

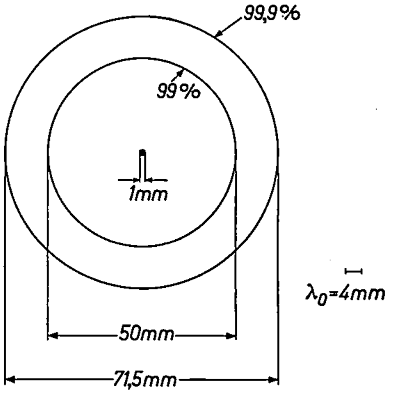
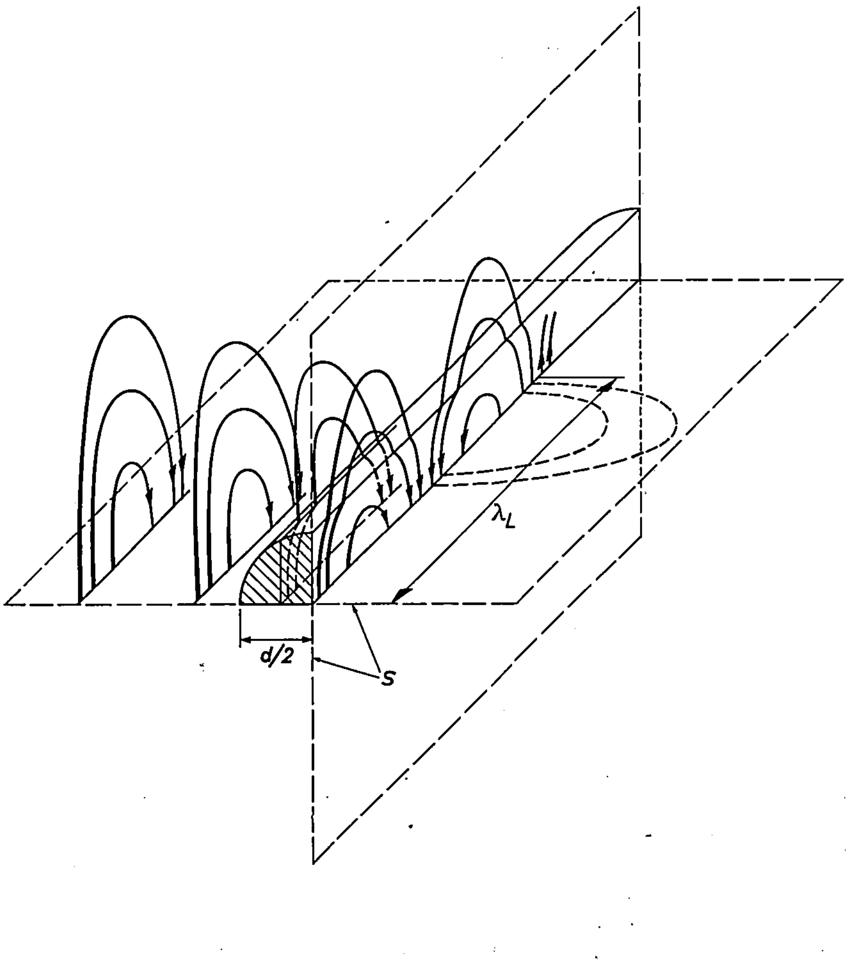


Fig. 2. Energy distribution around a dielectric string (polyethylene) of 1 mm diameter at a wavelength of 4 mm. The circles give the boundaries for 99% and 99.9% of the power transmission.

→

Fig. 3. Field lines for the HE_{11} mode. A quarter of the cross-section of the dielectric string is shown. The full lines denote the electric field, and two magnetic lines of force are indicated on the right by dashed lines. There are two planes of symmetry s, one of which (the horizontal one) is cut vertically by the electric lines of force. The wavelength on the line is denoted by λ_L.



that good use can also be made of dielectric string with other cross-sections. Here, too, the dominant mode without a cut-off frequency can occur. As long as the linear dimensions of the cross-section do not exceed the wavelength, the properties of the wave are mainly governed by the area of the cross-section.

At a wavelength of 4 mm good results have been obtained with a flat bunch of dielectric threads, as illustrated in *fig. 4*. The field here extends just as far as with the string of circular section, also shown in the figure, but the attenuation was found to be only half as great, i.e. 0.05 dB per metre. For this arrangement the space around the threads had to be kept free up to a distance of 10 cm.

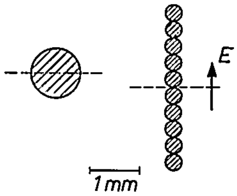


Fig. 4. Cross-section of two dielectric lines of equal field extent but different attenuation. The bunch of thin dielectric threads has roughly half the attenuation of the single round dielectric string. As in fig. 3, the electric field E is perpendicular to a horizontal plane of symmetry.

The dielectric image line

The dielectric image line makes use of the symmetry of the field, with respect to a horizontal plane through the axis of the dielectric filament as seen in fig. 3. Due to the symmetry, the lower half of the field can be replaced by the mirror image of the upper half, obtained from a metal surface coincident with this plane of symmetry. This is possible because the electric lines of force are perpendicular to this plane, so that the boundary conditions are fulfilled [1].

Fig. 5 shows four practical forms of the dielectric image line. The first case (*a*) corresponds to the cylindrical configuration, which can be calculated. The next two forms (*b*) and (*c*) have the advantage that the dielectric string does not have to be held to the metal surface with adhesive but can be clamped to it. The fourth case (*d*) has the virtue of particularly low attenuation [2].

The finite conductivity of the metal plate, as may be expected gives rise to extra attenuation. At millimetre waves these additional losses are relatively small, but they increase as the square of the frequency. Whereas with the free cylindrical dielectric string the attenuation can in principle be reduced to an arbitrarily low value, with the image line the attenuation can never be lower than that for the surface wave on a flat metal plate [3]. At a wavelength of 1 mm the attenuation is roughly 0.01 dB/m.

A practical advantage of the image line is the ease with which the metal plate can be mounted. Since, however, dielectric lines are usually used for applications where lengths of only a few metres are required, the freely suspended string seldom leads to practical difficulties. In the applications to be discussed we shall therefore confine ourselves to this arrangement.

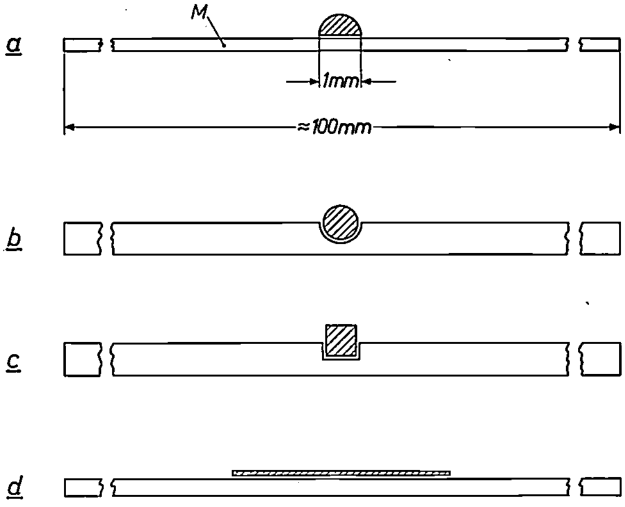


Fig. 5. Various forms of the dielectric image line. The upper face of the metal plate M always coincides with the horizontal plane of symmetry in fig. 3.

Excitation of the surface wave

Fig. 6 illustrates two methods of exciting a surface wave on a dielectric line. Fig. 6*a* shows how a rectangular waveguide (*1*) is connected by means of a transition (*2*) to a horn (*3*) of circular cross-section, the axis of which coincides with that of the dielectric string. The

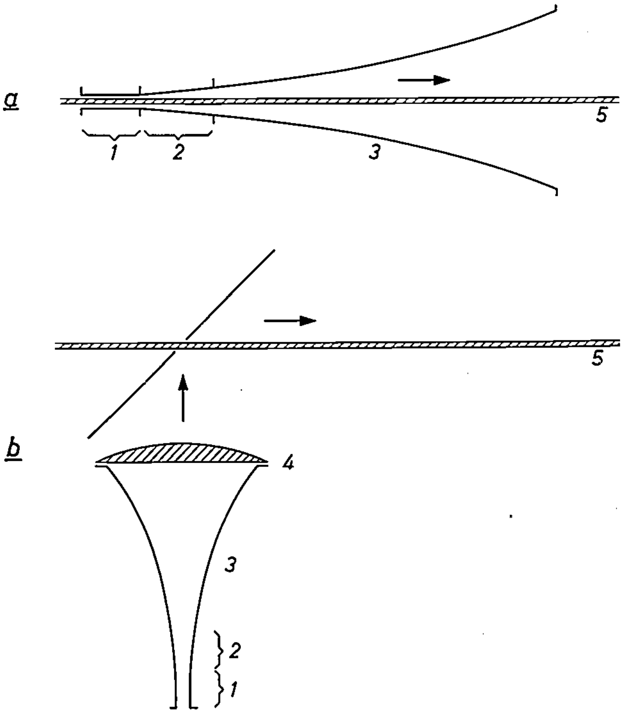


Fig. 6. Two methods of exciting the surface wave on the line. In *a*) the dielectric string *5* enters a rectangular waveguide *1* which terminates in a "rectangular-to-round" transition *2* and a circular horn *3*. In *b*) the energy is similarly radiated by a circular horn *3*, collimated by means of a dielectric lens *4* and reflected by means of a 45° reflector on to the dielectric line *5*.

transition tapers gradually from the rectangular to the circular cross-section. The dielectric string (*5*) should extend a fair distance into the waveguide. It can either taper to a point or can be led outside through a small opening in the waveguide wall. With this arrangement a mode conversion efficiency of up to 90% of the energy can be achieved. With the image line a similar arrangement can be used, but with the horn and the waveguide of semicircular cross-section.

Another method of exciting the wave is illustrated in fig. 6*b*. With the aid of a horn antenna fitted with a dielectric lens (e.g. of polyethylene) a plane wave is first produced. This wave is then directed along the dielectric string by reflection from a metal reflector. An advantage of this method compared with that shown in fig. 6*a* is that the fixing of the string presents no problems. The efficiency of the mode conversion is however somewhat smaller.

Sensitive detectors for millimetre waves exist in the form of crystal diodes and bolometers, built into waveguide. In order to use these to detect a surface wave its energy must first be reintroduced into a waveguide. This can also be done with the two arrangements illustrated, the wave then taking the reverse direction.

Circuit elements for the dielectric line

To carry out measurements on a dielectric line certain circuit components are needed, such as short-circuits, directional couplers and standing-wave indicators. Because the surface wave extends into the free space around the dielectric string, the appearance of these components is quite different from the corresponding types used in conventional waveguide techniques. Some of the circuit components will now be discussed.

a) *Reflecting short-circuit*

The reflecting short-circuit is a flat circular metal plate, placed perpendicular to the string, with the string passing through a hole at the centre. The string is fixed at some distance behind the plate. The phase of the reflected wave can be varied by moving the plate along the line.

b) *Non-reflecting termination*

A non-reflecting termination corresponds to the resistance equal to the characteristic impedance used for matched termination of an open-wire system or a coaxial cable. A termination of this kind must completely absorb the energy travelling on the line. This can be achieved with the dielectric line by terminating the line in a pyramid or cone of lossy material, such as wood. A pyramid of this kind can be seen in fig. 9 on the right. For easy mounting of the string the pyramid is in two halves, with the string running in a groove along the axis.

c) *Directional coupler*

When a wave travelling along the line is completely or partly reflected by an obstacle, a standing wave arises. One way of measuring the energy ratio of the waves travelling in the two directions is by determining the intensity of each wave separately. This can be done by means of a device which taps off part of the energy of the wave travelling in one direction. Devices of this kind are known as directional couplers. They are also used when it is required to divide between two lines the energy of a wave on a single line.

The principle of a directional coupler is illustrated in *fig. 7*. A foil or plate of dielectric material *3* is mounted at an angle of 45° to the string. A wave coming from the right is partly reflected towards a horn *2* and can be measured in a waveguide connected to the horn. A wave coming from the left, however, is not coupled to the horn, as the reflected part of this wave is absorbed by an absorption plate *4*.

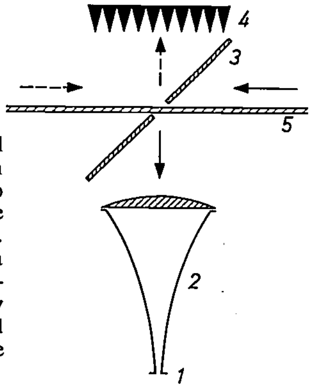


Fig. 7. Example of a directional coupler. A wave coming from the right is partially reflected to the circular horn *2* by a plate or foil of dielectric material *3*. The horn is connected to a waveguide at *1*. A wave coming from the left is also partially reflected, but this reflected energy is taken up by the absorbing element *4*.

The relative fraction of the energy reflected from the wanted wave is called the *coupling factor* s^2. This can be varied by altering the thickness d of the plate. If the thickness is sufficiently small compared with the wavelength λ, and if the reflecting plate is perpendicular to the plane of symmetry, assumed to be horizontal in fig. 3, one can write to a first approximation:

$$s = \sqrt{2}\,(\varepsilon_r - 1)\pi d/\lambda,$$

where ε_r is the dielectric constant of the plate material.

For various reasons a very small part of the unwanted wave will appear at the horn. The ratio of the wanted to the unwanted fraction is called the *directivity* of the coupler. A value of more than 40 dB can easily be achieved.

[1] S. P. Schlesinger and D. D. King, Dielectric image lines, IRE Trans. MTT-**6**, 291-299, 1958.

[2] J. C. Wiltse, Some characteristics of dielectric image lines at millimeter wavelengths, IRE Trans. MTT-**7**, 65-69, 1959.

[3] G. Schulten and H. Severin, Dämpfungsarme Leitungen für Millimeterwellen, Nachrichtentechn. Fachber. **23**, 20-23, 1961.

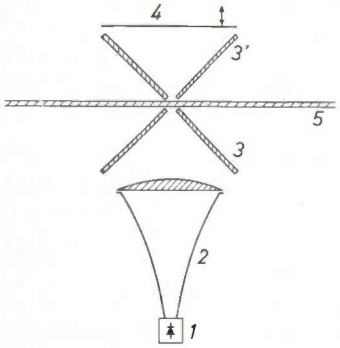


Fig. 8. The standing-wave indicator resembles the directional coupler (fig. 7), except that now the circular horn *2* receives an identical fraction of the waves coming from left and right. The distance from the reflecting plate *4* to the dielectric line *5* can be varied. The received energy is measured by a detector *1*. The whole arrangement can be moved parallel to the line.

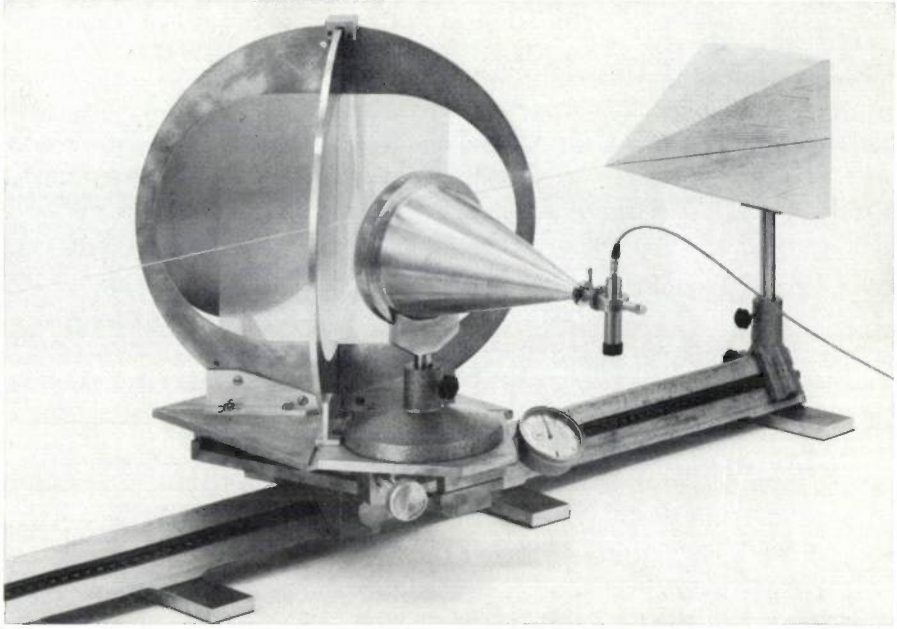


Fig. 9. A standing-wave indicator as in fig. 8. On the right can be seen a lossy wooden pyramid, which acts as a nonreflecting termination for the dielectric line.

d) *Standing-wave indicator*

When there are two waves travelling along a line in opposite directions, there are local maxima and minima of the field, because the contributions of the two waves reinforce or counteract one another. By measuring the magnitudes of these maxima and minima with an instrument that can be moved along the line a probe one can also determine the standing-wave ratio. A device has been designed for this which to a certain extent resembles the directional coupler described above (fig. 7) [4]. The principle of this device is illustrated in *fig. 8*.

The probe can be regarded as two intersecting directional couplers. One of the dielectric plates, e.g. *3* in fig. 8, reflects towards the horn a small fraction of the wave coming from the left. The other plate *3'* sends an identical fraction of the same wave to the reflecting metal plate *4*. The wave reflected by this plate also reaches the horn, after slight attenuation, and joins the fraction reflected by *3*, with a certain phase difference. By varying the distance of the metal plate *4* from the string, both waves can be brought into phase. In this way the probe can be tuned to a particular frequency. As the arrangement is symmetrical, an identical portion of the wave coming from the *right* reaches the horn in an entirely analogous manner. The waves coming from left and right are not, of course, generally in phase.

When the whole device is moved along the string, the phase between the two oppositely directed waves is altered. The detector *1* connected to the horn then records the maxima and minima, thus enabling the standing-wave ratio to be determined.

A photograph of the complete arrangement is shown in *fig. 9*. The thin plates (foils) of dielectric material are in the form of close meshes of thin polyethylene threads 0.3 mm thick. At millimetre waves these have almost the same effect as an equally thick homogeneous dielectric foil. This solution was adopted because it is not practicable to make two intersecting foils as in fig. 8. The lossy wooden pyramid on the right has already been mentioned.

e) *Resonator*

Any cylindrical line for surface waves acquires the characteristics of a resonator if it is placed between two reflecting short-circuits. Resonance occurs if the surface wave is in phase with the incident wave after two reflections. For this the length L of the line has to be equal to a whole number of half wavelengths. The effects are entirely analogous with those produced with a transmission line system used as a resonant circuit or a waveguide used as a resonant cavity.

Fig. 10 illustrates the principle of an experimental arrangement for setting up a resonator of this kind. A photograph of the arrangement is shown in *fig. 11*. The wave is introduced by means of a directional coupler as described under (c). A sufficiently small coupling factor — 20 to 30 dB — is obtained by using a very thin polyethylene foil as a coupling element. The device can be tuned to resonance by moving one

[4] G. Schulten, Eine neue Messleitung für dielektrische Oberflächenwellen-Leitungen, Nachrichtentechn. Z. **14**, 445-448, 1961.

[5] G. Schulten, Messung der Eigenschaften von dielektrischen Leitungen bei Millimeterwellen in einem optisch angekoppelten Resonator, Archiv elektr. Übertr. **14**, 163-166, 1960.

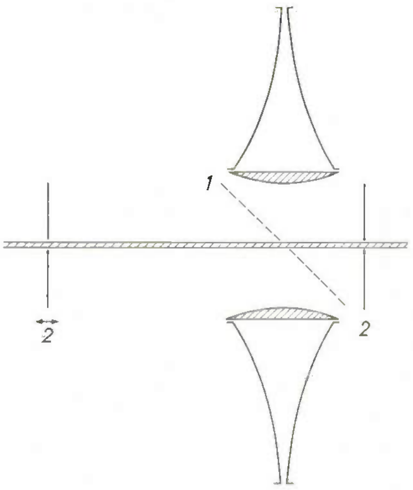


Fig. 10. Resonator consisting of a dielectric line and two circular metal end-plates *2*, the distance between which can be varied. Energy is radiated by one of the horns, and the energy absorbed can be determined by a detector connected to the other horn. The degree of coupling can be altered by suitably chosing the thickness of the polyethylene foil *1*.

of the reflecting plates along the string. At resonance maximum absorption of the wave transmitted from one horn to the other occurs. This can be ascertained by an energy measurement in the other (receiving) horn.

The quality of such a resonator is given by the *quality factor Q*. This can be defined as the relative bandwidth, if one takes for the bandwidth Δf the difference between the frequencies at which the energy taken up has dropped to half the maximum at resonance:

$$\frac{\Delta f}{f} = \frac{1}{Q}.$$

Since the foil couples the resonator with the two horns, the measured Q is lower than that of the unloaded resonator. Theoretical considerations show that the measured Q ("loaded Q") can be expressed by the following relation:

$$\frac{1}{Q} = \frac{\alpha\lambda}{\pi} + \frac{p\lambda}{2\pi L} + \frac{s^2\lambda}{2\pi L}. \quad . \quad . \quad . \quad . \quad (3)$$

Here λ is the wavelength measured on the line, α is the attenuation per unit length of the line, L the length of the line, s^2 the coupling factor of the directional coupler and $1 - p$ the reflection coefficient of the two plates. The first two terms on the right-hand side of equation 3 give the unloaded $Q = Q_0$ and the last term expresses the effect of the coupling ($1/Q_k$).

Owing to the influence of the dielectric, the wavelength on the line is always somewhat shorter than the free space wavelength. This implies that the wave does not travel along the line at the speed of light. The phenomenon of frequency-dependent phase velocity is referred to as dispersion. The arrangement described can be used for measuring both attenuation and dispersion [5]. In the resonator shown in fig. 11 the line consists of bunched dielectric threads as in fig. 4. At a wavelength of 4 mm the measured Q of this resonator was 100 000. *Fig. 12* shows a more finished version of a resonator for the same wavelength, with a Q_0 of about 200 000.

Applications

On the basis of what we now know about the properties of dielectric lines, the following brief account can be given of their possible applications. These lines are hardly suitable for long-distance communications. Although sufficiently low attenuation might be achieved, this would only be possible with a considerable radial extent of the field. Apart from the practical drawbacks of this, it would also involve difficulties in connection with screening and fixing the string. Moreover, with an extensive radial field the dielectric should run perfectly straight, or energy would be radiated laterally. For bridging short distances, for use in the laboratory and for special measurements, dielectric lines can, however, be very useful

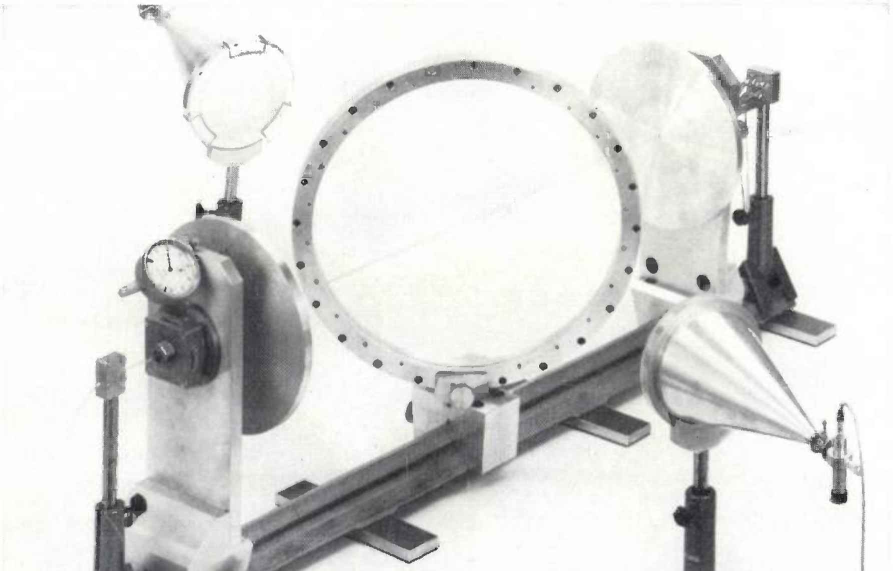


Fig. 11. Experimental form of the resonator illustrated in fig. 10. The horns have been displaced a little to show the details more clearly. The dielectric line uses the flat bunch arrangement of fig. 4.

Fig. 13 shows the principle of a measuring arrangement in which a dielectric line is used for the spectroscopic analysis of gases [6]. The string is enclosed in an air-tight tube which is filled with the gas to be analysed. The centre part of the tube is made of glass, making it possible to fit electrodes on the outside of the tube. By means of these electrodes an alternating electric field of relatively low frequency can be generated in the gas-filled tube for studying the Stark effect. The surface wave, which is set up on the line in the same way as in fig. 6*b*, is modulated in frequency [6]. If a molecular resonance occurs in the frequency interval covered, it shows up as an absorption line in the measured output voltage. Since the microwave energy is distributed over a greater cross-section than in waveguides, saturation effects of the molecular resonances become noticeable only at higher powers.

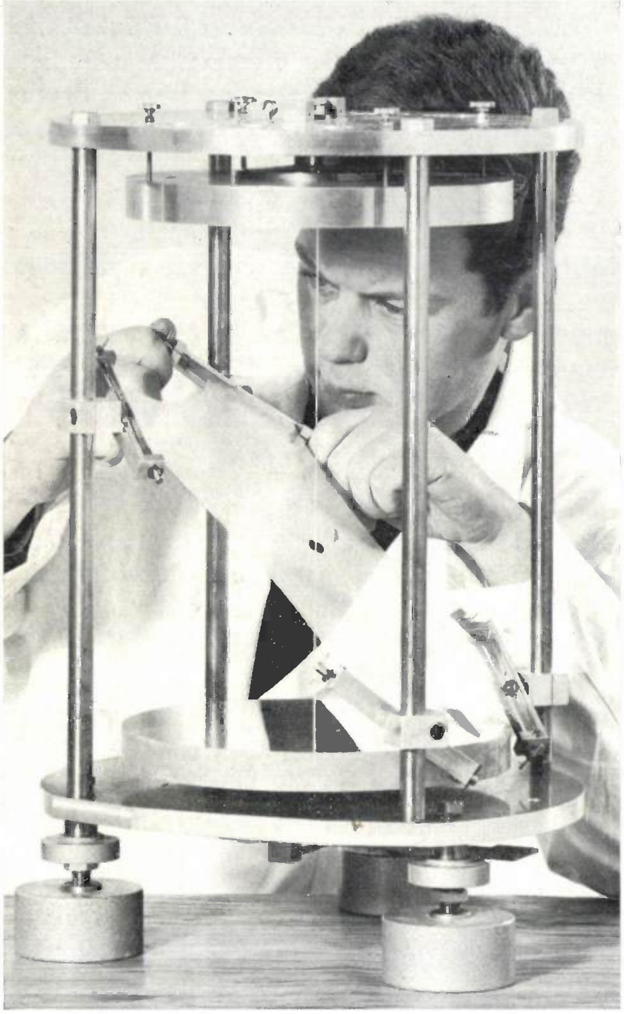


Fig. 12. Another form of a resonator with a dielectric line.

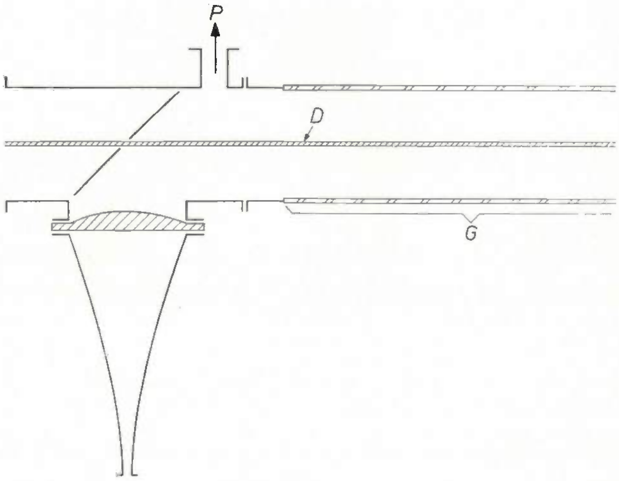


Fig. 13. Application of a dielectric line D in an absorption tube for gas spectroscopy. The tube is connected at P to a pump and is filled with the gas under investigation. Absorption of the surface wave at a particular frequency is an indication of molecular resonance. Since the tube consists partly of glass (G), external electrodes can be fitted for studying the Stark effect.

The simple construction and high Q ($\approx 10^5$) of the resonators shown in fig. 11 and fig. 12 make them suitable for many applications, such as for example wavelength measurements, klystron stabilization, measurement of the dielectric constant of gases, etc. Since they are open on all sides, they are also particularly suitable for gas masers.

[6] See also: C. W. van Es, M. Gevers and F. C. de Ronde, Waveguide equipment for 2 mm microwaves, II. Measuring set-ups, Philips tech. Rev. **22**, 181-189, 1960/61.

Summary. The fact that electromagnetic surface waves are possible on cylindrical lines can be turned to especially good use in view of millimetre wave transmission. Dielectric filaments prove to have such favourable characteristics for this purpose that there are good prospects of their use at even higher frequencies. Such dielectric lines can also be used in the form of what has been called the dielectric image line. The dimensions are always chosen so that only that mode of the electromagnetic field can occur which is known, for a circular cross-section, as the HE_{11} mode. This is the only mode that has no cut-off frequency (dominant mode). The main properties of this mode: field configuration, radial distribution of the energy and attenuation are briefly discussed. It is then shown how the wave is excited on the line and the design of appropriate circuit elements is described. Particular attention is paid to a resonator with which very high Q values can be obtained. In conclusion some possible applications are indicated, one being in an absorption tube for gas spectroscopy.

Volume 26, No. 11/12, 1965 pages 321-356 Published 31st May 1966